Felipe Martínez Rizo

EL NUEVO OFICIO DEL INVESTIGADOR EDUCATIVO

Una introducción metodológica

Felipe Martínez Rizo

EL NUEVO OFICIO DEL INVESTIGADOR EDUCATIVO

Una introducción metodológica

Felipe Martínez Rizo

EL NUEVO OFICIO DEL INVESTIGADOR EDUCATIVO

Una introducción metodológica

El nuevo oficio del investigador educativo. Una introducción metodológica. / Felipe Martínez Rizo —
México: Universidad Autónoma de Aguascalientes, 2019.
380 p.; 15.5 x 22.5 cm. — (Colec. Serie Investigación Educativa SIE)
ISBN 978-607-7923-29-9

1. Investigación. 2. Investigación educativa. 3. Metodología de la investigación. 4. Educación.
5. Felipe Martínez Rizo.

Diseño de la colección
Estudio Sagahón / Leonel Sagahón

Cuidado de la edición

Germán Álvarez Mendiola

Felipe Martínez Rizo

Corrección de estilo y de pruebas
Germán Álvarez Mendiola

Felipe Martínez Rizo

Imagen de portada
Leonel Sagahón

Formación y captura
Carmina Salas

Primera edición
Abril de 2020

ISBN: 978-607-7923-29-9

Hecho en México / *Made in Mexico*

ÍNDICE

Capítulo 4. Análisis de la información

Conclusión General

Para María Elena,
para mis hijos y nietos,
para mis alumnos pasados y ¿futuros?

En 1986 apareció la primera edición de la obra que antecedió a esta, titulada *El oficio del investigador educativo*. En 1996 vio la luz una segunda edición, con cambios no menores respecto a la primera. Una revisión de ambos textos muestra que, ya en aquellas fechas, me preocupaba la débil preparación que yo creía apreciar en los egresados de muchos posgrados que pretendían formar investigadores educativos, posgrados que, en aquel entonces, eran sobre todo gran número de maestrías y unos cuantos doctorados.

Con no poca audacia de mi parte, el libro pretendía contribuir a reforzar la formación de quienes quisieran dedicarse a la investigación de temas educativos, en particular en cuestiones metodológicas, que a mi juicio presentaban especial debilidad.

El relativo éxito que tuvieron las dos ediciones de la obra, pese a la limitada capacidad de difusión que caracterizaba los esfuerzos editoriales de las universidades públicas, junto con los comentarios que recibí de algunos lectores, confirmaron mi opinión de que un esfuerzo así era necesario.

Más de tres décadas después, el oficio del investigador educativo ha cambiado. En los países altamente industrializados, el campo ha visto transformaciones especialmente notables en lo que toca a las técnicas de análisis de información cuantitativa, con el uso cada vez más frecuente de técnicas complejas, gracias a la difusión de las computadoras personales y los paquetes de programas estadísticos.

También han sido importantes los avances en las técnicas de obtención de información y en las teorías aplicables a temas educativos desde disciplinas como la psicología, la sociología, la antropología, la economía y, desde luego, la pedagogía.

Avances aún más recientes abren perspectivas impensables hace muy poco tiempo, como la posibilidad de acceder a grandes bases bibliográficas en línea y de utilizar las

redes digitales para obtener enormes cantidades de datos. Al mismo tiempo, el campo ha vivido nuevas escaramuzas de las viejas guerras paradigmáticas, que las ciencias sociales y de la conducta conocen hace muchos años.

Con cierto desfase temporal y en grados diferentes, los cambios en México reflejan los ocurridos internacionalmente, además de los locales importantes. Cuando preparaba la primera edición de *El oficio* el Departamento de Educación de la Universidad Autónoma de Aguascalientes (UAA) compraba la primera computadora personal que hubo en la institución (con un disco duro de 10 MB y discos flexibles de 5¼ pulgadas) y la versión 2.0 del SPSS. La primera versión de este paquete de *software*, para computadoras *mainframe*, con tarjetas perforadas, apareció en 1968.

En esas fechas acababa de nacer el Sistema Nacional de Investigadores (SNI), y algunos años después el Padrón de Posgrados de Excelencia del CONACyT, antecedente del ahora llamado Programa Nacional de Posgrados de Calidad (PNPC). Con sus limitaciones, ambos muestran un innegable avance en la profesionalización de la investigación científica y en la formación de investigadores.

A lo largo de estos años, una parte significativa de mi trabajo profesional se dedicó a la formación de investigadores educativos, en los programas de maestría y doctorado de la Universidad Autónoma de Aguascalientes, en los que tuve a mi cargo buena parte de los cursos de carácter metodológico.

Durante mi paso por el Instituto Nacional para la Evaluación de la Educación, de 2002 a 2008, trabajé en temas de medición, principalmente en el contexto de la elaboración de pruebas en gran escala para primaria y secundaria, y en el estudio y medición de las prácticas docentes.

Gracias a ello hoy puedo apreciar, con más claridad que antes, las limitaciones de *El oficio del investigador educativo*, pero también que sigue vigente la necesidad de fortalecer la formación metodológica de quienes quieren dedicarse a la investigación educativa, cuya preparación en muchos casos me sigue pareciendo insuficiente, sobre todo ante los avances en la materia.

En *El oficio* cité una frase de William Spady:

> Los investigadores que no sean capaces de utilizar alguna forma de análisis multivariado con sus datos (para, por lo menos, detectar el carácter espurio de las relaciones que parezca haber entre las variables clave) deberían regresar a su alma mater y pedir que les devolvieran su colegiatura. (1970)

La frase de Spady me impactó porque hace 30 años yo creía que esa era la situación de muchos posgrados en ciencias sociales, que no aportaban a sus egresados no ya los rudimentos del análisis multivariado, sino cuestiones metodológicas más elementales.

Hoy me parece que muchas maestrías, y no pocos de los ya numerosos doctorados siguen en esa situación, en especial los que no han conseguido su aceptación en el PNPC. Por ello, y sin duda con tanta audacia como hace 30 años, preparé este libro, que no es una nueva edición de aquel, sino que pretende ser distinto, aunque tiene el mismo propósito, por lo que decidí llamarlo *El nuevo oficio del investigador educativo*.

El subtítulo de la obra (*Una introducción metodológica*) destaca que no pretende ser suficiente por sí sola, sino solo introducir a un campo cuyo dominio pleno supondría trabajar la extensa bibliografía citada, reconociendo que la variedad de acercamientos metodológicos, y el gran número de técnicas particulares que se pueden emplear, hacen impensable que una sola persona tenga un conocimiento exhaustivo.

El libro quiere ser un mapa de tan extenso territorio, que ayude a quienes quieran explorar alguno de sus rincones a ubicarlo en el conjunto, y a identificar guías que lo describan en detalle. Por ello en cada capítulo se destacan obras clave para ampliar los conocimientos de los principales temas tratados. La traducción de todos los textos que se citan de originales que no están en español es mía.

Reconozco que muchos de los ejemplos que presento se refieren a investigaciones sobre las prácticas docentes de los maestros, por ser el tema que me ha ocupado mayormente en los últimos años, aunque todo el contenido de la obra pretende ser aplicable a cualquier área de estudio en el campo de la investigación social y educativa.

Por otra parte, es claro que la obra centra la atención en los acercamientos que se suelen llamar cuantitativos, y aborda de manera solo tangencial los denominados cualitativos. Reconociendo sin ambages esta circunstancia, añado que me parece que los principios básicos de ambos enfoques no pueden ser diferentes, postura que intentaré justificar más adelante.

La gestación del libro ha sido larga. Las actividades en que he debido ocuparme en estos años, además de la docencia, no me permitieron terminarlo antes, y solo ahora lo consigo, una vez jubilado de la que fue mi casa durante 42 años.

Agradezco las observaciones que recibí de cuidadosos lectores: dos queridas colegas del Departamento de Educación de la UAA, Guadalupe Ruiz Cuéllar y María Guadalupe Pérez Martínez; un exalumno que ahora domina la metodología mejor que yo, Adán Moisés García Medina; los buenos amigos y reconocidos investigadores Rollin Kent y Romualdo López; otros dos excelentes amigos que se han especializado

en estudios de tipo interpretativo e intensivo, Genaro Zalpa de la UAA, y Eduardo Weiss del DIE (QEPD). Y sobre todo los numerosos, rigurosos y precisos señalamientos que me hizo José Felipe Martínez Fernández, mi hijo mayor, ahora jefe de la División de Metodología de la Investigación Social y profesor del Programa de Métodos Cuantitativos Avanzados en Investigación Educativa de la Escuela de Posgrado en Educación y Ciencias de la Información de la Universidad de California en Los Ángeles (UCLA), con el rigor que le permite su excelente formación metodológica y el cariño que se desprende de nuestra cercanía.

Dedico el libro a mi esposa, que durante este lapso me ha acompañado y apoyado sin fallar un solo día. A mis hijos, que padecieron mis pasiones intelectuales. Y a mis nietos, que tal vez algún día entiendan un poco de qué trataban los libros entre los que corrían cuando jugábamos a las escondidas. A mis alumnos, con quienes fueron tomando forma los capítulos de la obra, y que en varios casos son ahora colegas muy apreciados, de quienes aprendo mucho. A los alumnos de las generaciones más jóvenes, que no trataré en forma personal, pero que espero encuentren útiles sus páginas.

Aguascalientes, septiembre de 2019

Referencias

Spady, W. G. (1970). Dropout from Higher Education: An Interdisciplinary Review and Synthesis. *Interchange*, Vol. 1, No. 1.

CONTENIDO

Pedagogía experimental e investigación educativa
Concepciones de la investigación y de su metodología
Definiciones de investigación educativa
Organización de la obra

Pedagogía experimental e investigación educativa

Lo que llamamos *investigación educacional* o *educativa* comenzó a tomar su forma actual a fines del siglo XIX, aunque entonces se designaba más bien como *pedagogía experimental*. Como pasó con las disciplinas sociales y de la conducta, y las ciencias naturales, eso ocurrió en Europa occidental, en el Reino Unido, Francia, Holanda y otros países, pero en especial en Alemania, cuyas universidades eran entonces los más importantes centros de actividad intelectual del mundo.

La expresión refleja, por una parte, la influencia de educadores como Kant, Herbart y Pestalozzi o Rousseau, y por otra la de Claude Bernard, cuya *Introducción al estudio de la medicina experimental*, publicada en 1865, marcó la visión de la investigación científica, y en particular la de la naciente psicología, con el laboratorio de psicología experimental creado en Leipzig en 1879 por Wilhelm Wundt.

Según de Landsheere, la paternidad de la *pedagogía experimental* corresponde a Ernst Meumann y Wilhelm May, que en los últimos años del siglo XIX y primeros del XX acuñaron la expresión, y en 1905 crearon juntos la revista *Die Experimentelle Pädagogik*. Con la influencia de Wundt y estos dos autores alemanes, varias formas de *pedagogía experimental* se desarrollaron en las primeras décadas del siglo XX en Francia (Alfred Binet, Théodore Simon), Suiza (Édouard Claparède, Pierre Bovet, Jean Piaget), Bélgica (Médard Schuyten, Ovide Decroly, Raymond Buyse) y otros países del mundo, incluyendo a Argentina y Chile. (De Lansheere, 1986: 41-55 y 90-130)

La influencia de Wundt fue particularmente notable en los Estados Unidos, cuando sus discípulos directos o indirectos establecieron laboratorios en universidades como John Hopkins, con Stanley Hall (1882); Pensylvania (1887) y Columbia (1891) con J.

McKeen Cattell; Cornell (1891) y Stanford (1893) con J. R. Angell; y varias más en los años siguientes. (De Landsheere, 1986: 55-78)

La expresión, sin embargo, presenta dos problemas: el sustantivo *pedagogía* refiere en particular a trabajo con niños, mientras la educación se dirige también a personas de cualquier edad; y el adjetivo *experimental* designa una forma de hacer investigación importante, pero no la única.

Por ello, parece razonable que se impusiera la expresión *investigación educacional*, cuyos dos elementos son más amplios: investigación no se reduce a la experimental, sino que incluye cualquier acercamiento riguroso; y educación no se limita a los niños ni a las actividades de enseñanza y de aprendizaje, sino que incorpora trabajos con personas de cualquier edad, desde perspectivas didácticas, psicológicas, sociológicas o de otras disciplinas.

La segunda expresión prevaleció en Estados Unidos, donde el peso de filósofos y pedagogos como William James y John Dewey, no fue mayor que el de psicólogos como Edward Thorndike, el de Alfred Binet con los primeros test, y el de la estadística inglesa de Francis Galton y Karl Pearson. Desde fines del siglo XIX los pioneros estadounidenses de la investigación educativa enfrentaron la disyuntiva de definirse como disciplina científica o humanística. En el prefacio de su historia del campo, E. Condliffe Lagemann advierte:

> En enero de 1891, cuando apareció el número inaugural de la Educational Review, su primer artículo era firmado por el filósofo de Harvard Josiah Royce, y se titulaba "¿Hay una ciencia de la educación?... (2000: IX)

De Landsheere propone una periodización del desarrollo histórico de la investigación educacional, que comprende cinco etapas: a) Pre-científica, de fines del s. XVIII a fines del XIX; b) Florecimiento de la investigación cuantitativa, de fines del XIX a mediados de la década de 1930; c) Reflexión y luego estancamiento, hasta mediados de los años 1950; d) Los dorados 60s; e) Interrogación epistemológica y reconciliación entre la filosofía y las ciencias de la educación, hasta mediados de 1980. (1986: 24-27)

Esta periodización del desarrollo de la investigación educacional, que de Landsheere reconoce está inspirada en la clásica obra editada por Lee Cronbach y Patrick Suppes (1969), advierte la tensión entre la perspectiva científica y la de las humanidades, una de cuyas manifestaciones será la polémica que enfrenta a los enfoques metodológicos designados con las etiquetas simplistas de cuantitativos y cualitativos.

En las últimas décadas, el desarrollo de la investigación educativa ha continuado con importantes avances teóricos y técnicos, y también controversias que han enfrentado a los partidarios de distintos enfoques, en las *guerras paradigmáticas*, a las que se refirió N. L. Gage en la conferencia inaugural del congreso de 1989 de la *American Educational Research Association* (AERA).

En esa fecha Gage creía que *las guerras paradigmáticas habían llegado a un sanguinario climax*, y consideraba como posibles escenarios futuros del campo para el 2009, a) que triunfara la postura cualitativa; b) que, superadas las diferencias, los dos bandos colaboraran; y c) que persistiera la confrontación. (Gage, 1989)

En 2019 podemos ver que, al final del siglo pasado y en los primeros años del actual, avances técnicos en el campo de la psicometría y la computación hicieron posible la extensión masiva de pruebas en gran escala, como las conocidas con la sigla PISA de la OCDE, lo que contribuyó a exacerbar antiguas críticas a los sistemas educativos, y a que los gobiernos de muchos países adoptaran medidas con las que esperaban mejorar la calidad educativa a partir de los resultados de tales pruebas.

En el caso de los Estados Unidos la adopción de una ley, a la que se hará referencia más abajo, hizo que la política gubernamental solo considerara investigación educativa seria la de tipo experimental, con la consecuente reacción de la comunidad académica.

En México las décadas transcurridas de 1981 han visto consolidarse la investigación educativa, a partir del primer congreso nacional, y luego con la fundación del Consejo Mexicano de Investigación Educativa (COMIE). El número de personas dedicadas profesionalmente a esta actividad aumentó considerablemente, al igual que el de los centros especializados, los proyectos, los libros y las revistas que difunden resultados, entre otras cosas. Las distintas áreas temáticas tienen, desde luego, un desarrollo desigual, las polémicas entre partidarios de los distintos enfoques siguen vivas, si bien parecen quedar atrás las posturas radicales. (*Cfr*. por ejemplo López, Sañudo y Maggi, 2013)

Concepciones de la investigación y de su metodología

De Landsheere recuerda la diferencia que hay entre acciones muy distintas, que en español se designan con un mismo verbo *experimentar*. En el sentido que implica el calificativo *experimental*, y en el que se refiere a una *experiencia* cualquiera.

La diferencia entre *experimentar* y tener *experiencias* (en francés *expériencier*), es la que hay entre un trabajo hecho con rigor y otro sin él. Calificar de *experimental* la pedagogía implicó que la nueva disciplina adoptaba la rigurosa metodología de las

ciencias exactas, distinguiéndose del escaso rigor de las muchas experiencias que podían tener lugar en educación, pero no se sometían a ningún tipo de control.

A fines del siglo XIX el término *experimento* no tenía la precisión que Fisher le dio a mediados de la década de 1920, con la manipulación de una o pocas variables, el control de las demás con un grupo de tratamiento y uno de contraste, y la asignación aleatoria de casos a uno y otro. La dificultad que muchas veces se encuentra en ciencias sociales y de la conducta para hacer experimentos hizo que el diseño de investigación más frecuente en estas disciplinas de 1930 a 1960 fuera el de los estudios de tipo encuesta.

En la segunda mitad del siglo XIX la visión dominante en filosofía de la ciencia era el positivismo, no tanto de Comte, aunque se le suele atribuir, sino de científicos y filósofos como Claude Bernard, Henri Poincaré, Pierre Duhem o Ernst Mach, cuya visión de la ciencia era bastante rica, pero prevaleció una versión empirista, según la cual sólo es conocimiento sólido el que se basa en las percepciones que el cerebro recibe de los sentidos. Y si la ciencia es un conocimiento de especial calidad, en la perspectiva de la versión simple del positivismo quiere decir que se basa en percepciones de especial calidad también.

En esa perspectiva, la forma de trabajo de la ciencia, la *metodología científica* se concibe como un esfuerzo por asegurar la calidad de las observaciones, evitando que sean contaminadas por las ideas que tengan los investigadores, distintos tipos de *ídolos* o prejuicios identificados por Bacon en su *Novum Organum* (1620): en vez de atenerse a los hechos, hacer caso a la tradición (*idola theatri*); a lo que dice la mayoría de la gente (*idola fori*); a las ideas personales favoritas (*idola specus*); o a los prejuicios genéricos (*idola tribus*).

La versión simple de empirismo-positivismo lleva a reducir la metodología al cuidado de las técnicas de obtención de información, que en las ciencias naturales suele implicar el desarrollo de instrumentos que permiten observaciones cada vez más precisas. En ciencias sociales y de la conducta, la postura empirista-positivista simple que predominó hasta la década de 1950 también identificó la metodología con el uso de técnicas estadísticas de obtención y tratamiento de datos, e incluso la redujo a la aplicación rutinaria de un esquema rígido de pasos. Los estudios más usuales no eran experimentos, sino encuestas, por lo que la receta que encarnaba el *método científico* incluía: a) seleccionar un tema; b) plantear un problema; c) redactar un marco teórico; d) formular una hipótesis; e) seleccionar o diseñar un instrumento de obtención de datos; f) diseñar una muestra; g) aplicar el instrumento; h) procesar la información; i) analizar los datos.

Desde la década de 1950, y con fuerza a partir de la de 1960, la versión simple del positivismo-empirismo fue cuestionada, dando lugar a la etapa que de Landsheere llama de *interrogación epistemológica*.

Algunos críticos del positivismo sostienen que no habría razón para considerar que las ciencias pueden producir conocimientos mejores que el sentido común o las llamadas pseudo-ciencias, y que la idea misma de verdad debería abandonarse, dejando lugar a posturas relativistas y constructivistas extremas. La metodología de las ciencias sociales no tendría que ver con la de las ciencias naturales, y habría que abandonar cualquier pretensión de identificar una metodología de cualquier ciencia, dado que la única regla aceptable sería *todo vale*. (Feyerabend, 1970)

Una postura epistemológica intermedia entre el positivismo-empirismo simple y sus críticos extremos, sustenta también una visión equilibrada de la metodología de las ciencias, que no se reduce a técnica alguna en particular, a un conjunto de técnicas o una secuencia de los pasos a seguir para asegurar la calidad de un trabajo.

Esta postura intermedia considera que el conocimiento no es ni sola percepción, ni sola interpretación, sino una integración de ambos elementos, y el método se entiende como *la sistematicidad de la relación entre percepción e interpretación* o, si se quiere, entre teoría y experiencia.

En esta obra la metodología de la ciencia se entiende como:

> [...] un proceso en el que [...] no se observa lo primero que cae en el ámbito de la visión, y no se especula en forma independiente, estableciendo raramente la comparación o la contrastación entre lo que se capta por la experiencia y lo que se especula; sino que, por el contrario, se piensa previamente qué es lo que hay que observar y se observa a partir de lo que se piensa: la observación es guiada y orientada por la interpretación, y esta, a su vez, se ve enriquecida, o por el contrario cuestionada por la experiencia [...]
>
> [...] Ninguna técnica en particular es indispensable para el trabajo científico; ningún paso es absolutamente indispensable, ni es forzoso seguir un orden determinado en todos los casos. Lo único que es indispensable es que exista una concatenación, una retroalimentación continua y un apoyo mutuo entre pensamiento y observación: que se piense que se va a observar, que se observe en función de lo que indicó el pensamiento, que se retroalimente al pensamiento con el fruto de la observación, que se enriquezcan la observación y sus procedimientos a partir de las consideraciones hechas por el pensamiento, y así sucesivamente. Este va y viene entre experiencias e interpretación, deberá continuar permanente, tenaz, machacona y, a la vez, creativamente. (*Cfr*. Martínez, 1997:152-153)

Definiciones de investigación educativa

Un ejemplo de noción restrictiva de la investigación educativa, importante por sus repercusiones prácticas para el financiamiento de proyectos, es la que ofrece la ley conocida con la expresión *No Child Left Behind* (NCLB), que aprobó en 2001 el Congreso de Estados Unidos. Según esta definición:

> La expresión "investigación fundamentada científicamente" significa una investigación que implica aplicar procedimientos rigurosos, sistemáticos y objetivos para obtener conocimiento confiable y válido que sea relevante para actividades y programas educativos, e incluye la investigación que emplea métodos empíricos sistemáticos basados en observación y experimentación; implica análisis de datos rigurosos [...]; se basa en métodos de medición u observación que produzcan datos confiables y válidos [...]; se evalúa mediante diseños experimentales o cuasi-experimentales [...]; asegura que los estudios experimentales se presenten con suficiente detalle para permitir que sean replicados [...]; y ha sido aceptada por una revista arbitrada o aprobada por un panel de expertos independientes [...]. (US Congress, 2001, en Kelly, 2006: 55)

La reacción de la comunidad académica norteamericana se refleja en una obra publicada un año después de la aprobación de la Ley NCLB. La obra es el resultado del trabajo de un comité formado por la División de Ciencias Sociales y de la Conducta y de Educación del Consejo Nacional de Investigación (*National Research Council*, *NRC*) de los Estados Unidos, que desarrolla y discute una concepción más comprensiva de la investigación educacional.

El comité, encabezado por Richard Shavelson, e integrado por otros 15 reconocidos especialistas de orientaciones diversas, buscaba ofrecer elementos para fortalecer la investigación, con visión inclusiva que incorporara trabajos sólidos, experimentales o de otros tipos.

El Comité delinea seis principios fundamentales de la investigación científica, incluyendo la educacional:

- Plantear preguntas significativas que puedan investigarse empíricamente.
- Relacionar la investigación con la teoría relevante.
- Utilizar métodos que permitan la investigación directa de la pregunta.
- Ofrecer una cadena de razonamiento explícita y coherente.
- Tener la posibilidad de replicación y generalización entre estudios.

- Dar acceso a la investigación para propiciar escrutinio y crítica profesional (Shavelson y Towne, 2002: 3-5)

Enfatizando que estos principios no deben entenderse limitando su aplicación a los enfoques experimentales y los llamados cuantitativos, el Comité precisa:

> Las características de la educación, con los principios rectores de la ciencia, definen las fronteras para el diseño de investigaciones científicas en educación. El diseño de un estudio no lo hace científico por sí mismo. Hay una amplia gama de diseños legítimos que se pueden usar en la investigación educativa que van de un experimento con asignación aleatoria para estudiar un programa de bonos educativos, a un estudio de caso etnográfico en profundidad de unos maestros, o a un estudio neurocognitivo de cómo se aprenden los números, utilizando tomografía por emisión de positrones para formar imágenes del cerebro. (Shavelson y Towne, 2002: 6)

En 2008 la *American Educational Research Association* adoptó también una noción de *investigación fundamentada científicamente* que no la reduce a lo experimental y cuantitativo, sin abrir la puerta a relativismos extremos. Según la AERA:

> La expresión investigación fundamentada científicamente denota el uso de metodologías rigurosas, sistemáticas y objetivas para obtener conocimiento válido y confiable. Específicamente, tal tipo de investigación requiere:
>
> - Desarrollar una cadena de razonamiento lógico basado en evidencias.
> - Métodos apropiados para responder las preguntas planteadas.
> - Diseños observacionales o experimentales e instrumentos que ofrezcan hallazgos confiables y generalizables.
> - Datos y análisis adecuados para apoyar los hallazgos.
> - Explicación clara y detallada de procedimientos y resultados, precisando la población a la que se pueden generalizar los hallazgos.
> - Adhesión a las normas profesionales de la revisión por pares.
> - Diseminación de los hallazgos para contribuir al conocimiento científico.
> - Acceso a los datos para que se pueda hacer análisis secundarios de ellos, replicarlos y para tener la oportunidad de construir a partir de los hallazgos. (AERA, 2008)

A lo anterior, la AERA añade dos precisiones, una que reconoce el papel de los diseños experimentales, y otra que subraya que la definición anterior es aplicable a los diversos tipos de investigación:

> El examen de preguntas causales requiere diseños experimentales que usen asignación aleatoria, o diseños cuasi-experimentales u otros que reduzcan substancialmente explicaciones alternativas plausibles de los resultados. Estos diseños incluyen, sin limitarse a ellos, estudios longitudinales, métodos de control de casos, apareamiento estadístico o análisis de series de tiempo. Este estándar se aplica especialmente a estudios que evalúen el impacto de políticas y programas sobre los resultados educativos.
>
> La expresión "investigación fundamentada científicamente" incluye la investigación básica, investigación aplicada e investigación evaluativa, en las que la justificación (rationale), el diseño y la interpretación de los resultados se desarrollen de acuerdo con los principios científicos antes enumerados. El término es aplicable a todos los mecanismos de apoyo federal a la investigación, tanto la iniciada en el terreno como la dirigida. (AERA, 2008)

Una versión operacional de esta definición puede encontrarse en otro texto de la AERA (2006): los estándares que la Asociación decidió utilizar como criterios para aceptar publicar en sus revistas reportes de investigaciones empíricas. En el Recuadro I.1 se puede ver una síntesis de ese documento, preparada por el autor de esta obra.

N. B. En la Introducción, los Capítulos 1 a 4 y la Conclusión de la obra, los recuadros, tablas y gráficas se enumeran con un formato que inicia con 1, precedido en cada parte por la letra I o C, o los números 1 a 4: I.1, 1.1, 2.1, 3.1, 4.1, C

RECUADRO I. 1. ESTÁNDARES PARA REPORTAR INVESTIGACIONES EMPÍRICAS EN PUBLICACIONES DE LA AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

Se aplican a reportes basados en las tradiciones empíricas, y cubren métodos cuantitativos y cualitativos. Aunque pueden ser valiosos, los estándares no se aplican a revisiones bibliográficas, ensayos teóricos, conceptuales o metodológicos, críticas a tradiciones y prácticas de investigación, o de historia, filosofía, análisis literario o estudios artísticos.
Los estándares se basan en dos principios:

- Que los informes deben estar respaldados, presentando evidencia adecuada para justificar los resultados y conclusiones.

- Que deben ser transparentes, explicitando la lógica de la investigación y los pasos que llevarondel desarrollo del problema o pregunta de investigación a resultados, pasando por la definición, recolección y análisis de los datos o la evidencia empírica.

Se describen en detalle los siguientes puntos que los textos deberán cubrir:

1. Formulación del problema: propósito; contribución a conocimiento del campo; revisión de investigación relevante; justificación de orientación conceptual, metodológica, teórica con referencias pertinentes de lo que otros han escrito; justificación del grupo estudiado, en especial en relación con rasgos históricos, lingüísticos, sociales, culturales.

2. Diseño preciso e inequívoco; lógica del estudio: del planteamiento a resultados.

3. Fuentes de evidencia: unidades de estudio, forma en que fueron elegidas; cómo y cuándo se recabaron los datos o los materiales empíricos.

4. Medición/clasificación: desarrollo de instrumentos; esquemas de clasificación, ejemplos del rango de fenómenos; características de la medición: estadísticas descriptivas, escalas e índices, con medidas de confiabilidad, técnicas de reducción; convenciones y símbolos si se usan transcripciones; justificar relevancia de medición/clasificación.

5. Análisis e interpretación:
- En general: procedimientos utilizados; técnicas de análisis; forma en que resultados apoyan conclusiones; circunstancias que afecten interpretación; conclusiones y su relación con problema o pregunta.
- Estudios cuanti: análisis usados, pertinencia; estadísticas descriptivas-inferenciales; circunstancias en recogida o análisis de datos que puedan comprometer validez de inferencias; para c/ resultado importante, índices de su distribución en estudios descriptivos o relación con otras variables, con indicaciones de incertidumbre y nivel de significación; interpretación cualitativa en relación con las preguntas del estudio.
- Cualitativos: proceso de desarrollar descripciones interpretaciones; la evidencia que soporta cada afirmación; prácticas usadas para desarrollar y mejorar los soportes; comentarios interpretativos.

6. Generalización. Información de participantes, contextos, recolección y manejo de datos; propósito de generalización, si procede; lógica para aplicar hallazgos según propósito.

7. Ética del reporte. Consideraciones éticas de recolección, análisis y reporte resultados. Evidencias de respetar compromisos de consentimiento informado. Posibles sesgos o conflictos de intereses. Garantías de que no se manipuló la información. Accesibilidad de material empírico relevante para réplicas. Información de fuentes de financiamiento.

8. Título que informe sobre el contenido del trabajo; resumen autocontenido, breve y preciso; encabezados-subtítulos que evidencien lógica subyacente.

AERA (2006). STANDARDS FOR REPORTING ON EMPIRICAL SOCIAL SCIENCE RESEARCH IN AERA PUBLICATIONS. EDUCATIONAL RESEARCHER, VOL. 35 (6): 33-40.

King, Keohane y Verba muestran que principios metodológicos considerados propios de estudios cuantitativos se aplican en los cualitativos; a su juicio las inferencias que se deben hacer en toda investigación adquieren un papel central, y precisan:

> Como debería ser claro, no consideramos más científica la investigación cuantitativa que la cualitativa. Una buena investigación —o sea la investigación científica— puede ser cuantitativa o cualitativa en cuanto a estilo. Pero en cuanto a diseño, la investigación tiene estas cuatro características:
>
> - El objetivo es la inferencia...
> - Los procedimientos son públicos...
> - Las conclusiones son inciertas...
> - El contenido es el método... (King, Keohane y Verba, 1994: 7-9)

Según Vogt (2007: 6) hay tres elementos clave de toda investigación: diseño, medición y análisis. Se asocian a ellos cuatro tipos de inferencia, cuya validez debe cuidarse. Esto implica que el conocimiento humano no se reduce a percepciones; debe partir de ellas, pero debe interpretarlas la mente, que necesariamente infiere: a partir de información parcial, da el salto a conclusiones más amplias. El salto se puede dar bien o mal, y la inferencia puede ser válida o no; puede ser de distinto tipo, pero no se puede prescindir de ella. Está presente en el sentido común y en la investigación más rigurosa. En el conocimiento cotidiano no hay un cuidado sistemático de la solidez de las inferencias que se hacen continuamente. En el conocimiento científico debe haberlo: hay que identificar distintos tipos de inferencia, las amenazas que pueden afectarlos y la forma de reducir su impacto. *La metodología de investigación se puede entender como la sistematización de las formas de cuidar la calidad de las inferencias, su validez.*

La tradición experimental distingue cuatro tipos (Shadish, Cook y Campbell, 2002: 38):

- Validez de constructo: solidez de las *inferencias descriptivas* que se hacen a partir de las mediciones, según la calidad de la operacionalización de los constructos y las técnicas de obtención de datos.
- Validez de conclusiones: solidez de las *inferencias asociativas* que concluyen que hay correlación, según las técnicas utilizadas y la forma de usarlas.

- Validez interna: solidez de las *inferencias causales*, según calidad del diseño, variables incluidas y controladas, sujetos, hipótesis, condiciones, tiempo, etc.
- Validez externa: solidez de las *inferencias generalizadoras* de los hallazgos, debido a la representatividad de la muestra utilizada.

La relación de los tipos de validez de la tradición experimental con los que se definen con base en el tipo de inferencia al que se refieren puede expresarse como sigue:

TABLA I. 1. DOS FORMAS DE CLASIFICAR LOS TIPOS DE VALIDEZ

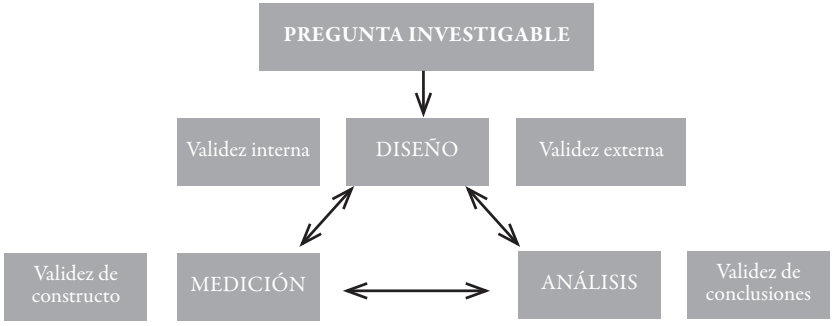
Terminología según tipo de inferencia	Tradición experimental
Validez de inferencias descriptivas relativas a la medición (incluye confiabilidad)	Validez de constructo
Validez de inferencias asociativas	Validez de conclusiones
Validez de inferencias causales o explicativas	Validez interna
Validez de inferencias generalizadoras	Validez externa

FUENTE: ELABORACIÓN PROPIA.

Es importante subrayar que hay distintas formas de hacer investigación, y que los distintos tipos de inferencia no están presentes en todas ellas:

- No toda investigación tiene propósitos explicativos ni tampoco generalizadores; muchos trabajos pretenden sólo describir cómo se comportan una o más variables, limitándose a los casos estudiados.
- Otros estudios descriptivos también buscan generalizar los hallazgos a una población o universo.
- Los trabajos que sí buscan identificar causas suelen ser estudios experimentales, aunque también lo pueden pretender trabajos correlacionales.
- En uno y otro caso, se puede o no buscar también generalizar los hallazgos.
- Los elementos de Vogt (inspirados en Pedhazur y Schmelkin, 1991), y los tipos de validez relacionados, se pueden presentar gráficamente como sigue:

GRÁFICA I. 1. LA *SAGRADA TRINIDAD* DE LA INVESTIGACIÓN Y LOS TIPOS DE VALIDEZ



FUENTE: ELABORACIÓN PROPIA A PARTIR DE VOGT (2007: 6).

Para definir lo que se entiende en esta obra por investigación educativa, de lo anterior se concluye que el alcance del término *investigación* no debe restringirse a estudios experimentales o que usen estadísticas complejas, pero tampoco puede aplicarse a cualquier trabajo, aunque carezca de un mínimo de rigor. No hay que limitar el sentido del término *educación* a la formal o escolar, pero tampoco extenderlo a cualquier tipo de proceso de socialización o aculturación. Por ello se retoma la definición de la obra que precedió a esta, entendiendo por investigación educacional:

> El conjunto de estudios de carácter básico o aplicado, desde el punto de vista de cualquier disciplina, no necesariamente sofisticados, pero siempre rigurosos, que utilicen cualquier tipo de metodología o enfoque particular siempre y cuando tenga fundamentación teórica y consistencia metodológica, sobre cualquier aspecto del fenómeno educativo en el sentido escolar formal, no formal o informal. (Martínez, 1997: 17)

RECUADRO I. 2. MI POSTURA SOBRE LOS ENFOQUES CUANTI-CUALI

No creo adecuadas las etiquetas cuantitativo y cualitativo. En sentido estricto, cuantificar se refiere al nivel de medición de las variables y, sea cual sea el enfoque de una investigación, las variables pueden medirse a nivel nominal, ordinal o métrico. El enfoque cuali se define como interpretativo, ya que en ciencias sociales es fundamental entender el sentido que tiene un fenómeno para los actores. Esto es verdad, pero no implica sostener concepciones epistemológicas incompatibles. En ambos enfoques se debe reconocer que no hay "datos" totalmente seguros, ni basados en percepciones sensoriales (empirismo) ni en impresiones subjetivas (hermenéutica).

El conocimiento supone siempre inferencias de varios tipos, y la metodología finalmente busca asegurar su solidez. La diferencia entre enfoques tiene que ver más con el número de casos y aspectos que se incluyan, y la profundidad y amplitud con que se haga, contrastando un enfoque sintético e intensivo y uno analítico y extensivo.

Quienes se oponen a los enfoques convencionales de manera radical postulan que habría diferencias irreconciliables en cuanto a la concepción misma del objeto de estudio y su conocimiento, en los niveles ontológico y epistemológico. Con Shavelson y Towne (2002) y otros autores sólidos (Cfr. conclusión) creo equivocadas las posturas radicales. Pienso que las posturas cualitativas (interpretativas e intensivas) son perfectamente compatibles con las cuantitativas, analíticas y extensivas. Más allá de obvias diferencias técnicas, en ambos casos se aplican las mismas concepciones filosóficas y epistemológicas. Las diferencias son de énfasis, y se refieren a las técnicas de obtención y análisis de la información. Son estilos complementarios; en ambos hay trabajos de poca y mucha calidad. Sintéticamente, justifico mi postura como sigue:

1. El conocimiento humano es inevitablemente imperfecto y analítico. Es imposible captar la totalidad de la realidad en forma estrictamente holística. Solo podemos conocer **aspectos** de la inmensa totalidad de la realidad, siempre relacionados con todos los demás, pero que hay que abstraer para poderlos captar.

2. Toda investigación comienza con la precisión de aspectos de la realidad que interesa estudiar, aspectos que no son inmutables, sino cambiantes: aumentan o disminuyen; comienzan o terminan (son **variables**), y eso es lo que hace interesante su estudio.

3. Según el propósito que se persiga, el estudio de los aspectos (variables) que interesan en la inmensidad inabarcable de la realidad implica al menos algunos de estos pasos:

- Obtener información de cada aspecto observándolo o midiéndolo en forma precisa.
- Describir cada aspecto, detectando que tan homogéneo o heterogéneo es.
- Identificar patrones en la forma en que cambia cada aspecto en el tiempo.
- Identificar patrones de asociación entre pares de aspectos.
- Detectar qué pasa con esas asociaciones si se hacen intervenir otros aspectos.
- Identificar conjuntos de aspectos que se relacionan de cierta forma.
- Explicar en términos causales ciertos aspectos a partir de otros, asegurando que sean esos aspectos y no otros los que incidan en los de interés.
- Analizar el impacto de ciertas intervenciones en aspectos de interés.
- Generalizar los hallazgos de un estudio a un universo más amplio.

4. Dependiendo del tipo de aspectos de que se trate, todos estos pasos se pueden dar de diversa forma, utilizando técnicas cuantitativas o cualitativas de obtención o análisis de la información, pero la lógica a seguir para describir, identificar patrones temporales o asociaciones, explicar o generalizar es fundamentalmente la misma, y siempre hay que cuidar la solidez de las inferenciasque inevitablemente hay que hacer.

Organización de la obra

Una investigación debe partir de preguntas susceptibles de respuesta empírica, relacionarse con la teoría, usar métodos que permitan investigar la pregunta, ofrecer una cadena de razonamiento explícita y coherente, poder replicarse y generalizarse, y abrirse al escrutinio y crítica (Shavelson y Towne, 2002). Los elementos esenciales de una investigación, a partir de una pregunta investigable, son diseño, medición y análisis (Vogt, 2007). Los capítulos del libro se organizan con base en la concepción de la investigación y su metodología que resumen esas propuestas. Esa visión puede expresarse en la forma de una lista de habilidades que debe dominar un investigador, que se sintetizan en el Recuadro 2 y se desarrollan en seguida.

RECUADRO I. 3. HABILIDADES QUE DEBE DOMINAR UN INVESTIGADOR

Para el inicio de una investigación
- Identificar fenómenos y aspectos a estudiar
- Formular preguntas susceptibles de respuesta empírica
- Revisar estudios previos y elaborar un marco teórico
- Seleccionar un diseño adecuado para responder las preguntas

Para la obtención de información empírica
- Seleccionar acercamiento(s) para obtener información
- Recolectar la información (observar/medir, registrar)
- Cuidar la calidad de la información obtenida

Para el tratamiento de la información empírica
- Describir, comparar, clasificar
- Identificar patrones, tendencias y asociaciones
- Controlar variables y proponer explicaciones causales
- Extrapolar, generalizar

Para la etapa final de la investigación
- Argumentar con base en evidencias
- Producir materiales de difusión de varios tipos

FUENTE: ELABORACIÓN PROPIA.

Habilidades para el inicio de un proceso de investigación

- Identificar fenómenos y aspectos a estudiar (*construir el objeto*)

Una investigación comienza cuando algo llama la atención y hace preguntar qué pasa allí, cómo o por qué ocurre. Dado que todo fenómeno forma parte de la inmensa com-

plejidad de la realidad, el primer paso de una indagación se da cuando se identifica uno o varios *aspectos* de ese todo inmenso, que son los que se buscarán entender o explicar. Además, si un aspecto de la realidad llama la atención es que no está simplemente allí, sin cambio: unas veces está y otras no; aparece o no; nace o muere; aumenta o disminuye; crece o decrece; cambia de forma o color; hace ruido o no; en síntesis, *varía*. Por ello esta habilidad se puede definir como *identificación de variables*.

- Formular preguntas susceptibles de respuesta empírica

El siguiente paso es la *formulación de preguntas* sobre el fenómeno y sobre los aspectos de interés. También se puede hablar de *hipótesis*. Investigar no es una actividad caprichosa; comienza planteando preguntas o explicaciones tentativas. Hay preguntas vagas o triviales, y otras más precisas que abren caminos a la búsqueda y dan lugar a observaciones, cuyo resultado puede ser acorde o contrario a lo que se esperaba, en tanto que una pregunta vaga no puede ser confirmada ni refutada. Una buena pregunta es susceptible de ser respondida gracias al resultado de observaciones; una inadecuada no puede dar lugar a observaciones que la respondan. Preguntarse si los niños aprenden más que las niñas o viceversa, o si el aprendizaje tiene que ver con las condiciones del hogar o con las prácticas docentes, se pueden responder con apoyo en observaciones. Pero no es posible responder con observación empírica si es aceptable o inaceptable tratar igual a niños y niñas.

- Revisar literatura de estudios previos y redactar un *marco teórico*

Esta habilidad tiene que ver con la relación con la teoría, a partir de la idea de que la ciencia avanza aprovechando hallazgos previos y no comienza de cero. Incluye dos habilidades particulares: *identificar fuentes de información* sobre el tema, incluyendo enciclopedias, libros generales y revistas, textos electrónicos, y conocimientos de personas con experiencia, como maestros; y *distinguir fuentes de distinta calidad*, pues no todos los libros o revistas, ni los sitios de Internet, son de la misma calidad, e incluso los mejores pueden equivocarse; tampoco todas las personas son igual de competentes; hay que comparar críticamente diversas fuentes para identificar las más consistentes.

- Seleccionar un diseño adecuado para responder las preguntas

Técnicas de obtención o de análisis de información hay centenares. Diseños, en el sentido de una forma particular de organizar el estudio, utilizando ciertas técnicas, no

hay muchos, y es importante que el investigador conozca sus características, para que, teniendo en cuenta los recursos de que dispone, pueda escoger el más adecuado, o los más adecuados, para responder sus preguntas, evitando quedarse limitado a las falsas disyuntivas de encuesta vs estudio de caso, o enfoque cuanti vs cuali.

Habilidades para la obtención de información empírica

- Seleccionar acercamiento(s) para obtener información

Hay también cientos de técnicas para ello, que se pueden agrupar en tres grandes conjuntos: las que se basan en las respuestas de algunas personas; las que emplean la observación de las conductas de los sujetos estudiados; y las que recaban información a partir del análisis de materiales generados por la actividad de dichos sujetos. Es importante que el investigador conozca las características generales de los instrumentos de esos tres grupos, que sepa utilizar algunos, y sepa cómo conseguir más información sobre otros, cuando llegue a necesitarla.

- Recolectar la información (observar/medir, registrar)

Medir es una forma de observar, cuando se observa algo cuantificable. Según una vieja definición, *cantidad es lo que puede aumentar o disminuir*, que *varía*; todo aspecto observable puede medirse, teniendo en cuenta que se puede hacer en varios niveles; que una medición puede ser más o menos precisa; que hay distintas razones por las que eso ocurre (fuentes de error); y que hay formas de mejorar la precisión. Para poder analizar los resultados y llegar a conclusiones, hay que consignar cuidadosamente los resultados de observaciones o mediciones, buscando la máxima fidelidad, sea que esa información se codifique con palabras, números o imágenes.

- Cuidar la calidad de la información obtenida

Todo investigador, independientemente de su preferencia por uno u otro tipo de diseño y de técnicas, deberá entender las nociones de confiabilidad y validez, así como la forma de cuidar ambas, de manera congruente con el tipo de información de que se trate.

Habilidades para el tratamiento de la información empírica obtenida

Obtenida la información, hay que analizarla para obtener conclusiones. Según la naturaleza de la información, el análisis se hará con herramientas cuantitativas o cualitativas, pero en uno u otro caso implicará habilidades particulares.

- Describir, comparar, clasificar

Un primer nivel de análisis consiste simplemente en la descripción de lo observado/medido en cada uno de los aspectos particulares considerados, sin relacionar un aspecto con otros ni pretender explicar a qué se debe. Un paso más consistirá en la comparación de sujetos y/o aspectos, llegando a algún tipo de agrupamiento de casos que tengan parecido o compartan ciertos rasgos: una clasificación o tipología.

- Identificar patrones, tendencias y asociaciones

La identificación de secuencias típicas o patrones puede referirse a un solo aspecto o variable observando, por ejemplo, si muestra tendencia creciente o decreciente, regular o irregular o con cierta periodicidad. Un tipo de patrón se refiere a dos o más aspectos o rasgos, en una asociación del tipo *cuando tal rasgo aumenta, tal otro aumenta también o disminuye*. Identificar este tipo de relación puede hacerse con técnicas estadísticas de correlación.

- Controlar variables y proponer explicaciones causales

El investigador debe tener claro que *correlación no implica necesariamente casualidad*. Esto conlleva a comprender la noción de correlación espuria, y que para poder atribuir efecto causal a una variable hay que descartar que ese efecto se deba realmente a otra u otras, para lo que hay que *controlarlas* todas, con excepción de aquella cuyos efectos se quiere estudiar, y que esa es precisamente la lógica de un experimento, que muchos consideran el diseño ideal para llegar a conclusiones en términos causales. Sin olvidar que no toda investigación debe responder preguntas causales, el investigador deberá entender la lógica del control de variables, las amenazas que pueden minar la solidez de conclusiones, y las aproximaciones cuasi-experimentales.

- Extrapolar, generalizar

Las observaciones/mediciones se pueden referir a todos los sujetos de una población o a parte de ellos. Un investigador deberá conocer las nociones de censo y muestra, advirtiendo que, si unas observaciones solo se hicieron sobre una muestra, generalizar los hallazgos a toda la población implica una inferencia, cuyo fundamento dependerá al menos de cuántos sujetos se observaron, de cómo se escogieron, y de qué tan parecidos o diferentes son en comparación con los sujetos no considerados en la observación.

Habilidades para la etapa final de la investigación

- Argumentar con base en evidencias

Las conclusiones de una investigación deben basarse en la discusión de los resultados del análisis, sustentándose en la evidencia disponible. Esta etapa implica la habilidad general de redacción, precisando que no es lo mismo un texto bien escrito, incluso bello, que un texto sólido. Importa aprender a distinguir explicaciones aparentes, que utilizan una palabra "científica" para dar cuenta de un fenómeno sin entenderlo realmente (*pseudo-explicaciones*), frente a explicaciones sólidas, que realmente den cuenta de un fenómeno.

- Producir materiales de difusión de varios tipos

La apertura a la crítica de los pares es clave para el avance de la ciencia, por lo que la difusión de resultados en canales de tipo académico es necesaria. Para otras audiencias, como las de maestros, autoridades educativas o público en general es necesario otro tipo de productos.

A partir de lo anterior, y de que los tres elementos clave de toda investigación, para tratar de dar respuesta a una pregunta investigable —diseño, medición y análisis— los capítulos de la obra se organizan de la siguiente manera.

El Capítulo 1, *Construcción del objeto de estudio*, se refiere a las habilidades para iniciar una investigación, las que se concretan en la formulación de una pregunta investigable, para lo que es necesario presentar las nociones de variables, preguntas e hipótesis, y lo relativo a la revisión de literatura para obtener información teórica.

El Capítulo 2, *Diseños de investigación*, parte de que, además de centenares de procedimientos particulares para obtener información empírica o analizarla (técnicas), hay un número reducido de grandes tipos o formas de organizar un estudio (diseños), que se distinguen por combinar en forma especial algunas de las muchas técnicas para responder cierto tipo de preguntas. En el capítulo se presentan 14 diseños, a sabiendas de que un investigador no podrá dominarlos todos, pero sí deberá conocer las características de todos, para que pueda seleccionar el apropiado a fin de responder sus preguntas, sin reducir su formación al conocimiento de alguno de esos diseños, como la encuesta o el estudio de casos.

El Capítulo 3, *Obtención de información empírica*, tras discutir las nociones de observación y medición, presenta tres grandes familias de procedimientos, basados respectivamente en interrogación, observación y análisis de evidencias, terminando con una discusión sobre la forma de cuidar la calidad de la información obtenida.

El Capítulo 4, *Análisis de la información*, presenta primero los fundamentos de las técnicas que luego se describen, con bastante detalle en lo que se refiere a técnicas básicas, y en forma muy breve en lo que respecta a las más avanzadas, poniendo el énfasis en la comprensión de los conceptos y la lógica subyacente.

La *Conclusión* inicia con reflexiones sobre los problemas que enfrenta la formación de investigadores educativos y la posibilidad de hacerlo, seguidas por consideraciones sobre los aspectos cognitivos de esa formación —en particular epistemológicos— así como los aspectos valorales y actitudinales, y por último los de carácter afectivo.

Para terminar esta introducción, algunas precisiones:

Sobre lo que no pretende la obra. Es claro que no bastará para alcanzar un dominio pleno de los diseños y las técnicas que se presentan, para lo que será indispensable estudiar textos especializados, revisar trabajos en los que se utilicen, analizando la manera de hacerlo, y practicar su aplicación pasando de casos simples a los más complejos. Normalmente esto deberá hacerse con la guía de un investigador con experiencia en el uso de los diseños y las técnicas de que se trate.

Sobre lo que sí pretende la obra. Por una parte, dar una visión de conjunto del vasto campo de la metodología de la investigación educativa y social, que no proporcionan obras que exploran con mayor detalle alguna de sus muchas regiones. Por otra parte, ayudar a los lectores a evitar errores importantes, por no tener una buena comprensión conceptual de nociones fundamentales; es frecuente, en efecto, que por la facilidad que ofrece hoy la tecnología, investigadores con débil formación metodológica utilicen técnicas muy complejas sin entender bien lo que están haciendo, lo que fácilmente lleva a interpretaciones erróneas de los datos que analizan.

Sobre la secuencia de capítulos y la forma de trabajarlos. El lector podrá tener dificultad para entender unas partes del texto, que aclaran otros pasajes ubicados en otros capítulos. La organización del contenido es, sin duda, una entre muchas posibles. Lo que no es posible es incluir en cada punto todos los elementos relativos al tema de que se trate, y cualquier intento en este sentido haría la obra más larga, repetitiva y no más clara, sino todo lo contrario. Se trata de un caso del *círculo hermenéutico*: la comprensión (*Verstehen*) del todo supone la de cada una de sus partes, pero la comprensión de cada parte supone, a su vez, la del todo. Por ello una buena comprensión de la obra requerirá de más de una lectura. Esto dificulta la tarea del lector, sobre todo del que aborde la obra por sí solo, sin la ayuda de un buen instructor. El ir y venir entre los capítulos será inevitable, si bien quiero pensar que podrá ser también

enriquecedor. En mi defensa me remito a lo que dice un gran investigador, que cito en seguida, en el Recuadro I.4.

Sobre el lenguaje utilizado. De acuerdo con la Real Academia Española (RAE), considero correcto el uso genérico del masculino como término no marcado y que, por la importancia que tiene el principio de economía lingüística, el desdoblamiento de los géneros masculino y femenino solo se justifica cuando no usarlo traería consecuencias discriminatorias. Por eso haré uso de dicho desdoblamiento solo excepcionalmente, manifestando expresamente que no considero que eso reduzca en lo más mínimo mi compromiso con la igualdad jurídica y social de varones y mujeres, y mi rechazo contundente de toda discriminación.

RECUADRO I. 4. UNA OPINIÓN SOBRE LA FORMA DE APRENDER

[…] actualmente es difícil encontrar un libro que pueda introducirte efectivamente en un tema. La mayoría de la gente solo aprende haciendo cursos. Tienen que ir a clase, tienen que ir a un curso de verano. Ahora la idea de aprender por uno mismo está ausente. Cuando yo quería aprender sobre un tema nuevo, como informática, debía leerme el libro hasta que lo entendía. El autor pensó que él tenía la forma correcta de enseñarme ese tema. Pero yo tuve que empezar leyendo el capítulo 7, y entonces la primera mitad del capítulo 1, y después el capítulo 10, etcétera. De esa manera yo era capaz de construir mi propia ruta a través del nuevo tema. Pienso que esto es muy importante.

SYDNEY BRENNER, PREMIO NOBEL 2002

Referencias

AERA (2006). Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher*, Vol. 35(6): 33-40.

AERA (2008). *Definition of Scientifically Based Research*. Consultada el 25/09/2017, en: http://www.aera.net/About-AERA/Key-Programs/Education-Research-Research-Policy/AERA-Offers-Definition-of-Scientifically-Based-Res.

Brenner, S. (2006). *Mi vida en la ciencia. Las aportaciones de un biólogo excepcional*. (Autobiografía narrada a Lewis Wolpert. Editores Erol C. Friedberg y Eleanor Lawrence). Valencia: Publicacions Universitat de Valencia. (Edición original en inglés, The Science Archive, 2001).

Condliffe Lagemann, E. (2000). *An Elusive Science. The Troubling History of Education Research*. Chicago: The University of Chicago Press.

Cronbach, L. J. y Suppes, P. (Eds.). (1969). *Research for Tomorrow's Schools. Disciplined Inquiry for Education*. Report of the Committee on Educational Research, National Academy of Education. London: Macmillan.

De Lansheere, G. (1986). *La recherche en éducation dans le monde*. París: Presses Universitaires de France. Traducción al español *La investigación educativa en el mundo. Con un capítulo especial sobre México*. México: Fondo de Cultura Económica, 1996.

Feyerabend, P. (1970). *Against Method. Outline of an Anarchistic Theory of Knowledge*. Minneapolis: University of Minnesota. Traducción al español, *Contra el método. Esquema de una teoría anarquista del conocimiento*. Barcelona: Ariel: 1974.

Gage, N. L. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Researcher* 18(7): 4-10.

Kelly, G. J. (2006). Epistemology and Educational Research. En Green, Judith L., G. Camilli y P. B. Elmore (Eds.). *Handbook of Complementary Methods in Education Research* [pp. 33-55]. New York, Routledge.

King, G., Keohane, R. O. y Verba, S. (1994). *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton: Princeton University Press. Traducción al español *El diseño de la investigación social. La inferencia científica en los estudios cualitativos*. Madrid: Alianza, 2000.

López Ruiz, M., Sañudo Guerra, L. y Maggi Yáñez, R. E. (2013). *Investigaciones sobre la investigación educativa 2002-2011*. México: COMIE-ANUIES.

Martínez Rizo, F. (1997). *El oficio del investigador educativo*. Aguascalientes: Universidad Autónoma de Aguascalientes, 2ª edición.

Pedhazur, E. y Schmelkin, L. (1991). *Measurement, Designa and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., y Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston-New York: Houghton Mifflin Co.

Shavelson, R. J. y L. Towne (Eds.). (2002). *Scientific Research in Education*. Washington. National Research Council. National Academy Press.

U. S. Congress (2001). *No Child Left Behind Education Law*. Pub. L. N° 107-110, 115 Stat. 1457. Washington: Author.

Vogt, W. P. (2007). *Quantitative Research Methods for Professionals*. Boston: Pearson Education.

CAPITULO 1
CONSTRUCCIÓN DEL OBJETO DE ESTUDIO

CONTENIDO

Introducción

Antes de que se puedan poner en juego los elementos básicos de la investigación —diseño, medición y análisis— es necesario haber precisado *qué* se va a investigar; solo después se podrá pensar en *cómo* se hará. Por ello el primer apartado de este capítulo se refiere a lo que suele llamarse *la construcción del objeto de estudio*, y comprende las primeras habilidades que debe dominar un investigador, relativas al inicio del proceso de investigación.

Este primer grupo de habilidades comprende identificar fenómenos susceptibles de indagación a partir de *preguntas investigables* y aspectos de ellos, o variables; formular preguntas que sean susceptibles de ser respondidas empíricamente; y obtener información de investigaciones previas. Esta última remite a la revisión de literatura, e incluye las habilidades particulares de identificar fuentes de información sobre el tema, y la de distinguir fuentes de distinta calidad.

El objeto de estudio en general y su acotamiento

La primera manera de expresar el objeto que se quiere estudiar en una investigación puede consistir en la formulación de un *tema* de interés, en términos más o menos generales. Otras expresiones de un mismo objeto de estudio pueden tener la forma de un *problema* a resolver, de una *pregunta*, o de un *objetivo* o propósito.

Las frases siguientes son ejemplos de maneras equivalentes para formular el objeto de estudio de una posible investigación, en una primera aproximación:

- Tema muy general: La calidad de la educación mexicana.
- Tema algo más preciso: Nivel de calidad de las primarias de México.

- Problema: El nivel de calidad de las primarias mexicanas es preocupante.
- Pregunta: ¿Cuál es el nivel de calidad de las primarias mexicanas?
- Objetivo: Identificar el nivel de calidad de las primarias mexicanas.

Estas formulaciones generales no bastan para definir suficientemente el objeto de estudio de una investigación. Para ello es necesario precisar o acotar el tema, lo que debe hacerse de dos maneras. Para presentarlas daremos antes un rodeo, con el objetivo de observar cómo se delimita el objeto de estudio de algunas disciplinas.

Para distinguir los tipos de seres que se encontraban en la naturaleza, desde Aristóteles hasta Linneo, se les clasificaba distinguiendo tres grandes reinos: uno de seres inertes, el reino mineral, y dos de seres vivos, el vegetal y el animal. Para estudiar cada uno de estos reinos había otras tantas disciplinas, la mineralogía, la botánica y la zoología. Los avances del estudio de la naturaleza han llevado a que, solo en lo relativo a seres vivos, y sin entrar en las discusiones entre especialistas, se distingan los reinos de animales, plantas, hongos, protozoarios, cromistas, bacterias y arqueas.

Según la clasificación tradicional, en el reino animal se distinguían vertebrados e invertebrados, y entre los primeros, los mamíferos, las aves, los reptiles, los anfibios o batracios y los peces. Algunas disciplinas tradicionales corresponden a algunos de estos grupos: la mastozoología es el estudio de los mamíferos; la ornitología, de las aves; la herpetología, de los reptiles y los batracios; y la ictiología, de los peces.

Hay disciplinas que estudian animales ubicados en niveles taxonómicos inferiores, que atrajeron el interés de estudiosos que pusieron un nombre a lo que hacían: la primatología estudia los primates, la entomología los insectos y la nematología los nemátodos, los gusanos redondos (nematelmintos), sin que el estudio de los planos (platelmintos) haya conseguido un rango similar.

Lo que tienen en común estas disciplinas es que definen su objeto de estudio según un conjunto empírico, un grupo de animales de ciertas características. Sin embargo, siguiendo esta lógica se llega a resultados que muestran que no es el mejor camino.

El grupo de los primates incluye cientos de especies, desde lémures hasta gibones, orangutanes, gorilas, chimpancés y bonobos, y los humanos. Posiblemente no haya muchas objeciones a definir la antropología como la disciplina que estudia a estos últimos, pero ¿por qué no lemurología, golirología o bonobología? Las aves incluyen águilas y cóndores, zopilotes, avestruces y pingüinos, cenzontles, canarios y periquitos, ¿por qué no aguilología, zopilotología, periquitología?

Si el objeto de una disciplina se define con base en el conjunto empírico de los seres que estudia, entonces deberíamos terminar con una ciencia de cada animal, de cada planta, de cada hongo y de cada bacteria.

Otras disciplinas, en cambio, definen su objeto no con base en un conjunto empírico, sino en un aspecto de la realidad de muchos objetos empíricos.

La morfología y la fisiología estudian la estructura y procesos de cualquier animal, o en general de cualquier ser vivo. La definición de la física y la química se hace también con base en cierto tipo de propiedades o fenómenos que pueden darse en cualquier ser, inerte o vivo. Se pueden estudiar los aspectos físicos o químicos de un ser humano, de cualquier otro animal, de una planta, o de una piedra.

La distinción fundamental es la de objetos empíricos y objetos teóricos o construidos teóricamente. En una terminología tradicional, objetos materiales y formales.

Después de este rodeo, apliquemos la distinción a temas educativos. En este campo, *los objetos empíricos* a estudiar pueden ser alumnos, maestros, planteles, el sistema educativo como un todo, el currículo, los libros de texto, entre otros. *Los objetos construidos teóricamente*, en cambio, son los aspectos de diversos objetos empíricos susceptibles de estudiarse, por ejemplo el nivel de aprendizaje de un grupo de alumnos, las prácticas de evaluación de ciertos maestros, el clima o el ambiente que prevalece en algunos planteles escolares; la eficiencia del sistema educativo y procesos como la transición de unos grados a otros o la deserción; relaciones entre aspectos o posibles explicaciones de procesos, como causas de deserción, o factores que inciden en el aprendizaje y en el clima escolar.

La construcción del objeto de estudio, primera etapa de toda investigación es, pues, el paso de un tema muy general a un planteamiento más preciso de lo que se quiere estudiar. Este proceso de acotamiento incluye dos tipos de delimitación:

- Una empírica (más fácil): preciar qué sujetos u objetos se estudiarán, de qué características, cuántos, dónde, cuánd. Por ejemplo, estudiantes de cuarto grado de primaria de México, en el ciclo escolar 2017-2018.
- Una delimitación teórica (más difícil): qué aspectos de esos sujetos u objetos empíricos se desea estudiar, qué procesos, qué relaciones entre aspectos, qué tipo de explicación de procesos. Por ejemplo, los factores del hogar de esos alumnos, y las prácticas docentes de sus maestros, que inciden en el nivel de aprendizaje que alcanzan en matemáticas.

Las variables y la operacionalización

La distinción entre objetos empíricos y construidos alude a otro concepto importante en investigación: el de *variable*.

En una primera aproximación, variable es *algo* que varía, que toma diversos valores, aumenta o disminuye; es lo opuesto a algo que no varía, a una constante.

Las obras de metodología suelen distinguir tipos de variables según varios criterios: cualitativas (categóricas, dicotómicas, politómicas) y cuantitativas (discretas o continuas...); referidas a individuos (sexo, edad, estatura, cociente intelectual...) o a colectivos (religión mayoritaria, preferencia electoral, promedio de ingreso); independientes, dependientes, intervinientes o extrañas.

En línea con lo apuntado en el inciso anterior, se propone una definición alternativa de variable: *un rasgo o aspecto particular de los sujetos u objetos empíricos que se estudiarán, que puede tomar distintos valores, e interesa por sí mismo y/o por su relación con otros aspectos.*

Antes de identificar diferentes tipos de variables parece importante precisar qué es ese *algo* que varía, ya que no se trata de un objeto empírico que existe o está dado en sí mismo, sino de un objeto construido mentalmente por nosotros.

La definición propuesta supone una manera de entender la forma de conocer de los seres humanos, que no permite captar la totalidad de la realidad, y ni siquiera la totalidad de cualquier objeto empírico, sino que inevitablemente es parcial y selectiva: centra la atención en unos aspectos y los capta, al tiempo que desatiende otros. En la realidad cualquier aspecto existe junto a todos los demás del objeto empírico de que se trate, está ligado a ellos (*todo está relacionado con todo*), pero es posible distinguirlo de los demás analíticamente, *abstraerlo*; de hecho, esto es inevitable, pero se puede hacer de manera irreflexiva o reflexiva.

Si el objeto de un estudio sobre calidad educativa se delimita empíricamente ciñéndolo a las escuelas de cierta región, o sus alumnos, es necesario acotarlo también teóricamente, precisando el aspecto que interesa, por ejemplo, la eficiencia terminal de las escuelas, o el aprendizaje de las matemáticas de los alumnos.

No sobra añadir que ningún aspecto de la realidad tiene en sí mismo y siempre el carácter de variable. Un mismo elemento puede ser tratado como variable en una investigación y como constante en otra, por ejemplo, si en un estudio se decide que solo se incluya a las alumnas, o únicamente a escuelas públicas.

Las variables (o aspectos de la realidad) que interesa estudiar se pueden definir de manera suficientemente concreta para poder observarlos con cierta facilidad, o de manera

tan abstracta que no es posible captarlas sin antes traducirlas a versiones más observables. A esto se refiere el importante concepto de *operacionalización*.

RECUADRO 1.1. EJEMPLOS DE OPERACIONALIZACIÓN

Con ocasión de la primera visita a México del papa Juan Pablo II, un chico oyó decir a otro: *el papa quiere mucho a los niños*. No muy seguro de lo que tal cosa quería decir en concreto, preguntó: *¿da balones de fut?* La noción de *amor por los niños* es muy abstracta; el número de balones de fútbol que alguien regala es más fácil de apreciar con precisión: al menos en opinión del niño que hizo la pregunta, una persona que regale muchos balones de fútbol tendrá más amor por los niños que otra que no lo haga.

Ejemplos del campo de investigación, en varias áreas de las ciencias sociales y en educación:

- ***Producto Interno Bruto* (PIB), global o por habitante (*per capita*).** Este es un dato muy utilizado como indicador del desarrollo de un país, pero también uno cuyas limitaciones son claras, ya que solo toma en cuenta una dimensión del complejo concepto de desarrollo, la económica, y únicamente en la forma más simple, total o agregada, sin distinguir por ejemplo la forma en que se distribuyen los bienes económicos, y muchos otros aspectos.
- ***Índice de Desarrollo Humano* (IDH) de la ONU.** En un esfuerzo por contar con un indicador mejor, que capte al menos algunas de las muchas dimensiones del *desarrollo de un país*, la ONU construyó el *Índice de Desarrollo Humano*, compuesto por tres indicadores: uno sobre la dimensión económica del concepto (PIB per capita), otro sobre la dimensión de salud (el promedio o *esperanza* de vida de la población), y uno más sobre la dimensión educativa (el porcentaje de alfabetización y el de la población con primaria).
- **Mediciones de pobreza del CONEVAL.** En México son conocidos los trabajos para operacionalizar y medir la pobreza, con las dimensiones de pobreza alimentaria, pobreza de capacidades y pobreza patrimonial, con varios indicadores de tan complejo fenómeno: rezago educativo, acceso a servicios de salud, a seguridad social, a servicios básicos de vivienda y a la alimentación, así como el grado de cohesión social (CONEVAL, 2009).
- **Clima escolar.** Para hacer observable esta noción, un trabajo identificó 4 dimensiones, cada una con algunas subdimensiones o índices, y sus respectivos indicadores, asociados a los ítems de un cuestionario: *organización escolar* (control de conducta y uso efectivo del tiempo); *convivencia entre profesores y estudiantes* (los alumnos se llevan bien con los docentes, estos les brindan apoyo); *clima positivo* (según percepción de los alumnos); y *clima negativo* (frecuencia de molestias, agresiones, robos…). (Treviño, Place y Gempp, 2013)

Los anteriores son ejemplos de *operacionalización*, proceso por el que se identifican manifestaciones más fáciles de captar de una variable que no lo es directamente, llamada también variable latente o constructo.

En la tradición metodológica de Lazarsfeld (1973) a esas nociones más fácilmente apreciables se les designa con el término de *indicadores*, en tanto que se reserva el de *variables* para denotar las nociones más abstractas y difíciles de captar.

Como muchas variables que interesa explorar en la investigación social y educativa se refieren a aspectos de la realidad que no resulta sencillo captar directamente, la operacionalización constituye una etapa fundamental en ese tipo de pesquisas, que antecede a la fase de construcción de los instrumentos con los que se obtendrá la información necesaria para cualquier estudio. Para desarrollar tales instrumentos es indispensable, en efecto, pasar de los conceptos más abstractos a los indicadores más concretos, después de identificar las dimensiones de los primeros y, en su caso, antes de integrar algunos indicadores en índices.

Siguiendo a Vogt *et al.* (2012: 325-329) se pueden distinguir cuatro tipos de indicadores:

- Componentes del concepto: aspectos particulares de una dimensión o subdimensión; la existencia de votaciones libres es un componente de la noción de democracia, como la libertad de prensa o la división de poderes.
- Síntomas o efectos de lo que designa el concepto: puede ser más fácil verificar la presencia de un efecto, y de ella inferir la de la causa. Durkheim utilizó cambios en las leyes para rastrear cambios en los valores morales de una sociedad.
- Causas del concepto: la irracionalidad como indicador de prejuicio. Si se quiere estudiar el prejuicio, definido como *creencia negativa irracional sobre algo o alguien, pese a las evidencias en contrario*, puede ser más factible estudiar manifestaciones empíricas de irracionalidad y de allí inferir la presencia de prejuicios, causados por la irracionalidad.
- Aproximaciones: si se trata de un concepto que denota una realidad muy amplia y multidimensional (*v.gr.* la calidad de los servicios médicos de un país), es posible utilizar un concepto bastante simple como aproximación (indicador aproximado o *proxy*), por ejemplo, la tasa de mortalidad infantil.

Preguntas precisas e hipótesis

Al inicio de este capítulo se dijo que una primera expresión del objeto de estudio de una investigación podía tener la forma de un tema, un problema a resolver, un objetivo o una pregunta. En este último caso se pensaba en una pregunta general, como la que se ofreció como ejemplo: ¿cuál es el nivel de calidad de las primarias mexicanas?

Consideremos nuevamente algunos enunciados que presentan de varias maneras el objeto de estudio de una posible investigación:

- Factores que inciden en niveles de aprendizaje en las primarias mexicanas.
- Los niveles de aprendizaje en las primarias mexicanas son preocupantes.
- Mejorar los niveles de aprendizaje en las primarias mexicanas.
- ¿Por qué son bajos los niveles de aprendizaje en las primarias mexicanas?
- La pobreza del hogar incide en los niveles de aprendizaje en las primarias.

La primera de estas formulaciones es un tema de investigación; la segunda, un juicio —razonable o no— que refleja un problema; la tercera, un propósito que podemos o no considerar loable; la cuarta, una pregunta que coincide con el tema, pero que no apunta a ninguna pista para responderla; la quinta formulación, por su parte, es una aseveración que tiene una posible explicación del fenómeno al que se refieren el tema, el problema, el objetivo y la pregunta. En la jerga de la metodología de la investigación se dice que esa formulación es una *hipótesis*.

Hay posturas opuestas en relación con las hipótesis. Para unos su formulación sería un paso indispensable en el proceso de investigación; para otros sólo serían útiles en los estudios llamados cuantitativos, mientras que no lo serían en los cualitativos; y para algunos serían incluso un elemento que debería descartarse por completo.

La postura que sostengo es que, si se identifica el término con hipótesis *estadística*, que implica estimar la probabilidad de que la información recabada haga descartar o no la *hipótesis nula* con una prueba de significación, entonces, desde luego, no toda investigación tiene por qué incluir este elemento.

Pero hay otra manera de entender la noción: un concepto de hipótesis según el cual deben verse como *pistas de búsqueda* para orientar la obtención de información, a partir de lo que ya se sabe sobre el tema. Como se verá con mayor amplitud después, en el sentido de puente entre teoría y experiencia, las hipótesis son un elemento importante en una investigación, evitando comenzar cada vez de cero, como si nadie hubiera explorado antes el terreno.

La desconfianza ante las hipótesis se ve alimentada por las visiones estrechas que consideran que su redacción siempre debe tener una forma que supone un estudio de tipo experimental, o del tipo *a más de tal cosa más de tal otra*. Unos ejemplos permitirán entender que lo importante en una hipótesis no es la forma, sino el fondo.

Supongamos que se trata de un trabajo que busca entender a qué se deben los niveles de aprendizaje que se encuentran en las escuelas primarias de una región, que se considera son malos.

Expresado en forma interrogativa (¿a qué se deben...?), el objeto de estudio del trabajo es inteligible, pero no distinto de un tema general como *Factores que inciden en los niveles de aprendizaje;* en ninguno de los dos casos se nos da una pista que nos oriente sobre la información que debemos recoger. Si se pasa directamente al trabajo de campo sin precisar la cuestión, el resultado sería seguramente pobre.

La revisión más somera de la literatura sobre el tema mostrará que los factores que inciden en el aprendizaje en las escuelas se suelen clasificar en dos grupos: los factores del hogar y, en general, el entorno extraescolar de los alumnos; y los factores escolares, que incluyen los de la escuela y del aula. Una revisión más amplia permitirá conocer las opiniones sobre el peso relativo de unos factores y otros, la complejidad de ambos, sus intricadas relaciones, entre otras cosas.

Veamos ahora tres formulaciones que pueden aspirar a llamarse hipótesis, para orientar el trabajo de un imaginario investigador:

- Las condiciones del hogar son los factores que influyen más en los niveles de aprendizaje en las primarias mexicanas.
- Las condiciones de la escuela y el aula son los factores que influyen más en los niveles de aprendizaje en las primarias mexicanas.
- ¿Qué tan importantes serán las condiciones del hogar, comparadas con las de la escuela y el aula, como factores que inciden en los niveles de aprendizaje en las primarias mexicanas?

Para algunas personas las dos primeras fórmulas podrían considerarse hipótesis, pero la tercera no, ya que es una pregunta. Así es, pero a diferencia de la pregunta inicial (¿a qué se deben los niveles de aprendizaje...?) en este caso sí da una pista para responderla, ya que presenta dos tipos de factores que podrían incidir en los niveles de aprendizaje, pero de los que no sabe en qué medida lo hará cada uno.

De hecho, las formulaciones son equivalentes, si lo que se espera de todas es que orienten la subsecuente recolección de información.

Contra lo que podría pensarse a primera vista, suponiendo que se acepte que todos los factores que inciden en el aprendizaje son extraescolares o escolares, las dos primeras formulaciones no son hipótesis diferentes, sino la misma, ya que las dos lleva-

rían a recolectar la misma información. No se trata de alternativas entre las que habría que escoger, apostando a una u otra, y esperando a ver el resultado del estudio para ver quién acertó. Las dos primeras fórmulas orientan la recolección de información en la misma dirección, exactamente igual que la tercera.

Teniendo en cuenta que los fenómenos sociales y educativos siempre tienen una gran complejidad e involucran numerosos aspectos (variables), con una intricada red de relaciones entre ellos, si las hipótesis debieran formularse siempre en forma de aseveraciones simples, que postulen cierta relación entre variables de dos en dos, cualquier estudio medianamente complejo debería tener decenas de hipótesis de ese tipo, que además siempre serían insuficientes para captar la complejidad del fenómeno. Por ello es razonable preferir un buen conjunto de preguntas, que recoja de la mejor manera posible lo que se sepa de trabajos anteriores, gracias a una buena revisión de literatura, de lo que tratará el siguiente inciso.

Pero no hay que perder de vista una ventaja de la forma de redactar las hipótesis como oraciones aseverativas: una pregunta que incluye una pista de respuesta, como *¿qué tan importantes serán los factores extraescolares, comparados con los escolares, para explicar los niveles de aprendizaje…?*, se transforma fácilmente en una aseveración, del tipo *(yo creo que) los factores extraescolares son más importantes que los escolares, para explicar niveles de aprendizaje.* En cambio, la pregunta *¿a qué se deben los niveles de aprendizaje…?* puede sonar interesante, aunque no dé pista alguna para responderla; si se expresa en forma aseverativa, quedaría como *(yo creo que) los niveles de aprendizaje se deben a algo…*, lo que deja clara su insuficiencia para orientar el trabajo de recolección de información.

Esto lleva a preguntarse cómo hacer buenas hipótesis. Para responderla importa distinguir dos sentidos del adjetivo *buenas*: hipótesis *correctas* e hipótesis *fecundas*.

Es posible aprender a hacer hipótesis correctas, formulaciones que efectivamente sean pistas de búsqueda para el trabajo de obtención de información y no otra cosa, temas o preguntas generales, preocupaciones razonables, o propósitos loables. En cambio, no hay receta para hacer hipótesis fecundas, que lleven a hallazgos importantes: sólo la solidez de la formación teórica, el conocimiento del campo y la experiencia del investigador pueden producirlas.

La expresión *formación teórica*, por otra parte, no debe entenderse como limitada al dominio de las teorías más generales y abstractas de un campo, como conductismo o teorías cognitivas en psicología, funcionalismo o interaccionismo simbólico en sociología. En esta obra se entiende como el conjunto de conocimientos sobre un tema particular, aunque no alcancen el nivel de una gran teoría, pero integran lo que se sabe

sobre el tema concreto de que se trate, de manera similar a lo que plantea la noción de *teorías de alcance medio* de Merton, y la de *grounded theory*.

Esas microteorías recogen los conocimientos sobre un tema que se han acumulado gracias al esfuerzo de los investigadores que lo han estudiado previamente, así como de la experiencia de quienes trabajan en el campo (en cuestiones educativas, ante todo los maestros), cuyo saber práctico debe aprovecharse. Debe reconocerse, desde luego, que esos conocimientos previos son de desigual calidad, pueden estar mal integrados e incluso contradecirse.

De nuevo un rasgo fundamental de la investigación científica es que no parte de cero, sino que toma en cuenta los hallazgos previos para enriquecerlos, ampliarlos, acotarlos o cuestionarlos. Por ello es fundamental la revisión de la literatura, sin olvidar que una parte importante del conocimiento previo no está plasmado en artículos de revistas, sino en la experiencia de los que trabajan en el campo, que también habrá que tratar de recuperar, de otras maneras.

La revisión de la literatura

La elaboración del *marco teórico*, como reza la expresión consagrada de lo que el investigador debe hacer una vez seleccionado y acotado el tema, no debe referirse a las reflexiones más abstractas del campo de que se trate, sino simplemente a lo que se sabe sobre el mismo, para no perder tiempo recorriendo senderos trillados o metiéndose en callejones que ya han demostrado no tener salida.

> La revisión de literatura es importante en varios sentidos. Es esencial tanto si se trata de dar ideas para transformar un tema general de investigación en preguntas de investigación más específicas, como de evitar que trates de estudiar una cuestión que ya ha sido ampliamente resuelta, o para alejarte de caminos metodológicos equivocados. La revisión de literatura impide que trates de reinventar la rueda y, tan importante como eso, que trates de reinventar la rueda ponchada. (Vogt, Gardner y Haeffele, 2012: 9)

Es importante conocer lo que ya se ha encontrado sobre un tema y partir de allí para llegar más lejos. Para ello lo primero es localizar dónde se encuentra la información. ¿Cómo lograrlo, si hay miles de investigadores que producen decenas de miles de artículos de revista, libros, documentos de circulación restringida y demás?

Contra lo que ocurría hace tres décadas, cuando el primer problema que enfrentaba el investigador era que no encontraba casi nada sobre su tema en las precarias biblio-

tecas institucionales y sin Internet, hoy debe enfrentar el problema opuesto: gracias a las nuevas tecnologías, desde cualquier lugar puede acceder al acervo de las mejores bibliotecas y a inmensas bases de información bibliográfica. Sin una buena estrategia de búsqueda, el investigador se encontrará con una masa enorme de material de desigual calidad, cuyo procesamiento será difícil y poco productivo.

La estrategia que se propone parte de una idea básica: para ser productiva, la búsqueda bibliográfica debe seguir un orden *inverso* al seguido para la producción.

Los primeros escritos derivados de una investigación son los reportes internos que se presentan a las instancias patrocinadoras del trabajo. En seguida se presenta una o más ponencias en congresos especializados, los mejores de los cuales suponen *arbitraje* (*peer review* o revisión por pares), de preferencia sin conocer la identidad del autor (ciego). El autor puede así recibir retroalimentación sobre su trabajo.

Con ello puede mejorarse el trabajo y generar el tercer tipo de producto, el artículo, que se enviará a una revista especializada. Las buenas revistas utilizan el sistema de revisión ciega por pares, por lo que la aceptación del texto por una revista de prestigio es señal de que es fruto de un trabajo de buen nivel.

Cuando un investigador trabaja por un tiempo suficiente (años) en proyectos de una misma línea, y su trabajo va alcanzando la calidad que da la experiencia, y el prestigio que da el reconocimiento de los colegas, puede generar el cuarto tipo de producto: el libro especializado o monográfico sobre su tema.

El quinto producto de una investigación es un capítulo en una obra de referencia especializada. Este tipo de obras incluye enciclopedias, diccionarios, manuales, tratados y similares: obras de gran tamaño, en cuyos artículos se concentran los conocimientos fundamentales sobre un campo. Por su naturaleza a estas obras suelen ser colectivas, y los autores de capítulos son seleccionados por invitación entre los estudiosos más reconocidos. Otro producto final puede ser también un documento de revisión de la literatura sobre un tema, que puede ser publicado en forma independiente, en revistas de tipo general, o en las especializadas como la *Review of Research in Education* o la *Review of Educational Research*.

RECUADRO 1.2. SECUENCIAS DE LA PRODUCCIÓN Y DE LA BÚSQUEDA

La secuencia de la producción científica:

- Ponencias en congresos

- Artículos en revistas
- Reportes internos
- Libros monográficos
- Obras de referencia y artículos de revisión de literatura

La secuencia invertida de la búsqueda bibliográfica:
- Obras de referencia y artículos de revisión de literatura
- Libros monográficos
- Artículos Reportes internos
- en revistas
- Ponencias en congresos

El primer lugar al que hay que acudir para buscar información no son las revistas ni los libros ordinarios —*monográficos*, porque su contenido centra la atención en un solo tema— sino un tipo particular de libros, que son las obras de referencia.

Las revistas contienen el material más actualizado sobre un tema, pero también el más desigual y perecedero. Se consultan para obtener información *muy reciente*, del año en curso, o de pocos años de antigüedad. Los libros monográficos reúnen material menos actual, pero más decantado y, por ello, más resistente al tiempo, cubriendo probablemente un lapso de 5 a 10 años atrás. Los artículos particulares de las obras de referencia especializadas, en cambio, concentran los elementos fundamentales de un tema, aunque no los más recientes. Por eso hay que comenzar la búsqueda con ellas, para obtener visiones panorámicas sobre el tema de interés que fueron elaboradas por una autoridad en la materia, por lo que serán nuestra mejor introducción al tema.

Las obras de referencia —enciclopedias, *handbooks* o manuales y similares— suelen ser de gran formato y tener muchas páginas, e incluso varios volúmenes. También son poco numerosas, pero es raro el tema importante sobre el que no hay alguna (como los *handbook* sobre currículo, sobre investigaciones relativas a enseñanza en general, o a la enseñanza de matemáticas o ciencias), además de las de cobertura más amplia, como la *International Encyclopedia of Education*, o la *Encyclopedia of Educational Research*, en sus sucesivas ediciones.

Los capítulos de las obras de referencia, firmados por una autoridad en el tema, son síntesis breves del mismo, e incluyen bibliografía con obras importantes al respecto.

Por ello, estos artículos son una valiosa fuente de información sobre tres elementos que permitirán orientar de manera muy precisa los siguientes pasos de la búsqueda: *descriptores*, palabras clave o *key words* relativos a conceptos importantes; *autores*,

nombres de los investigadores que han hecho aportaciones básicas al campo; y *títulos* de libros monográficos y de publicaciones periódicas importantes.

El otro tipo de producto muy adecuado para iniciar una búsqueda bibliográfica es el formado por los artículos de revisión de la literatura sobre un tema, también designados como estados del conocimiento. Esta última expresión comenzó a usarse en México para designar las revisiones de literatura promovidas por el Consejo Mexicano de Investigación Educativa, y es una adaptación de la expresión *state of the art*, que se solía traducir, erróneamente, como *estado del arte*.

En realidad, *state of the art* no es una expresión de carácter sustantivo sino adjetivo; se trata de un modificador del sustantivo *review*, que designa simplemente lo más reciente del campo que designe el sustantivo al que modifica. Así, *state of the art lap top, o state of the art technology*, significa simplemente computadora portátil de última generación, o tecnología de última generación. Por su parte, *review* no debe traducirse como revista (*journal*) sino como revisión. La expresión completa en inglés que se traduce como *estado del conocimiento* es *state of the art literature review*, revisión de la literatura más reciente sobre cierto tema.

Por lo anterior, las revistas antes citadas que se dedican a publicar revisiones de literatura, ambas editadas por la *American Educational Research Association*, llevan el término *review* en sus respectivos títulos: *Review of Research in Education* (que aparece anualmente desde 1973) y *Review of Eductional Research*, de aparición trimestral desde 1931. Encontrar en ellas un artículo sobre el tema de investigación que interesa, es una excelente forma de iniciar una búsqueda bibliográfica.

Los DAT —Descriptores, Autores y Títulos— encontrados en capítulos de obras de referencia, o en artículos de publicaciones especializadas en revisiones de literatura son pistas que facilitarán la búsqueda en catálogos de biblioteca y bases en línea de bibliografía, evitando que el investigador se pierda en los innumerables senderos del territorio que comienza a explorar.

Con ello podrá pasar a la búsqueda de libros monográficos y artículos en revistas, utilizando nombres de autores y títulos de obras esenciales, así como descriptores de subtemas, conceptos y corrientes teóricas centrales, para explotar los catálogos de autor, título y tema de bibliotecas, fuentes bibliográficas secundarias, y bancos de información bibliográfica en línea.

Al presentar las habilidades para el inicio de una investigación se dijo que *revisar literatura* comprende dos habilidades particulares: *identificar fuentes de información* sobre el tema, y *distinguir fuentes de distinta calidad*. La secuencia recomendada para

la búsqueda bibliográfica se basa en la idea de que es más seguro que sean de gran calidad los artículos de una obra de referencia importante, firmados por los expertos más reconocidos de cada tema, que los de otras obras.

En este sentido es importante llamar la atención de los investigadores en formación sobre un efecto negativo que acompaña el crecimiento explosivo de publicaciones académicas: el incremento de publicaciones de baja calidad, que aprovechan la tendencia internacional a valorar la *productividad* de un investigador, para efectos de contratación y promoción, según el número de sus publicaciones.

Lo anterior incluye editoriales y revistas que publican libros o artículos con rapidez, pero sin arbitraje o dictaminación de la calidad del contenido, a cambio de un pago por parte del autor o autores. En el caso de las revistas, el nombre de la publicación se escoge cuidadosamente para dar una falsa impresión de seriedad, que engaña a personas no familiarizadas con un campo del conocimiento: *American Journal of..., European Journal of..., International Journal of...*

Las empresas que publican tales revistas no están basadas en instituciones serias de tipo académico, aunque pueden tener páginas web igualmente engañosas; los eventuales miembros de los consejos o comités editoriales no suelen ser personas que tengan amplio reconocimiento en las respectivas comunidades académicas, si bien en algunos casos puede tratarse de investigadores que aceptan de buena fe aparecer en esos cuerpos, sin que en realidad cumplan la tarea de cuidar la calidad que teóricamente les corresponde.

El número de este tipo de revistas está aumentando tanto que se ha acuñado para designarlas la expresión *revistas depredadoras* (*predatory journals*), y es motivo de preocupación en las instituciones académicas más serias, por las distorsiones que pueden provocar en los procesos de contratación y promoción de investigadores, con un lucro no despreciable para quienes las administran. La preocupación se ha manifestado incluso en medios periodísticos (Silió, 2019).

Hay sitios en Internet con listas de revistas sospechosas de ser depredadoras, así como de las empresas que las publican. Esas listas, desde luego, pueden cometer errores de Tipo I y de Tipo II, al incluir títulos serios, y al dejar de incluir títulos realmente depredadores, pero pueden ser un inicio a partir del cual no es difícil que quien esté interesado en publicar recabe información sobre la seriedad de un editor o un título. La condición de que se pague por la publicación de un texto debería ser ya un primer motivo de sospecha.

RECUADRO 1.3. SITIOS CON LISTAS DE REVISTAS SOSPECHOSAS

Sitios con listas: *Cabell's International* (https://www2.cabells.com/); ***Beall's List of Predatory Journals and Publishers*** (https://beallslist.weebly.com/standalone-journals.html); ***Scholarly Open Acces*** (https://scholarlyoa.com/individual-journals/). La Universidad de Barcelona ofrece la Matriz de ***Información para el Análisis de Revistas*** (http://miar.ub.edu/idioma/es).

La redacción del apartado de referentes teóricos

Localizados los materiales (libros, artículos, *papers*) con información relevante para el tema, para extraerla es necesario leerlos. En el sentido de no solo decodificar los símbolos básicos, las letras, sino en el de apropiarse realmente del contenido de un texto de extensión y profundidad mínimas, leer es una tarea compleja, que muchas personas no dominan, aunque hayan terminado la educación obligatoria. Un investigador debe tener la capacidad para hacerlo, pero no sobra recordar unas orientaciones básicas para la lectura, tomadas de una obra clásica al respecto.

Después de las reglas *para descubrir sobre qué trata un libro*, Adler y Van Doren (2000: 171) dan otras *para interpretar su contenido*, importantes, aunque parezcan obvias:

> Llegar a acuerdo con el autor interpretando las palabras clave; comprender las proposiciones más destacadas reflexionando sobre las oraciones más importantes; conocer los argumentos, en secuencias de oraciones o a partir de estas; determinar qué problemas ha resuelto el autor y cuáles no.

En seguida, los autores citados enuncian reglas *para criticar un libro,* que incluyen:

- Máximas de etiqueta intelectual: no decir que se está de acuerdo o no, ni suspender el juicio, hasta poder decir "lo comprendo"; no disentir por puro afán de polémica; reconocer la diferencia entre conocimiento y opinión personal aportando buenas razones para cualquier juicio crítico.
- Criterios especiales para puntos de crítica, mostrando dónde el autor está desinformado, dónde está mal informado, dónde es ilógico, y dónde es incompleto su análisis. (Adler y Van Doren, 12000:171)

La lectura que debe hacerse del material identificado en la búsqueda bibliográfica sobre el tema de interés se caracteriza por ser selectiva (es orientada por un interés preciso, plasmado en un esquema) y por dejar rastro, por producir un conjunto de *fichas* que recogen fragmentos de información para tener a mano en forma práctica y flexible en el momento de redactar. Adler y Van Doren hablan de lectura paralela.

El esquema de trabajo como guía para la lectura

Además de dar elementos (DAT) para orientar la búsqueda, comenzar la lectura con capítulos sobre el tema en obras de referencia especializadas permite elaborar un primer esquema de lectura, con los principales apartados del tema.

La ventaja de tener una guía para abordar la lectura de un material abundante es clara: con un buen esquema es posible leer selectivamente, sin perder tiempo en lo irrelevante, pero sin dejar pasar lo importante. Por otra parte, ningún esquema es perfecto, y uno malo será una mala guía. Pero ¿cómo tener un esquema adecuado sin haber leído todo, y cómo leer materiales abundantes y desiguales sin tener un esquema? Para escapar al círculo vicioso no hay que olvidar que un esquema debe estar abierto a las modificaciones que el avance en la lectura aconseje. Es mejor comenzar la lectura de materiales adicionales con un esquema sencillo, elaborado con base en las obras de referencia (y con el consejo de un buen tutor), que hacerlo sin esquema alguno. Es indispensable tener flexibilidad, partiendo de un esquema inicial que se irá enriqueciendo con la lectura de libros y artículos especializados.

Las fichas de lectura

Una pieza siempre presente en el trabajo intelectual tradicional consistía en varios tipos de *fichas*, término con el que se designaban unas tarjetas de cartulina de ciertas características. Las obras de metodología del estudio solían dar indicaciones en cuanto a rasgos formales y convencionalismos aceptados sobre la forma de hacerlas: las referencias comienzan siempre con el apellido del autor principal del trabajo, seguido por su nombre y la fecha de publicación. Eso basta para llenar una *ficha bibliográfica*, pero no para hacer *fichas de trabajo*, que son las que importan en una lectura orientada. Estas fichas recogen ideas relativas al tema de investigación de que se trate, con las que se elaborará el apartado de referentes.

Gracias al esquema inicial, por sencillo y provisional que sea, se podrán revisar una gran cantidad de los materiales identificados en la búsqueda, deteniéndose en aquellos que aporten elementos para la investigación, y pasando rápidamente por los que no lo

hagan. A medida que avance la lectura se irán registrando —en tarjetas de cartulina o, como es lógico ahora, con la ayuda de un procesador de texto— las ideas relevantes, teniendo en cuenta que es necesario hacer una *ficha por idea*, lo que quiere decir *tantas fichas como ideas relevantes distintas* se encuentren.

Cada ficha deberá incluir lo necesario para identificar de dónde se tomó, lo que será necesario para incluir en el texto resultante las referencias precisas y completas de todos los elementos tomados de cualquier autor, para dar el reconocimiento debido y no incurrir en la grave falta que es el plagio. Muchos investigadores hemos vivido la experiencia de tener una cita importante para incluir en un texto, en la que no se anotó la referencia, y que no localizamos por más esfuerzos que hacemos.

El trabajo de lectura y fichado debe interrumpirse en algún momento. No hay que caer en el error de querer leer todo, porque es imposible y lleva al conocido caso del investigador que nunca pasa del *marco teórico*.

El esquema con el que se inició la lectura se habrá modificado a lo largo de la misma, y se tendrá un buen número de fichas de trabajo. Procesada y asimilada, la información obtenida debe traducirse en un texto propio. No se trata de escribir muchas páginas, citando autores sin ton ni son, sino de plasmar en forma clara las ideas que sean relevantes para la investigación.

Para terminar el capítulo conviene decir que en muchos aspectos hay criterios precisos para saber si lo que hizo un investigador satisface o no los estándares de calidad que se consideran aceptables, como coeficientes de confiabilidad o pruebas de significación estadística. En cambio, no hay criterios tan precisos para acotar bien un objeto de estudio ni para hacer preguntas ricas que orienten la indagación y se muestren fecundas. La única orientación al respecto es que para ello lo que se necesita es saber lo más posible sobre el tema, gracias a una buena revisión de literatura, y a la familiaridad con los conocimientos prácticos —muchas veces tácitos— de los que trabajan al respecto en educación, en especial los maestros.

La importancia de la revisión de la literatura se puede apreciar considerando que la calidad de tal revisión es un indicador claro de la calidad de la investigación. Si el apartado del informe que reporta el resultado de dicha revisión no muestra que el autor está familiarizado con los principales autores del campo, y con las obras clave del mismo, habrá base para dudar de la calidad del conjunto del trabajo. Sin una buena revisión, es poco probable que los demás elementos considerados en este capítulo sean adecuados, incluyendo la construcción del objeto de estudio, la identificación de variables clave y la formulación de preguntas fecundas.

Conclusión

En la Introducción de la obra se comentó la definición de investigación educativa de la Ley *No Child Left Behind*, que parece reducirla a los estudios experimentales y cuasi-experimentales. Se comentó también la reacción de la comunidad de investigadores, en la forma de la definición mucho más amplia de la AERA, y en particular en el texto de Shavelson y Towne en el que se decía:

> [...] El diseño de un estudio no lo hace científico por sí mismo. Hay una amplia gama de diseños legítimos que se pueden usar en la investigación educativa que van de un experimento con asignación aleatoria para estudiar un programa de bonos educativos, a un estudio de caso etnográfico en profundidad de unos maestros, o a un estudio neurocognitivo de cómo se aprenden los números, utilizando tomografía por emisión de positrones para formar imágenes del cerebro [...]. (Shavelson y Towne, 2000: 6)

Esta es la idea que orientó este capítulo, coincidiendo con la obra de la que se han tomado muchas ideas, que advierte expresamente que los señalamientos sobre los límites de los estudios experimentales no se deben entender como una crítica de este tipo de diseño, pero sí de una forma de pensar sobre los métodos de investigación que parta de la idea de que hay un método superior, cuyas bondades deberían tratar de imitar todos los demás (*gold-standard thinking*):

> No sería un buen consejo para investigadores ni para tomadores de decisión elevar los modelos multinivel, las entrevistas semi-estructuradas, observación participante o cualquier otro buen método, al nivel de estándar único contra el que todos los demás deberían compararse [...] Partimos del supuesto de que todas las preguntas de investigación pueden abordarse de múltiples maneras, cada una de las cuales tiene ventajas y límites [...] la elección que uno haga de un diseño debe ser dirigida por la pregunta de investigación, por el contexto en que uno trata de responderla, y por los propósitos del estudio [...]. (Vogt, Gardener y Haeffele, 2012: 49)

Referencias

Adler, M. J. y Van Doren, Ch. (2000). *Cómo leer un libro. Una guía clásica para mejorar la lectura*. México: Plaza y Janés. Traducción de versión de 1972. Versión original 1940.

CONEVAL (2009). *Metodología para la medición multidimensional de la pobreza en México*. México: Consejo Nacional de Evaluación de Política de Desarrollo Social.

Lazarsfeld, P. (1973). De los conceptos a los índices empíricos. En Boudon, R. y P. Lazarsfeld, *Metodología de las ciencias sociales*, Vol. I. Barcelona, Laia, pp. 35-46.

Shavelson, R. J. y L. Towne (Eds). (2002). *Scientific Research in Education*. Washington: National Research Council. National Academy Press.

Silió, E. (2019). Revistas seudocientíficas para engordar currículos académicos. En *El País*, 14 de enero de 2019.

Treviño, E., Place, K. y Gempp, R. (2013). *Análisis del clima escolar. ¿Poderoso factor que explica el aprendizaje en América Latina y el Caribe?* Santiago: UNESCO-OREALC y Santillana.

Vogt, W. P., D. C. Gardner y L. M. Haeffele (2012). *When to Use What Research Design*. [Conclusion. Culmination of Design, Sampling and Ethics in Valid Data Coding, pp. 317-333]. New York: The Guilford Press.

Vogt, W. Paul (2007) *Quantitative Research Methods for Professionals*. Boston: Pearson.

CAPÍTULO 2
LOS DISEÑOS DE INVESTIGACIÓN

CONTENIDO

Introducción

Hay un gran número de técnicas de obtención de información (Cap. 3), y de análisis de la información obtenida (Cap. 4), pero en una investigación particular solo se utilizan unas cuantas, seleccionadas según el propósito que se persiga y la forma en que se haya acotado el objeto de estudio:

- Según el propósito: describir, comparar, identificar aspectos relacionados, seguir la evolución del fenómeno, explicar sus causas...
- Según la delimitación empírica: qué tipo de personas, casos, instituciones o procesos se estudiarán, cuántas o cuántos, situados en dónde...
- Y según la delimitación teórica: qué aspectos (variables) de esas personas, casos instituciones o procesos se considerarán...

En cambio, hay un número reducido de formas de organizar la investigación, cada una de las cuales utiliza preferentemente algunas técnicas de obtención y de análisis de la información. La noción de *diseños o tipos de investigación* alude a esas formas generales de organizar el trabajo.

En ciertos momentos alguno de esos diseños o tipos se ha llegado a considerar como la única forma de hacer investigación, en el contexto de trabajos inspirados en una u otra disciplina. Fue el caso de los experimentos, de los estudios de tipo encuesta, o de los estudios de casos. La investigación experimental, con la que Claude Bernard caracterizó la medicina científica, fue retomada por la psicología y la pedagogía desde la primera mitad del siglo XX. El desarrollo de trabajos sociológicos tras la

Segunda Guerra Mundial llevó a privilegiar un acercamiento extensivo, la encuesta; y el trabajo de los antropólogos hizo lo mismo con el acercamiento intensivo de los estudios de casos.

Experimentos, encuestas, estudios de casos son, pues, tres tipos de investigación entre varios posibles, que se llegaron a identificar con La investigación. Frente a lo limitado de esta concepción, se han hecho esfuerzos por mostrar que la gama de opciones en cuanto al diseño de una investigación es más amplia, que además de esos diseños hay otros, y que quienes pretendan dedicarse a la investigación deben tener al menos una idea de ellos, aunque solo manejen con soltura algunos.

La *American Educational Research Association* (AERA) ha difundido tres obras que presentan algunos diseños en el campo de la investigación sobre temas educativos, cuyo carácter no excluyente se destaca en el título de las tres obras: *métodos complementarios para la investigación en educación.*

La primera edición (Jaeger, 1988), después de una introducción sobre la naturaleza de la investigación educativa, firmada por Lee Shulman, presenta los *métodos* históricos, filosóficos y etnográficos, estudios de casos, encuestas, experimentos comparativos y métodos cuasi-experimentales. La segunda edición (Jaeger, 1997) añade un apartado sobre los métodos de investigación basados en las artes.

La tercera obra (Green, Camilli y Elmore, 2006) discute cuestiones filosóficas, epistemológicas y éticas; presenta técnicas como el análisis del discurso; distingue estudios definidos por su objeto, como sobre currículo o desarrollo humano; y trata también sobre diseños o tipos, incluyendo novedades como investigación de diseño (*design research*), método microgenético, estudios narrativos, metodología múltiple, investigación con los actores (*practitioner inquiry*) y síntesis de investigaciones.

Las ediciones del *Handbook of Research on Teaching* (HRT 1963, 1973, 1986, 2001 y 2016) dedican capítulos a cuestiones metodológicas, pero la más reciente las abordó de manera diferente. Según el Cap. 3, *Engaging Methodological Pluralism:*

> Es la primera vez que el HRT trata la metodología de investigación en un solo capítulo. Las ediciones anteriores tuvieron varios capítulos [...] muchos de los cuales podían ser ubicados en una u otra constelación de metodologías que se han etiquetado como cuantitativas-cualitativas [...] Si bien los términos cuantitativo y cualitativo pueden ser útiles para caracterizar algunos métodos, no representan de manera suficiente la rica gama de metodologías de investigación educativa, y dan una visión limitada de las diferencias básicas y los acercamientos únicos de ellas [...]. (Moss y Haertel, 2016: 127)

El capítulo distingue 10 tradiciones metodológicas: 1) Diseños experimentales y cuasi-experimentales; 2) Investigación etnográfica; 3) Estudios comparativos de caso (*small N)*; 4) Análisis del discurso; 5) Encuestas-mediciones; 6) Investigación basada en el diseño; 7) Análisis de redes sociales; 8) Investigación de sistemas complejos-adaptativos; 9) Teoría crítica de la raza; y 10) Investigación de acción participativa. A estos se añaden métodos mixtos y estudios multidisciplinarios, interdisciplinarios y transdisciplinarios. (Moss y Haertel, 2016: 131 y 133)

En el *Oficio del Investigador* propuse una tipología de diseños a partir de los rasgos que pueden tener según su propósito, fuente y forma de obtención de información, número de sujetos estudiados, participación del investigador, número de variables, nivel de análisis y dimensión temporal. Combinando estas dimensiones identificaba 10 diseños de investigación: 1) Histórica; 2) Situacional etnográfica; 3) De caso no situacional; 4) Comparativa; 5) Encuesta; 6) Explicativa no experimental; 7) Experimental; 8) Evaluativa; 9) Investigación y desarrollo; 10) Investigación acción.

La clasificación de diseños de investigación que se propone en seguida distingue estudios de *orientación básica* (cuyo propósito es solo ampliar o profundizar el conocimiento de algún fenómeno) y de *orientación aplicada*, que buscan aprovechar el conocimiento de un fenómeno para un propósito ulterior.

En los trabajos de orientación básica se distingue la *investigación viva,* que implica obtener información nueva, no recogida previamente, e investigación *documental*, que utiliza datos obtenidos y registrados antes. Se distinguen además los estudios *observacionales*, en que no se manipula ninguna variable, y los *no observacionales,* en los cuales algunas variables no solo se observan, sino que se manipulan.

La Tabla 2.1 sintetiza los rasgos básicos y algunas variantes de 14 diseños o tipos de investigación: siete de investigación básica viva (tres observacionales, dos no observacionales, uno longitudinal y uno de métodos múltiples), cuatro diseños de investigación básica documental, tres de investigación aplicada. Cada uno de estos 14 diseños se presenta luego con más amplitud, aunque una mejor comprensión se conseguirá con la lectura de las técnicas de obtención y análisis de información que se usan en cada diseño, que se verán en los capítulos 3 y 4.

TABLA 2.1. DISEÑOS DE INVESTIGACIÓN

Diseños	Características	Variantes
Investigaciones básicas vivas		
Encuestas	Estudian en forma extensiva algunos aspectos de números considerables de sujetos, aplicando cuestionarios o instrumentos similares	• Uni, bi, o multi-variadas • Descriptivas y asociativas • En casa, por teléfono, en línea
Estudios de caso	Estudian intensivamente pocos sujetos con entrevistas y otras técnicas	• Simples-múltiples; holísticos o multinivel; etnográficos
Estudios de observación visual	Indagan hechos y procesos sin limitarse a lo que dicen los sujetos, con base en la observación visual hecha por terceras personas de los fenómenos estudiados	• Narrativa • Estática-instantánea • Dinámica, de interacciones • Con videograbación
Experimentos	Buscan establecer relaciones causales comparando un grupo en que se altera(n) una(s) variable(s), con un grupo control, con asignación aleatoria de los sujetos	• Dos grupos, solo postest • Dos grupos pretest-postest • Varios tratamientos, factoriales • Longitudinales
Cuasi-experimentos	Se acercan a la identificación de relación causal controlando el efecto de terceras variables en formas que se aproximan a las de un experimento estricto	• Regression discontinuity • Propensity Score Matching • Experimentos naturales
Estudios longitudinales	Estudian el proceso de desarrollo de un fenómeno identificando los cambios que ocurren a lo largo del tiempo y tratando de explicar las causas que los producen	• Series de tiempo, interrumpidas • Edos. de cohortes o tendencias • Edos. transversales simultáneos • Estudios de panel • Análisis de la historia de eventos
Estudios mixtos	Combinan dos o más de los diseños anteriores y/o de los siguientes	• Encuesta + caso o viceversa • Combinación de más diseños
Investigaciones básicas documentales		
Investigación de archivo	Estudian fenómenos históricos o actuales con información guardada en archivos	• De archivos formales • De cartas, diarios y similares
Anál. secundario de bases de datos	Hacen análisis adicionales de información que se conserva en bases de datos	• Diversas posibilidades como en las encuestas
Análisis de evidencias	Infieren información a partir de materiales que no se generaron para ello	• Físicas (acumulación-desgaste) • Escritas, portafolios, tareas
Síntesis de investigaciones	Integran información de estudios previos buscando coincidencias y discrepancias	• Narrativas, de conteo de votos • Meta-analíticas

Investigaciones aplicadas		
Estudios de intervención	Buscan expresamente producir algún tipo de cambio en la realidad que estudian	• Intervenciones controladas, design research • Investigación acción, practitioners' research
Investigaciones evaluativas	Buscan llegar a juicio de valor en relación con un referente, sustentar decisiones o valorar el impacto de las ya tomadas	• Evaluación de personas • Evaluación de programas • Evaluación de políticas
Investigaciones metodológicas	Buscan desarrollar o mejorar técnicas de obtención o análisis de información, sin llegar a conclusiones sobre una realidad	• Diseño de instrumentos • Estudios de validación

FUENTE: ELABORACIÓN PROPIA.

Investigaciones básicas vivas

Encuestas

Aunque ya no se ven como el diseño de investigación por excelencia, las encuestas (*surveys*) siguen siendo uno de los más utilizados. Según Vogt, Gardner y Haeffele, *probablemente es exacto decir que en ciencias sociales se han recogido más datos utilizando encuestas que por medio de cualquier otro tipo de diseño* (2012: 15)

Las encuestas tienen en común que mediante ellas se busca recoger información sobre el tema de estudio por medio de las respuestas que un número considerable de sujetos dan a las preguntas que se les hacen. Por ello suelen usar cuestionarios que se aplican a gran número de sujetos (muestra amplia o censo), y analizan la información recabada con técnicas estadísticas. Este diseño es adecuado si la información necesaria para responder las preguntas es conocida por los miembros de la población de interés; si esa información no es demasiado compleja o abstracta, por lo que es posible explorarla con preguntas directas, razonablemente claras; y si no hay razones para temer que los informantes tengan reticencia para ofrecerla, dejando de responder o con respuestas que no correspondan a lo que piensan.

Las variantes de las encuestas tienen que ver con las técnicas que se usan en cada una. En cuanto a la obtención de información, es posible usar cuestionarios simples o que usen algún tipo de escala; con preguntas de respuesta estructurada o abierta; autoadministrados, aplicados en persona o por teléfono, o en línea. En cuanto al análisis que se quiere hacer, una encuesta puede servir para responder preguntas descriptivas sobre unas variables; para identificar asociaciones entre dos o más variables; para aproximarse a explicaciones causales tratando de controlar el posible impacto de algunas con procedimientos estadísticos. Será más frecuente usar encuestas en estudios transversales, en un solo momento, pero también

se podrán utilizar en estudios longitudinales, que se desarrollan a lo largo del tiempo, siguiendo un panel o una cohorte, o con cohortes sintéticas o aparentes. Las variantes que resultan de la combinación de posibles formas de obtener y analizar la información son numerosas, como señalan Vogt, Gardner y Haeffele:

> Simplemente con tres formas básicas de aplicar un cuestionario (en persona, por teléfono o autoadministrado), tres tipos de muestra (con panel, cohorte o transversal) y dos formatos de pregunta (abierta y de elección forzada), se llega a 18 combinaciones de posibles diseños [...]. (2012: 27)

Lo anterior basta para reconocer lo limitado de la caricatura que reduce la encuesta —o incluso toda investigación— a la secuencia rígida de pasos que va de plantear el problema, formular hipótesis, diseñar la muestra, elaborar un cuestionario, aplicarlo, capturar la información, analizarla, y redactar el informe con las conclusiones.

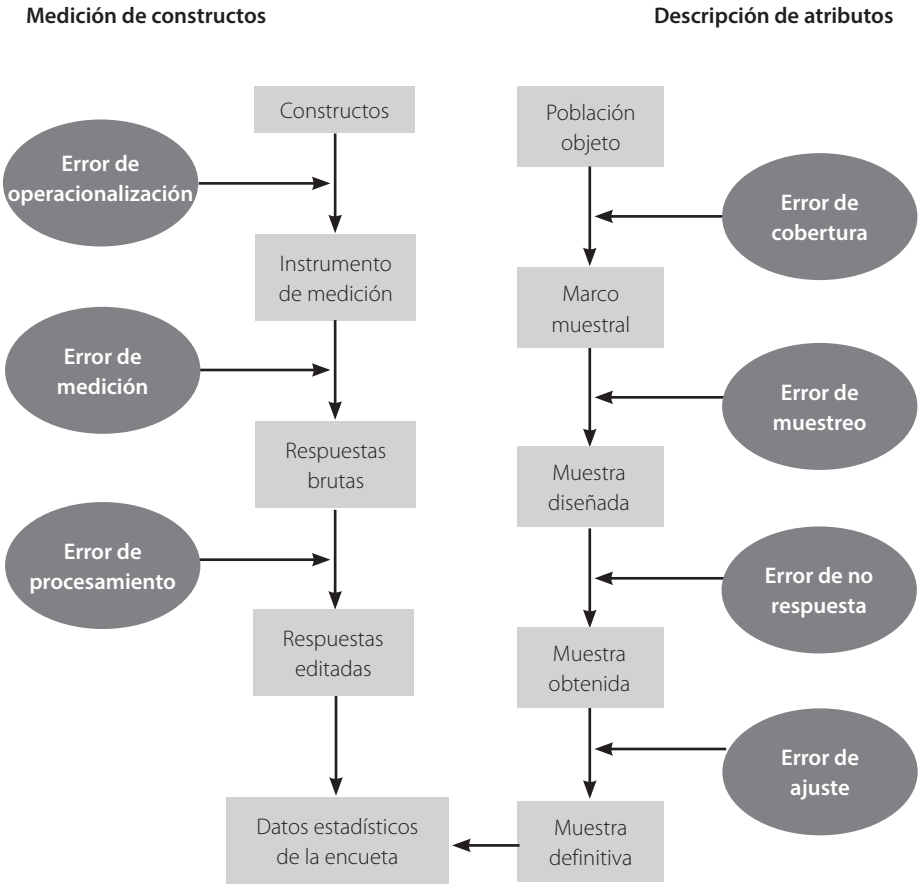
Planeación de una encuesta

Definidos los objetivos, las preguntas a responder, se deben precisar dos aspectos: el conjunto de sujetos del que se estudiarán todos o algunos casos (universo, marco de muestreo); y la técnica de obtención de información que se utilizará. Según Groves *et al.* (2009), se deben considerar diversos pasos, que se pueden organizar en dos conjuntos o secuencias:

- **Medición de constructos:** definir operacionalmente conceptos definidos de manera abstracta (*constructos*), para dar lugar a las preguntas o ítems de un instrumento con el que se obtendrán respuestas brutas, que serán editadas, de manera que constituyan la información sobre los constructos iniciales.
- **Descripción de atributos de una población:** definir la población objetivo de la que se quiere estudiar algo, con una muestra cuya obtención implica la conformación de un marco muestral; cubrir de manera más o menos amplia la muestra diseñada, lo que llevará a hacer ajustes a la muestra obtenida al depurar las respuestas.

El conocimiento obtenido nunca es perfecto. Cada paso implica cierta inferencia y tiene siempre algún margen de error, como sugiere la gráfica 2.2.

GRÁFICA 2.2. LA ENCUESTA EN LA PERSPECTIVA DE LA CALIDAD DE LA INFORMACIÓN



FUENTE: GROVES *ET AL.*, 2009: 48, FIGURA 2.5.

La inferencia en el lado de la medición permite pasar, de las respuestas que da un sujeto a las preguntas de un instrumento, a las características de ese sujeto que se quieren estudiar, a partir de unos constructos. En cuanto a la descripción de atributos, los pasos permiten inferir ciertas características de la población objetivo, a partir de las observadas en la muestra definitiva. (Groves *et al.*, 2009: 39-63)

Críticas fundadas e infundadas a las encuestas

Después de ser consideradas la forma por excelencia de investigación en ciencias sociales, las encuestas han sido objeto de críticas, algunas de las cuales cuestionan únicamente

la solidez de generalizar con base en datos de muestras de tamaño reducido, pero en ocasiones llegan a descalificar por completo estos estudios, lo que a su vez puede carecer de sustento sólido. Para ilustrar el punto, se retoma una discusión sobre los resultados de la Sexta Encuesta Nacional Sobre Inseguridad (ENSI-6 2009) difundida por el Instituto Ciudadano de Estudios sobre la Inseguridad, A. C., sobre la frecuencia de varios tipos de hechos delictivos.

La información se refería a asesinatos, secuestros y otros delitos, y concluía que su frecuencia no había disminuido, pese a los esfuerzos gubernamentales. Como el estudio identificaba las entidades federativas en que tales hechos eran más frecuentes, algunos gobernadores lo cuestionaron, diciendo que la información se refería a una fecha algo lejana; por otra parte, y principalmente, cuestionaban que se tratara de una *encuesta de opinión,* que no reflejaba necesariamente la situación objetiva, sino la *percepción* de los interrogados, influenciada seguramente por la excesiva atención que los medios de comunicación prestan a los hechos violentos, así como al carácter especialmente cruel que se presenta en ciertos casos.

Dejando aparte la cuestión de la demora en la difusión de los resultados, que puede ser lamentable, pero no es un defecto del estudio, el punto fundamental de la crítica no es adecuado, porque el estudio *no era una encuesta de opinión*, que explorara la percepción de los encuestados sobre el grado de violencia existente en su lugar de residencia; se trataba de una *encuesta de victimización*, cuya realización implica interrogar a una muestra sobre hechos muy precisos: si el sujeto o sus familiares cercanos han sido o no *víctimas* de ciertos delitos. Si una persona responde afirmativamente a la pregunta sobre un delito, se le hacen preguntas adicionales sobre sus circunstancias: cuántas veces, cuándo, en qué lugar, a qué hora, cuántas personas intervinieron, cuáles eran sus características. Una pregunta que no debe faltar es la relativa a si denunció o no el hecho, en caso afirmativo cuándo y dónde, y en caso negativo las razones por las que no lo hizo.

La diferencia es clara: una *encuesta de opinión* indaga sobre cuestiones subjetivas: si el encuestado cree que la violencia ha aumentado o disminuido, si cree es mayor en tal ciudad o en otra, si le preocupa mucho o poco, si le parece más grave un secuestro o una violación, o si le parece que es más preocupante la violencia o la situación económica. Una *encuesta de victimización* pregunta a cada persona si él o ella, o sus familiares cercanos, han sido víctimas de robo sin violencia, asalto a mano armada, secuestro, violación, asesinato u otro delito, sin indagar su juicio al respecto. Obviamente es posible que una persona responda falsamente preguntas de este tipo, pero varias razones muestran que se trata de una metodología que puede ser muy adecuada: la confiabilidad de las

respuestas puede aumentar mucho si la encuesta se hace con todo el cuidado que ha mostrado la experiencia al respecto, que tiene ya varias décadas; es posible también corroborar la veracidad de la parte de las respuestas que informen haber denunciado ciertos hechos, cruzando la información con los registros correspondientes.

Es sabido que una parte considerable de los delitos no son denunciados por su víctima por varias razones: desconfianza en la autoridad, vergüenza, desproporción entre el tiempo que se invertirá y el beneficio que se podrá obtener, u otras por el estilo. Por esta razón los registros de las autoridades judiciales y la policía no bastan para conocer con precisión la incidencia de los diversos delitos, ya que una parte importante, que no es fácil precisar, no se denuncia: *la cifra negra de delitos*.

Esa cifra negra es la que las encuestas de victimización estiman con cierta precisión. Parece la mejor forma de hacerlo, y se utiliza hace décadas en las sociedades más avanzadas.

> [...] aunque algunos crímenes son denunciados a la policía, muchos no lo son. La mejor manera de estimar la proporción de personas que son víctimas de ciertos crímenes es preguntar a una muestra sobre sus experiencias de victimización. (Fowler, 1995: 1)

Encuestas de este tipo llevan tiempo; no es posible hacerlas con mucha frecuencia; sin embargo, como la proporción de delitos denunciados es relativamente estable, la información de una encuesta de victimización permite estimar la cantidad de delitos de cierto tipo realmente cometidos en cierto lugar, a partir de la cifra de los que se hayan denunciado en el lapso de que se trate.

Supongamos que se encuentra que, en cierta ciudad, los asesinatos y los robos de automóvil se denuncian en 90% de los casos; los secuestros y las violaciones en 60%; los robos con violencia en 40%; y los robos sin violencia en 10%. A partir de ello se puede calcular sencillamente el número probable de cada delito que se haya cometido en cierto tiempo, con base en las denuncias presentadas.

Lo anterior no es una defensa del estudio de referencia, pero sí un señalamiento de que las críticas que se le hicieron carecen de sustento sólido, ya que se basaron en una idea equivocada de su naturaleza. El estudio pudo ser bueno o malo, pero su calidad no se debió a que fuera una encuesta de opinión, sino que depende de factores a los que las críticas no se refirieron y sobre los que la opinión pública no tuvo información:

- De qué tamaño fue la muestra utilizada.
- Cómo se calculó y cómo se seleccionó a los entrevistados.

- Cómo estaban formuladas las preguntas.
- Qué procesos de validación de las preguntas se hicieron antes de aplicarlas.
- Cómo se presentaron a los sujetos.
- Qué cuidados se tuvieron para asegurar la confidencialidad.
- Qué procedimientos de validación *a posteriori* se utilizaron.
- Cómo se procesó la información obtenida.

Si esos aspectos fueron bien tratados, podrá decirse que la información obtenida es de buena calidad, y refleja de manera razonable la situación del fenómeno. De lo contrario los resultados serán de baja calidad, pero la conclusión no debería ser que ese tipo de estudios no sirve, sino que hay que hacerlos bien, lo que implica tiempo y recursos, en especial personas calificadas para atender correctamente las exigencias metodológicas, técnicas y prácticas de cada aspecto relevante.

Ningún procedimiento de obtención de información es perfecto, ninguno es en sí mismo superior o inferior a otros, todos tienen pros y contras; unos son más apropiados para obtener cierto tipo de información, y otros para otros. La decisión de utilizar una u otra estrategia dependerá de lo que se busque y de las circunstancias en que deba obtenerse la información.

La encuesta es adecuada si la información a obtener puede consistir en respuestas breves a preguntas estructuradas; si se puede asumir que los sujetos a los que se preguntará conocen la respuesta y están dispuestos a darla; si el investigador tiene clara la relación de cada pregunta con alguno de los constructos que le interesa estudiar; y si es posible tener un marco muestral razonable, medios para llegar a un número suficiente de sujetos y lograr que respondan preguntas que se les formulen.

TABLA 2.2 COMPARA VENTAJAS Y DESVENTAJAS DE TRES VARIANTES DE ENCUESTA

Ventajas y desventajas	**Forma de aplicación**		
	Cara a cara	**Por teléfono**	**Auto-administrado**
1. Utilizable con sujetos que no saben leer	Sí	Sí	No
2. Aplicador puede explicar preguntas	Sí	Sí	No
3. Identidad del que responde	Segura	Probable	Dudosa

4. Interacción con aplicador	Alta	Media	Baja
5. Misma pregunta para todos	Dudoso	Probable	Seguro
6. Costo por respondente	Alto	Medio	Bajo
7. Tiempo por pregunta	Alto	Alto	Bajo
8. Posible aplicación grupal	Sí	No	Sí
9. Dificultad de muestra grande	Alta	Media	Baja
10. Acceso a áreas remotas	Difícil	Fácil	Fácil
11. Dependencia de correo, teléfono…	No	Sí	Sí

FUENTE: VOGT, GARDNER Y HAEFFELE, 2012: 20, TABLA 1.1.

Vogt, Gardner y Haeffele señalan que variantes de las encuestas se pueden usar para estudios de carácter longitudinal, con diseños de tipo panel, de cohortes, de estudios transversales repetidos, o con los métodos llamados *de la historia de los eventos* (2012: 23-25). También tratan cuestiones éticas que deberá tener en cuenta el investigador que haga encuestas, y que tienen que ver con la forma de asegurar el anonimato de quienes respondan un cuestionario, y con la importancia de obtener su consentimiento expresamente formulado, con información de qué implica el aceptar ser considerado en un estudio de esta naturaleza. (2012: 241-252)

Estudios de caso

Como las encuestas, los estudios de caso tienen un propósito básico y un carácter observacional, ya que no incluyen manipulación de variable alguna. A diferencia de las encuestas no pretenden alcanzar muchos sujetos, sino un número reducido, en el límite solo un caso, rasgo que da su nombre al diseño. La diferencia entre estudios de caso y encuestas se aproxima a la que opondría lo cualitativo a lo cuantitativo, pero estrictamente estos términos se refieren al nivel de medición de las variables, y no caracterizan correctamente los diferentes diseños, ya que en cualquiera de ellos es posible incluir variables medidas al menos ordinalmente, e incluso cardinalmente. Por ello aquí se prefiere la pareja de términos extensivo e intensivo, y se distinguen tres tipos de estudios de caso:

- Unos parten de conceptos cuyas dimensiones o categorías se especifican desde el inicio, lo que permite formular preguntas de investigación o hipótesis precisas, y usar técnicas de obtención y análisis de la información de carácter estructurado, lo que hace que se les considere cuantitativos.
- Otros parten de conceptos menos precisos y preguntas más generales, y proponen construir categorías a partir de información recabada con técnicas no estructuradas de obtención y análisis de información (*grounded theory*, etnografía), por lo que se les considera cualitativos; las concepciones del conocimiento subyacentes son compatibles con las del grupo anterior.
- Los trabajos del tercer grupo parten de concepciones sobre el conocimiento (epistemológicas) que sus defensores consideran incompatibles con las que suponen los otros dos tipos, y utilizan también técnicas cualitativas.

Estudios de caso estructurados

Una encuesta explora pocos aspectos o variables (de tres o cuatro a algunas decenas) de un número grande de sujetos (desde pocos centenares hasta muchos miles); un estudio de caso trata de indagar el mayor número posible de aspectos (aunque nunca sea exhaustivo), en un número reducido de sujetos (personas, pero también grupos o instituciones), en el extremo uno solo. (Ashley, 2012: 102)

Los estudios de caso difieren de las encuestas por la técnica de obtención de información que usan preferentemente. Cuando se busca explorar el conocimiento subjetivo de los involucrados, y se esperan respuestas amplias, que implican tiempo para reflexionar y elaborarlas, la entrevista es preferible al cuestionario. Por ello este diseño es apropiado si es más importante profundizar en el conocimiento de un fenómeno que generalizar lo que se encuentre a una población más amplia. (*Cfr*. Vogt, Gardner y Haeffele, 2012: 47)

Robert Yin desarrolla el tipo de estudios de caso de los que aquí se trata, señalando que no son solo una etapa preliminar de estudios más amplios, sino un diseño distinto, que puede ser exploratorio, o descriptivo, e incluso explicativo. Por su carácter intensivo son indicados no solo para responder preguntas sobre *qué*, sino también cuestiones sobre *cómo* y *por qué,* lo que los hace apropiados para estudios evaluativos y explicativos. (Yin, 1984: 41-54)

Cuando se busca encontrar explicaciones, el sustento de las conclusiones causales de un estudio de caso no es el control estadístico, con asignación aleatoria, de todas las variables que pueden explicar un resultado, sino el control directo de pocas variables

clave. La lógica de generalización del estudio de caso no es la estadística del muestreo, sino la analítica de la replicación. (Yin, 1984: 39-40 y 48-49)

Este punto de vista es apoyado nada menos que por Donald T. Campbell. En el prólogo de la obra citada de Yin, el principal expositor de los diseños experimentales subraya que lo esencial al buscar explicaciones causales no es la experimentación, sino descartar explicaciones alternativas. (En Yin, 1984: 7-8)

Yin distingue cuatro variantes de los estudios de caso, que resultan de combinar dos dimensiones, cada una de las cuales tiene dos posibilidades: por una parte, el número de casos considerados, uno o varios; por otra, si se considera como unidad de análisis solamente una, la totalidad del objeto de estudio, o bien si se consideran varias unidades de análisis, en varios niveles que forman una especie de estructura anidada, *v. gr.* la institución completa, dependencias particulares de la misma, y personas individuales en cada dependencia. Tenemos así:

- Estudios de caso simples (*Single-case designs*), en los que se distinguen los de tipo holístico (solo una unidad de análisis, que se aborda en forma global), y los de enfoque anidado (*embedded*), con múltiples unidades de análisis en una estructura de varios niveles.
- Estudios de caso múltiples (*Multiple-case designs*), en los que se puede distinguir también aquellos en que se trabaja con una sola unidad de análisis (holísticos), o con varias unidades anidadas (*embedded*).

Sturman (1997: 63) retoma de Stenhouse (1985) la distinción entre cuatro estilos de estudio de caso: etnográficos, de investigación acción, evaluativos y educativos.

En estudios de caso los preparativos para el trabajo de campo implican gestiones más o menos complejas de acercamiento a la institución o las personas de que se trate, para conseguir que acepten el esfuerzo que implicará atender las peticiones de los investigadores que recogerán información.

Yin advierte que para el proceso de recolección de información hay que aprovechar varias fuentes además de entrevistas: revisar documentos y archivos; observación (participante o no); y recolección de artefactos o evidencias. (1984: 55-98).

Los pasos que implica un estudio de caso son básicamente:

- Selección del caso y negociación del acceso.
- Trabajo de campo: entrevistas, observación participante o no,

grabación de audio o video, recolección de evidencia documental, entre otras técnicas.
- Organización de los registros.
- Redacción del informe. (Stenhouse, 1985)

Hay que añadir que una investigación de caso implica un proceso de estructuración progresiva, en el que se puede distinguir una primera fase exploratoria, seguida por una semiestructurada. (Ashley, 2012: 105)

En cuanto al análisis de información, habrá que apoyarse en la teoría que sustenta el estudio y desarrollar un marco en el que se encuadre la descripción del caso. Las estrategias analíticas particulares incluyen identificación y comparación de patrones en cuanto a variables dependientes o explicaciones alternativas; construcción de explicaciones; análisis de series de tiempo; análisis de unidades anidadas, cuando las hay; realización e observaciones repetidas; cuando se cuenta con un número considerable de estudios de caso, un análisis secundario del conjunto en cuanto a ciertos aspectos (*case survey*). (*Cfr.* Yin, 1984: 99-120)

Estudios de casos no estructurados

El segundo tipo de estudios de casos se distingue del anterior por su enfoque inductivo. Al no partir de conceptos con categorías definidas al inicio del proceso, por lo que no puede utilizar instrumentos estructurados, y recurre a acercamientos como la llamada teoría fundamentada (*grounded theory*) y la etnografía.

Subrayo que los trabajos que se incluyen en este grupo privilegian el uso de técnicas de obtención y análisis de la información cualitativas, pero comparten con el grupo anterior las ideas básicas sobre el conocimiento que se obtiene con la investigación y el conocimiento en general, aunque se usen formulaciones propias de nociones como objetividad, confiabilidad y validez, posibilidad de generalizar o causalidad.

Sin pretender dar cuenta de estos acercamientos, por mi insuficiente conocimiento, me limitaré a remitir al lector a obras como las de Delgado y Gutiérrez (1999); LeCompte, Millroy y Preissle (1992); y Lincoln y Guba (1985).

Llamo también la atención de los lectores sobre un asunto tratado en uno de los últimos textos publicados en vida por Eduardo Weiss. El texto parte de constatar que *la mayoría de los trabajos en la investigación social y educativa en México se realizan hoy mediante métodos cuantitativos*, y que en ellos se afirma que se hicieron *desde el enfoque de la teoría fundamentada*. Subraya la poca congruencia de estas afirmaciones con lo

que muestran luego muchos trabajos que parten de un marco teórico sobre grandes autores poco relacionados con los hallazgos centrales de la tesis, y buscan legitimar su metodología por haber usado un software como Atlasti o NVivo para la generación de categorías de análisis. (Weiss, 2017: 637)

Como anuncia el título de su artículo, Weiss destaca la necesidad de reconocer la importancia de las aportaciones a la metodología de la investigación social de la hermenéutica de Gadamer y la descripción densa de Geertz, frente al predominio de la codificación con software, y el uso no consistente de la teoría fundamentada. El trabajo cierra con una reflexión, en perspectiva epistemológica, sobre la relación entre comprensión e interpretación, descripción y explicación.

A partir de la convicción de que, bien entendidas, las concepciones epistemológicas de los enfoques estructurados y no estructurados son compatibles, el Recuadro 2.1 sintetiza los pasos a seguir en estudios de caso de corte cualitativo, como los presenta una especialista de este enfoque metodológico (Merriam, 2009). El lector podrá apreciar las similitudes con los pasos de un estudio más estructurado, que se suele considerar cuantitativo.

RECUADRO 2.1. PASOS DE UN ESTUDIO DE CASO CAULITATIVO

- Elección del tema. El primer paso no puede ser otro, pero como se señaló al tratar la construcción del objeto de estudio, es frecuente que esto se haga de manera vaga, lo que no es productivo. (2009: 55-58)
- Del tema al problema. Pasar del tema a definir el problema de investigación supone hacerlo en términos precisos, incluyendo la descripción del contexto en que se sitúa el tema-problema, su importancia, las lagunas que hay en lo que se sabe del mismo, y, finalmente, el propósito del estudio. (2009: 58-64)
- Marco teórico y revisión de literatura. La definición del problema y de su propósito deben situarse (*enmarcarse*) en lo que se sabe previamente. Es frecuente pensar que los estudios cualitativos no deben partir de un marco así, y que deben proceder inductivamente; según Merriam eso es un error, y retoma de Schwandt la idea de que *una investigación ateórica es imposible* (2009: 66). La construcción de un marco teórico implica, necesariamente, la revisión de la literatura relevante. Hay que añadir que esta revisión puede hacerse en paralelo al trabajo de campo, y no necesariamente antes, lo que también aplica para cualquier otro estudio. (2009: 71-76)
- Elección de la muestra. El término se asocia generalmente a la variante de una muestra calculada estadísticamente y extraída aleatoriamente de un marco muestral, para que sea representativa de cierta población, dentro de ciertos márgenes de error y con cierta probabilidad. Pero si se usa de manera más general, la palabra muestra denota simplemente el sujeto (o los sujetos) que se van a estudiar en una investigación, lo que sin duda debe precisarse.

- Se puede tratar de un solo caso o de varios (pocos) que se quiere comparar, pero no hay que perder de vista que, aunque se trabaje con uno o pocos casos, este o estos se han seleccionado por alguna razón, que puede ser el que se le(s) considera representativo(s) de alguna forma (si se quiere, cualitativamente) de un número mayor. Merriam menciona varias formas no aleatorias de elegir el o los sujetos a estudiar, con lo que se llama muestreo intencional, con variantes como el del caso típico, único, el criterio de máxima variación, el muestreo en bola de nieve, etc. (2009: 76-82)
- Trabajo de campo. La recolección de datos, con técnicas de entrevista, observación y revisión documental. (2009: 85-163)
- Análisis. Merriam señala procedimientos de análisis de datos cualitativos, formas de atender cuestiones de confiabilidad, validez y ética, y lo relativo a redacción del informe (2009: 169-264). Es importante al respecto la obra clásica de Miles y Huberman (1984, con versiones ampliadas recientes).

FUENTE: MERRIAM, 2009.

Sobre otros estudios de tipo intensivo

En estas páginas no se incluyen trabajos que parten de nociones epistemológicas incompatibles con las que aquí se consideran más sólidas, como se argumenta en la Conclusión. Esos trabajos usan también técnicas cualitativas e incluyen estudios de autoetnografía, de enfoque feminista o perspectiva de género y crítica, algunos estudios socioculturales, entre otros.

Estudios de observación visual

El término observación se puede entender en sentido amplio, como sinónimo de obtención de datos mediante cualquiera de nuestros sentidos, o en sentido estricto, cuando se trata de estudios en los cuales la información se obtiene precisamente con los sentidos de la vista y del oído. Este inciso trata de observación en el segundo sentido. Lo que caracteriza a un estudio de observación visual es, pues, el tipo de instrumento de obtención de información: la observación visual del fenómeno en estudio, en vivo o en forma diferida, mediante videograbaciones.

Este diseño es apropiado cuando el investigador se interesa por un fenómeno cuyo desarrollo puede atestiguar personalmente a medida que sucede, y las preguntas de investigación tienen que ver con sus variaciones, contexto y los aspectos que se quiere describir en detalles (*thick description*). También procede si se desea estudiar el fenómeno en profundidad para descubrir mecanismos causales o reconocer relaciones antes no advertidas entre variables, máxime si se trata de un fenómeno no bien comprendido. (*Cfr*. Vogt, Gardner y Haeffele, 2012: 68-73)

La observación puede ser participante o no, según que el investigador se incorpore o no a la actividad del grupo; además, el investigador puede ser identificado como tal por los actores del fenómeno en estudio, o no serlo. Combinando estos dos aspectos se distinguen cuatro tipos de diseño:

- Observación pasiva (naturalista), que puede ser cubierta (investigador no identificado como tal) o abierta (investigador identificado como tal).
- Observación activa (participante), igualmente cubierta o abierta.

La observación naturalista es adecuada si los procesos que interesan tienen lugar en espacios públicos, sobre todo si se trata de primeras aproximaciones a variables delicadas, y es importante no influir en quienes participan en el fenómeno, o si una observación participante es imposible, peligrosa o suscitaría problemas éticos. Una observación participante, en cambio, es adecuada si se busca entender cómo ocurre un fenómeno desde adentro, si interesa entender los puntos de vista de los actores a medida que se desarrollan y si se pretende influir de alguna manera en el fenómeno que se estudia. (Vogt, Gardner y Haeffele, 2012: 77-81)

En el terreno educativo, los estudios de observación visual son particularmente apropiados para estudiar procesos que tienen lugar en el interior de los salones de clase, tanto las prácticas de enseñanza (docentes) que ponen en juego los maestros, como las de aprendizaje (discentes) de los estudiantes. Estos aspectos son esenciales para explicar, por ejemplo, la diversidad de aprendizajes alcanzados por los alumnos, o para evaluar a los maestros; además son procesos sumamente complejos para cuyo estudio no bastan encuestas ni estudios de caso.

Por ello no sorprende que haya una larga tradición de investigaciones educativas basadas en observación visual (Medley y Mitzel, 1963; Rosenshine y Furst, 1973; Stallings, 1977; Everston y Green, 1986; Good y Brophy, 2000; Floden, 2001; Goe, Bell y Little, 2008).

Además de modestos trabajos de observación hechos por supervisores desde fines del s. XIX y principios del XX, en el Cap. 3 se identifican:

- Estudios de 1950 a 1980, cuando prevalecía la idea de evitar inferencias, limitándose a registrar manifestaciones visibles de la conducta de los sujetos.
- Estudios de 1990 a la fecha, para estudiar procesos complejos, reconociendo que al observar no se puede dejar de hacer inferencias, pero tratando de asegurar una

razonable confiabilidad por el diseño de los instrumentos y el control estadístico a partir de varios observadores.
- Estudios basados en la observación diferida de videograbaciones de los fenómenos a estudiar.

Como las encuestas y los estudios de caso, los estudios de observación visual comienzan con la definición del objeto de estudio y precisar las preguntas de investigación que se quiera responder, lo cual implica elaborar un marco de referencia teórico, basado en la correspondiente revisión de literatura.

Para precisar cuántos sujetos incluir en una observación visual estructurada, la principal consideración no es asegurar la representatividad de la muestra, de modo que puedan hacerse generalizaciones, sino la de contar con un número suficiente de casos para poder hacer los análisis estadísticos necesarios para verificar la calidad de la información que se obtiene.

Tanto en observaciones visuales en vivo, como de preferencia en observaciones retrospectivas de videograbaciones, no bastará un observador, sino que deberá haber varios, que codifiquen de manera confiable la información.

La principal diferencia se refiere a la preparación o adecuación del instrumento de obtención de información, que en estos estudios reviste particular complejidad. Las lecciones que dejan las experiencias recientes para desarrollar sistemas de observación complejos muestran que la tarea requiere muchos meses del trabajo de equipos conformados por especialistas en los muchos aspectos involucrados.

El análisis de la información que se obtiene con un estudio de observación visual reviste, también, particular complejidad. Las videograbaciones, cada vez más frecuentes, se usan por las ventajas que tienen, y por la creciente accesibilidad de la tecnología necesaria, pero traen complicaciones adicionales.

Experimentos

El desarrollo de la ciencia se asocia con la idea de *experimentar*, entendiendo esta palabra en sentido amplio, de *observar* la realidad para comprender cómo es y cómo funciona, frente a las posturas que apostaban a la lectura de textos sagrados, o autores consagrados, como fuente de conocimiento no solo de cuestiones relativas a otro mundo, el de los dioses, sino también de este mundo, el de los humanos, los animales, plantas, montañas y océanos. Después el término *experimentar* se entendió como observar una realidad que se había alterado o manipulado

de alguna manera, para observar lo que ocurría y ampliar así el conocimiento de la realidad.

En el lenguaje cotidiano, el término en cuestión se sigue usando en uno u otro de los dos sentidos mencionados, y es en el campo de la investigación y de su metodología donde ha tomado el sentido más estricto, de una observación que se hace sobre una realidad que se modifica de alguna manera, pero además con el requisito de que haya un referente que no se haya alterado, y que se asegure la equivalencia de las dos realidades que se contratan, la modificada y la no alterada. Este apartado busca dar cuenta de la experimentación en este sentido estricto.

En este inciso se presentan conceptualmente los experimentos, y se dejan para el Cap. 4 las herramientas para analizar los resultados de estos diseños, buscando establecer relaciones de causa-efecto, y no solo de correlación entre dos variables, como las tablas de contingencia y los coeficientes de correlación parciales, y las diversas clases de Análisis de Varianza y Análisis de Regresión.

Se señala que esas técnicas enfrentan el problema de controlar no solo unas variables cuya influencia se puede confundir con la de la variable que se cree *produce* cierto efecto, sino *todas las potencialmente intervinientes*, incluyendo aquellas en que las que el investigador ni siquiera ha pensado. Se dice también que los experimentos estrictos, con asignación aleatoria de sujetos a los grupos de tratamiento y control son la solución de ese problema y constituyen, en principio, la manera más firme de establecer causalidad.

En ese lugar el tema se trata desde la perspectiva de las técnicas de análisis que se pueden usar en estudios experimentales, pero no se trata de los tipos de estudios de este enfoque, los diseños, que son ahora el foco de atención.

En el sentido estricto que se da hoy al término en medios especializados, se espera que un experimento resuelva el problema de controlar toda variable que incida potencialmente en un fenómeno que se quiere explicar, incluso aquellas de cuya existencia y posible influencia no se tiene idea.

Para que tan ambiciosa pretensión tenga sustento, un experimento estricto no consiste en una observación cualquiera, ni siquiera en la observación de un fenómeno que se ha alterado de alguna manera, sino que requiere: tomar un buen número de sujetos o casos; formar dos grupos, asignando en forma completamente aleatoria, la mitad de los sujetos/casos a uno y la mitad al otro; modificar en un grupo la variable que se cree puede ser causa del fenómeno a explicar, en tanto que en el otro grupo no se toca; y observar lo que pasa después en ambos grupos con el fenómeno a explicar.

Si el número de sujetos/casos es suficiente, y se asignan a los grupos de manera realmente aleatoria, se asegura que los dos grupos sean similares en toda variable, conocida o no, y que las diferencias que se observen solo puedan atribuirse a la variable que se modificó en uno de los grupos. En los dos grupos aleatorios habrá una proporción similar de cualquier característica, trátese de rasgos que parezcan relevantes para explicar el fenómeno (sexo, nivel socioeconómico, grupo étnico, escolaridad de los padres o cierto rasgo de personalidad...), pero también rasgos que parezcan irrelevantes, como el color de ojos de los sujetos, o su signo zodiacal.

En términos más precisos, suponiendo definido el objeto de estudio (el aspecto de la realidad que se quiere explicar), para encontrar la(s) causa(s) que lo produce(n), un experimento estricto supone al menos:

- Dos grupos, experimental y control.
- Equivalencia de ambos, gracias a la asignación aleatoria de sujetos.
- Intervención en el grupo experimental, o tratamiento.
- Medición previa y/o posterior (*pretest-postest*) del aspecto(s) en cuestión.
- Comparación de los resultados.

Con estos elementos se puede configurar gran número de variantes, de las que en este inciso se presentan algunas, a partir de la obra de Shadish, Cook y Campbell (2002), que recoge los avances de esta área de la investigación, retomando la herencia de Donald Campbell (Campbell y Stanley, 1963).

En esa obra clásica se recuerda que las primeras ideas sobre diseño experimental, en agricultura, se suelen encontrar en una obra de Ronald Fisher publicada en 1925, pero que dos años antes William McCall había propuesto ideas similares, en el ámbito de investigación educativa. Campbell y Stanley (1963: 171-172 y 176-182) reflexionan sobre las limitaciones de estudios que pretenden llegar a conclusiones causales con base en la observación de un solo caso, aunque sea con mucho detalle, y observando lo que ocurría antes de cierto tratamiento y después del mismo, y subrayan la importancia de tener al menos un punto de comparación. Los diseños experimentales buscan atender esta recomendación; algunas variantes de esta familia de diseños son las siguientes.

Algunos diseños experimentales

▪ **Diseño de dos grupos solamente con prueba posterior (postest)**

Se forman dos grupos con asignación aleatoria; en uno (grupo experimental) se aplica

el tratamiento (intervención) y en el otro (grupo control) no; se observa lo que pasa con la variable que se quiere explicar en ambos grupos.

Los lectores que tengan idea de los diseños experimentales echarán de menos un elemento que se suele considerar necesario: el que consiste en la observación previa (*pretest*) de la situación de la variable, cuyo efecto se quiere analizar. Campbell y Stanley señalan que la prueba previa no es esencial en un experimento estricto, ya que la asignación aleatoria de sujetos a los dos grupos aseguraría por sí sola su equivalencia en cuanto a toda variable, incluyendo la que se quiere explicar (1963: 195-196).

▪ Diseño de dos grupos con pretest y postest

Se forman dos grupos con asignación aleatoria de sujetos; se observa la situación de la variable que se quiere explicar en ambos; se lleva a cabo la intervención o tratamiento, solo en el grupo experimental; finalmente se observa la situación de la variable a explicar en ambos grupos.

Este es el diseño experimental básico, que puede presentar un problema: el hecho mismo de observar la situación de los dos grupos antes de la intervención puede afectar el resultado (Campbell y Stanley, 1963: 183-194).

▪ Diseño de cuatro grupos de Solomon

Para atender ese problema, en este diseño se forman cuatro grupos con asignación aleatoria de sujetos; dos grupos serán experimentales y dos de control. Se realiza la observación previa de la variable a explicar, solamente en dos de los grupos, pero no en los otros dos. Luego se realiza la intervención en dos grupos (experimentales), uno en el que se haya hecho observación previa, y otro en el que no se haya hecho. En los otros dos grupos (control) obviamente no se hace la intervención. Por último, se observa la situación de la variable a explicar en los cuatro grupos.

De esta manera los posibles efectos de la observación previa se podrán detectar comparando los dos grupos en los que se hizo dicha observación con los dos en los que no se hizo. El impacto del tratamiento de detectará comparando los dos grupos experimentales con los dos grupos control, teniendo en cuenta, cuando sea el caso, el efecto de la observación previa.

Aunque obviamente es más costoso, este diseño corrige las posibles limitaciones de los dos anteriores, por lo que se le llegó a considerar el diseño ideal, el *estándar de oro* de la investigación (Campbell y Stanley, 1963: 194-195).

▪ **Diseños con tratamientos alternativos**

Cuatro décadas más tarde Shadish, Cook y Campbell ya no dan tal importancia al diseño de Solomon, pero mencionan otras opciones. Una es un diseño en el que la intervención no se reduce a la aplicación de un tratamiento, sino que se manejan varios tratamientos distintos, lo que implica formar un grupo experimental para cada uno. Se puede comparar dos o más grupos que reciben tratamientos distintos, además del grupo control que no obtiene tratamiento alguno, pero también es posible no incluir grupo control, sino solamente los que adoptan tratamientos alternativos, que funcionan como controles entre sí. (2002: 261-263).

▪ **Experimentos factoriales**

Otra posibilidad es la que consiste en que el tratamiento implique la modificación no solo de una variable (o factor), sino de dos o más, con dos o más posibilidades en cada caso. Shadish, Cook y Campbell dan un ejemplo relativo al impacto de un programa de tutoría en algún resultado de los alumnos, considerando por una parte la duración de la tutoría (una hora o cuatro en cierto período), y por otra quién sería el tutor (un compañero o un adulto). En este caso las opciones de cada factor deben combinarse con las del otro, habiendo varias formas de hacer esa combinación, con los diseños denominados anidados o cruzados. (2002: 263-266)

▪ **Experimentos longitudinales**

Otra variante importante es la que consiste en los experimentos longitudinales, que toman en cuenta el que el paso del tiempo, por sí mismo, puede afectar de manera importante la(s) variable(s) sobre las que se quiere estudiar el efecto de cierto tratamiento. En este caso pueden hacerse varias observaciones *antes* de aplicar el tratamiento, y *varias* después de hacerlo, de manera que su posible impacto pueda distinguirse de cambios debidos a otros factores, asociados con el paso del tiempo, teniendo en cuenta que el efecto de una intervención no es instantáneo, sino que puede llevar tiempo, y también que el efecto se puede desvanecer más o menos rápidamente. (Shadish, Cook y Campbell, 2002: 266-268)

Teóricamente es clara la ventaja de los diseños experimentales estrictos en lo que se refiere a sustentar interpretaciones causales de las relaciones que se observen entre ciertas variables, pero estos diseños tienen también debilidades importantes, en particular en la investigación educativa y social. Además de la dificultad práctica que supone la asignación aleatoria de sujetos al grupo experimental y control, y de las implicaciones

éticas, hay problemas técnicos que tienen que ver, por ejemplo, con fidelidad de la implementación y maduración.

En el Cap. 4, al tratar del análisis de datos derivados de trabajos experimentales, se podrá ver la manera en que los autores que se sigue en este inciso sistematizan las diversas amenazas que deben enfrentar estos diseños, y las formas en que proponen hacerlo. (Shadish, Cook y Campbell, 2002: 64-102)

Planeación de un estudio experimental

El atractivo teórico de los experimentos, como forma ideal de sustentar atribuciones causales, se contrapone a la dificultad práctica de reunir las condiciones que supone este tipo de diseño, como el que sea posible asignar aleatoriamente los sujetos o participantes a los grupos, que sea factible manipular las variables cuyo efecto se quiere explorar, o que la intervención no distorsione el proceso en cuestión.

Por lo que se refiere a la primera dificultad, en el campo educativo es frecuente que solo sea posible asignar aleatoriamente a cierto tratamiento escuelas o aulas completas, lo que en principio no es inadecuado, pero tiene la consecuencia de que el tamaño de la muestra se reduce mucho.

Imaginemos un estudio sobre un nuevo plan de estudios, en el que participan 1200 estudiantes, 600 de los cuales llevan el nuevo currículo y los 600 restantes el antiguo. Será muy diferente si no se asignan aleatoriamente los alumnos en lo individual, sino por aulas, considerando 30 de 40 alumnos cada una, con 15 aulas en el grupo que trabajará con el nuevo currículo y 15 con el anterior. El tamaño de muestra para los análisis que se hagan no será 1200, lo que sería suficiente para los análisis más complejos, sino 30, lo que difícilmente bastará aún para los más simples. (Vogt, Gardner y Haeffele, 2012: 163)

Los diseños experimentales pueden ser fuertes en validez interna —la solidez de las inferencias causales— pero suelen ser débiles en cuanto a validez externa, ya que por lo general sus hallazgos no pueden generalizarse más allá de los participantes.

Cuasi-experimentos

Un cuasi-experimento es una alternativa a un experimento estricto, por la dificultad de asignar los sujetos en forma aleatoria, lo que impide cumplir el tercer requisito de una relación causal, excluir explicaciones alternas; los otros dos —que haya correlación, y que la causa hipotética preceda al efecto— pueden cuidarse tanto en un experimento como en un cuasi-experimento. En vez de asignar aleatoriamente sujetos, en un cuasi-experimento se descartan las explicaciones alternas aplicando otros principios, como identificación

y análisis de amenazas plausibles a la validez interna; primacía del control mediante el diseño del estudio; comparación de patrones de relaciones coherentes con hipótesis causales complejas, que es improbable que satisfagan las explicaciones alternativas. Hay variantes de cuasi-experimentos análogas a las de los experimentos: sin grupo control; con grupo control pero sin prueba previa; o con grupo control y prueba previa. Shadish, Cook y Campbell presentan ejemplos (2002: 103-134 y 135-170).

Regresión con discontinuidad (regression discontinuity, RD)
El diseño más cercano a un experimento estricto es la *regresión con discontinuidad*, que se distingue porque usa una aleatorización implícita. Puede emplearse cuando los grupos a comparar se forman asignando sujetos con base en sus puntajes en una variable medida previamente, que tienen un margen de error, por lo que puntajes ligeramente arriba o abajo de un *valor de corte*, dentro del margen de error, se pueden considerar equivalentes, y la posición exacta de un sujeto arriba o abajo del punto de corte se debe en realidad al azar. Los grupos formados por quienes estén arriba o abajo del valor de corte se pueden considerar equivalentes, como si se hubieran formado aleatoriamente. Es posible aplicar un tratamiento a un grupo y no al otro, y los resultados se pueden analizar como los de sujetos de un grupo experimental y uno control estricto. (Shadish, Cook y Campbell, 2002: 207-245)

Diseños con grupos similares por emparejamiento
Una forma menos rigurosa de asegurar equivalencia es que el grupo experimental y control se parezcan al máximo en posibles variables explicativas. Si se hipotetiza que un método de enseñanza no solo está asociado con mejores resultados de aprendizaje, sino que los *produce*, se pueden controlar variables como experiencia del docente o tamaño del grupo, y descartar que la influencia causal se deba a variables de los alumnos, asegurando que en los grupos experimental y control haya la misma proporción de niños y niñas, de cierta edad o nivel socioeconómico, etc. La expresión *Propensity Score Matching* designa la estrategia de formar grupos *emparejados* sistemáticamente en el *puntaje* de variables que puedan marcar tendencia (*propensión*) en resultados. (Shadish, Cook y Campbell, 2002: 118-122)

Experimentos naturales
Se presentan cuando el grupo *experimental* se forma con sujetos que tienen en común una variable que podría explicar otras, sin intervención de los investigadores. Si se quiere

estudiar el efecto de la desnutrición en el aprendizaje, razones éticas impedirían formar un grupo en que se produzca artificialmente desnutrición, pero es posible encontrar niños que presenten desnutrición sin intervención de los investigadores. Se podrían así formar grupos experimentales y de control *naturales*.

El estudio del cambio

Los diseños experimentales y cuasiexperimentales tienen en común el propósito de sustentar explicaciones de un fenómeno en términos causales, lo que no pueden hacer encuestas, estudios de casos y de observación visual, los diseños llamados genéricamente *observacionales*.

El diseño de *series de tiempo* busca sustentar atribuciones de causalidad sin asignación aleatoria, con base en el estudio de sujetos a lo largo del tiempo, observando si cuando ocurre un cambio en un aspecto (que puede verse como tratamiento), se detecta un cambio significativo en otra(s) variable(s), que se puede interpretar como efecto del primero.

Cuando los datos de una investigación se recogen considerando un solo momento, el estudio se define como *transversal* (*cross-sectional*), en contraposición a estudios que utilizan datos obtenidos en varios momentos, que se definen de manera general como *longitudinales*, debiendo distinguir los estrictamente longitudinales, que obtienen datos de los mismos sujetos en varios momentos a lo largo del tiempo, y los que se aproximan a los anteriores recogiendo datos en un solo punto del tiempo, pero que buscan información de varios momentos. (Lietz y Keeves, 1997; Keeves, 1997)

Se distinguen tres diseños para estudiar cambios: de panel, de cohorte (llamados también de tendencias) y transversales simultáneos, a los que añaden los de análisis de la historia de los eventos (*Event-History Analysis*, EHA). (Vogt, Gardner y Haeffele, 2012: 23; Keeves, 1997: 141-143)

El concepto de *cohorte* de edad (*birth cohort*) denota un grupo de personas nacidas en un mismo lapso de tiempo, por lo que su edad aumenta en paralelo. Algunos historiadores utilizan el concepto de *generación*, y el término se usa desde hace mucho en demografía. (Ryder, 1965)

Al estudiar cambios en el tiempo se deben distinguir tres tipos de efecto:

- Efecto de la edad: que una persona sea joven o vieja afecta sus opiniones, actitudes y comportamiento. Es probable que personas de 75 años sean más conservadoras o más religiosas que las de 50, y estas que las de 25.

- Efecto de la cohorte: personas de la misma edad nacidas en distintas fechas pueden tener creencias, actitudes o comportamientos diferentes; las nacidas en 1945 (que, si pudieron, fueron a la escuela en tiempo de Ruiz Cortines y López Mateos, llegaron a 25 tras el movimiento estudiantil de 1968, y a los 50 al comenzar la alternancia democrática), posiblemente sean diferentes, a la misma edad, de las nacidas en 1970, que fueron a la escuela en tiempo de López Portillo y de la Madrid, cumplieron 25 al inicio de la alternancia democrática, y cumplirán 50 en la presidencia de López Obrador.
- Efecto del período en que se recoge la información: personas de igual edad y la misma generación o cohorte tendrán opiniones, actitudes y conductas diferentes en un momento de bonanza y tranquilidad, en contraste con las que manifestarán en una época de crisis económica o de seguridad.

El reto que enfrenta un estudio del cambio es distinguir el impacto de estos efectos. En estudios transversales simultáneos se mantiene constante el período y la edad varía; en los de tendencia es constante la edad y cambia el período; en las series de tiempo se estudia una sola cohorte y la edad y el período pueden variar. (Keeves, 1997: 144; Vogt, Gardner y Haeffele, 2012: 22-23).

Series de tiempo

Según Shadish, Cook y Campbell, una serie de tiempo consiste en *largas series de observaciones de la misma variable a lo largo del tiempo*, pudiendo tratarse de observaciones de los mismos sujetos observados repetidamente, pero también de observaciones de sujetos distintos pero similares. (2002: 172)

Un tipo de especial interés, por su potencial para detectar efectos casuales, es el de las series de tiempo interrumpidas (*Interrupted Time-Series,* ITS). En estas en cierto punto de la serie de observaciones se modifica algo (*se dio un tratamiento*, lo que constituye una *interrupción*), tras lo cual se observa un cambio en la variable dependiente, que puede ser interpretado como resultado o efecto.

La serie de observaciones puede mostrar, por ejemplo, que antes de la interrupción las mediciones del posible resultado tenían cierto nivel general (intercepto) y cierta tendencia, plana o más o menos rápidamente creciente o decreciente (pendiente).

Si tras la interrupción se observa un cambio anómalo en el intercepto y/o en la pendiente, será razonable atribuirlo al cambio en el que consistió la interrupción, y no a cualquier otra causa, y esto con mayor seguridad cuanto más larga sea la serie de

observaciones. El posible efecto puede ser brusco o paulatino, transitorio o permanente, y puede haber aspectos que impidan detectarlo con claridad, por ejemplo, si el cambio mismo que se considera tratamiento o interrupción no ocurre en un momento preciso, sino que se extiende por cierto tiempo.

Una variante de este diseño consiste en que, después de cierto tiempo, el cambio introducido se suspende; la interrupción se interrumpe, lo que permite observar si el presunto efecto también se revierte, lo que robustecería la interpretación causal.

Entre otras variantes que presentan Shadish, Cook y Campbell (2002: 171-206), se pueden introducir y suspender interrupciones varias veces; trabajar con dos grupos que reciban el tratamiento en distintos momentos, de manera que uno sirva como control del otro en cada ocasión; manejar como grupo control otra serie en la que no se haga la interrupción; o incluir más de una variable dependiente.

Estudios de cohortes o de tendencias

En vez de seguir en momentos sucesivos a un mismo grupo de sujetos, se sigue a muestras diferentes, pero representativas, de sujetos de una misma cohorte (*v.gr.* nacidos en 1998), recogiendo información en varios momentos, *v. gr.* los resultados en pruebas de rendimiento de 3° de primaria de 2006; en las de 6° grado de 2009, y en las de 3° de secundaria de 2012. El estudio es longitudinal en el sentido de que se recogen datos en varios momentos, pero no lo es estrictamente porque no se trata de los mismos sujetos. (Keeves, 1997: 141-143)

Sin embargo, como el intervalo entre las aplicaciones de las pruebas es de tres años, la mayoría de los niños nacidos en 1998 estará en los grados estudiados en los años mencionados; si la muestra está bien hecha, los sujetos considerados en el estudio no serán los mismos (no es un estudio de panel), pero serán similares, pues serán representativos de su cohorte. El error de muestreo podrá ser importante y deberá cuidarse, pero probablemente sea menor que el error debido a la pérdida de sujetos en un estudio de panel, dada la dificultad de seguir a un mismo grupo a lo largo de seis años. (Vogt, Gardner y Haeffele, 2012: 23-24)

Estudios transversales simultáneos

En este caso se recaba información en un solo momento, pero de varias cohortes de distinta edad, y se considera equivalente a la que se obtendría siguiendo una misma cohorte a lo largo del tiempo. Si se estudia en un mismo momento el nivel de aprendizaje alcanzado por muestras de estudiantes de distintos grados, como hace la *International*

Association for the Asessment of Educational Achievement (IEA) en los estudios TIMSS y PIRLS, las cohortes a que pertenecen los alumnos evaluados en cada grado considerado son diferentes, pero con la debida cautela se les puede considerar como equivalentes, si no ha habido cambios drásticos en la sociedad y el sistema educativo de que se trate. Por ello se les puede considerar como si fueran una sola cohorte, *sintética o aparente*. (Keeves, 1997: 140-141)

Estudios de panel

Estos estudios son los que se consideran estrictamente longitudinales, porque en ellos se recoge información de los mismos sujetos (el panel) varias veces a lo largo del tiempo. Combinando elementos de los diseños de cohorte y los transversales simultáneos, en este caso se siguen varias cohortes en varios momentos, gracias a lo cual es posible distinguir los efectos de edad, cohorte y período de recolección de información, para lo que desde 1965 se desarrolló un modelo estadístico que permite dar cuenta de cada efecto y de las interacciones de las combinaciones de dos efectos (Schaie, 1965, según Keeves, 1997: 144)

La tabla siguiente muestra el diseño de un estudio de tipo panel, en el que se daría seguimiento a personas de cinco cohortes distintas (nacidas entre 1980 y 2000) cuando tuvieran 5, 10, 15, 20 y 25 años, o sea cada cinco años, entre 1985 y 2025.

TABLA 2.3. DISEÑO DE PANEL CON LA EDAD DE LOS MIEMBROS DE CINCO COHORTES DE LAS QUE SE OBTENDRÁN DATOS CINCO VECES, EN INTERVALOS DE CINCO AÑOS

Fecha	**Fecha de la medición (en las casillas la edad de los sujetos)**									
nacim.	**1980**	**1985**	**1990**	**1995**	**2000**	**2005**	**2010**	**2015**	**2020**	**2025**
1980	0	5	10	15	20	25				
1985		0	5	10	15	20	25			
1990			0	5	10	15	20	25		
1995				0	5	10	15	20	25	
2000					0	5	10	15	20	25

FUENTE: KEEVES, 1997: 144, FIGURA 1

Es claro que estudios de esta naturaleza implican costos muy considerables, por lo que no son usuales, pese a sus ventajas teóricas.

Análisis de la historia de eventos (Event history análisis, *EHA*)

Seguir a ciertos sujetos en el tiempo es complicado, por lo que los diseños longitudinales por lo general recogen información solo en algunos momentos, como cada año, o cada cinco años, lo que tiene la desventaja de que entre una y otra recolección de datos pueden pasar cosas que no se detectan.

El enfoque de historia de eventos atiende este problema, con la ventaja adicional de que la información se recoge en un solo momento, aunque la estrategia enfrenta otras limitaciones.

En estos estudios se pide a los sujetos que recuerden y refieran sucesos ocurridos en distintos momentos del pasado, para reconstruir de alguna forma el proceso por el que se llegó a la situación presente. (Vogt, Gardner y Haeffele, 2012: 25)

Es posible estudiar el conjunto de la trayectoria laboral de una persona de 65 años que se entrevista después de jubilarse. Se le puede preguntar desde cuál fue su primer trabajo, cuándo lo consiguió, cuánto tiempo lo había buscado, cuánto duró en él, cuándo lo dejó y por qué, cuánto tiempo tardó en conseguir otro empleo, y así sucesivamente hasta el último.

Lo mismo puede hacerse sobre la historia sanitaria de una persona, o sobre su trayectoria escolar, indagando si fue a preescolar, a qué edad y cuántos años, en qué escuela, y luego en cuanto a primaria, secundaria y media superior, entre otras cosas.

En un estudio en el que se entreviste a las personas cada cinco años se perderá mucho de lo que ocurra entre una ocasión y otra, lo que se evita con el Análisis de la Historia de Eventos; pero en este caso la calidad de la información depende solo de la memoria de cada sujeto, que no siempre es confiable, como han mostrado diversos estudios, incluso tratándose de eventos de gran importancia, como intervenciones quirúrgicas o despidos laborales.

Debe distinguirse, además, el caso de eventos, más o menos importantes, que ocurren en un momento preciso, como los mencionados en los ejemplos anteriores, (eventos discretos), de aquellos que se desarrollan de manera ininterrumpida a lo largo de cierto tiempo (eventos continuos). (Willet y Singer, 1997: 513-519)

Estudios de métodos múltiples

Hasta ahora se han descrito seis diseños de investigación, todos de enfoque básico y que requieren obtener información no recabada antes; cada uno tiene variantes, y aún no se han considerado los diseños basados en información previamente obtenida, ni los de orientación aplicada. Para emprender un estudio hay, pues, una gama de posibilidades, y

ninguna es absolutamente superior a las demás: "Toda pregunta de investigación puede abordarse de múltiples maneras, cada una de las cuales tiene ventajas y límites". (Vogt, Garner y Haeffele, 2012: 49)

Es posible combinar varios diseños para responder una pregunta desde distintos ángulos. También es posible —y en temas complejos deseable— formular varias preguntas, cada una de las cuales podrá ser abordada con más de un diseño, por lo que se puede optar por combinar varios de los diseños anteriores.

Algunos autores sugieren que esto es indicado cuando se quiere corroborar, elaborar o aclarar los resultados que se han obtenidos con un diseño, seguir desarrollando el conocimiento sobre el tema a partir de ellos, abordándolo desde nuevos ángulos, identificando hallazgos inesperados o posibles contradicciones, para contar la historia completa e incluso llegar a desarrollar una teoría al respecto. (*Cfr*. Vogt, Gardner y Haeffele, 2012: 107, Tabla 6.2)

Para designar este diseño combinado se habla de *métodos mixtos*. Siguiendo un sugerente dicho de Frederick Erickson (*you mix drinks, but don't mix methods*), se prefiere la expresión *métodos múltiples*.

Según Erickson la expresión *mixed methods* (métodos mixtos o mezclados) lleva a pensar que en un mismo estudio se deben combinar las características de varios acercamientos diferentes, con el resultado de que no se atiendan adecuadamente las exigencias de cada uno. La expresión *métodos múltiples*, en cambio, transmite la idea de que un trabajo complejo puede comprender varios trabajos particulares, cada uno de los cuales utilice un acercamiento, respetando íntegramente las exigencias derivadas de sus características, y aprovechando los resultados de todos para conformar una visión más completa de los fenómenos que se estudien.

Dos ejemplos de combinaciones clásicas de diseños son:

- Iniciar con una encuesta, para tener una visión amplia, pero algo superficial, y continuar con un estudio de caso para profundizar; o bien,
- Iniciar con un estudio de caso para identificar variables importantes y estudiarlas en una encuesta con una muestra que pueda sustentar generalizaciones.

Los diseños combinados se podrán manejar secuencialmente o en paralelo; puede predominar uno y los otros tener menor importancia, o tener el mismo nivel; pueden ser conducidos por el mismo investigador (o el mismo grupo), o bien por personas o equipos distintos.

En todos los casos una noción importante es la de triangulación, o sea la contrastación de los resultados de un diseño con los del otro u otros, para corroborarlos o cuestionarlos. (Vogt, Gardner y Haeffele, 2012: 106-111)

La posibilidad de combinar diseños de investigación distintos para abordar un tema implica rechazar la oposición total entre un enfoque intensivo y uno extensivo, o uno cuantitativo y otro cualitativo, y consideran en cambio que es más ventajoso combinar varios acercamientos.

Tashakkori y Teddlie distinguen dos *períodos*, en un proceso iniciado hacia 1960, después de una larga etapa en la que solo se utilizaban enfoques cuantitativos:

- Emergencia de *métodos* múltiples, distinguiendo:
  - Combinación de diseños *Cuanti-Cuali* de estatus equivalente, sea en forma secuencial o en paralelo, iniciando cualquiera de los dos.
  - Combinación con un diseño dominante y uno subordinado, también en forma secuencial o en paralelo, iniciando cualquiera.
- Emergencia de *modelos* múltiples:
  - Aplicación única de técnicas *Cuanti-Cuali* en cada fase del estudio.
  - Aplicación múltiple de técnicas *Cuanti-Cuali* en cada fase. (1998: 15)

Otras dos perspectivas sobre el tema de los diseños mixtos pueden verse en Bericat (1998), y Gorard y Taylor (2004).

El recuadro siguiente presenta sintéticamente un ejemplo de que lo que puede ofrecer una combinación más compleja de diseños: un trabajo de Richard Nisbett y Dov Cohen (1996), sobre la posible explicación de los niveles particularmente elevados de violencia que hay en algunas regiones del sur de los Estados Unidos.

Las explicaciones de sentido común del fenómeno lo han atribuido a factores como clima cálido, presencia de descendientes de esclavos, pobreza y desigualdad, o concentración de población en grandes ciudades, todo lo cual estaría más presente en regiones del sur del país donde se registran niveles de violencia altos. La debilidad de estas explicaciones se puede detectar mostrando que la intensidad de la violencia varía entre regiones de manera compleja, y es diferente según el tipo de manifestaciones de que se trate. Al parecer el tipo de delitos que tiene más presencia en algunas zonas sureñas es el que tiene que ver con reacciones a lo que las personas involucradas consideran atentados a su honor.

El trabajo de Nisbett y Cohen cuestiona esas explicaciones y busca unas más sólidas, a partir de referentes teóricos de la psicología social y la antropología culturalista

sobre la llamada *cultura de honor*, que se ha encontrado en regiones de España, Italia y Grecia, las tribus kaibyles de Argelia, Irlanda, y otros lugares que tienen en común que la población se dedica a la ganadería o el pastoreo, y que no hay un Estado sólido cuyas fuerzas policiales garanticen la seguridad. En un lugar así la supervivencia implica saber defenderse, y la socialización de un niño incluye que aprenda desde chico a hacerlo, aplicando la violencia que haga falta, en vez de aprender "a resolver diferencias mediante el diálogo", como se busca en contextos con concepciones de la convivencia más acordes con las ideas actuales. El estudio de referencia muestra que esa socialización se traduce en las conductas adultas que caracterizan la subcultura de algunas regiones del sur estadounidense.

El Recuadro 2.2. permitirá tener una idea de la riqueza del trabajo.

RECUADRO 2.2. LA CULTURA DE HONOR EN EL SUR DE ESTADOS UNIDOS

- Se cuestionan explicaciones usuales analizando estadísticas: tasas de homicidios en el sur y otras regiones, controlando clima, nivel de ingreso y % de descendientes de esclavos; distinguiendo muertes relacionadas con asaltos o resultado de discusión; delitos cometidos por blancos, o con víctimas blancas en ciudades chicas, medianas y grandes. Conclusión: esas explicaciones usuales no resisten la prueba de los datos.
- Se exploró diferencia entre habitantes de norte y sur en actitudes ante violencia, como posibilidad de castigar físicamente a niños en el hogar y en la escuela, la gravedad de los insultos, etc. Con análisis secundarios de encuestas de opinión previas, se comparó % de personas que manifestaban acuerdo o desacuerdo respecto a si se justifica matar en defensa propia; si la policía solo debe estar vigilante o intervenir incluso disparando a matar para controlar tumultos; la preferencia por alternativas de reacción ante agresiones, ofensas a la esposa o novia, castigos aceptables para corregir a los niños, la posesión de armas, etc. Se constató la clara preferencia de residentes de regiones sureñas por las opciones que implican mayores dosis de violencia.
- *Etnografía experimental* para explorar reacciones emocionales ante un insulto: se informó a los sujetos que debían dar muestras de saliva después de llenar un cuestionario en un escritorio al fondo de un corredor estrecho por el que apareció un sujeto corpulento que caminaba rápidamente. Se observó la reacción de los sujetos: la distancia a la que se apartó del camino, medidas de su actitud corporal y su autoestima, niveles de cortisona y testosterona en saliva. En todos los casos las reacciones de los sureños se asocian con una reacción violenta.
- Se exploró la dimensión social de la *cultura de honor,* reflejada en disposiciones legales. Se encontraron diferencias importantes: leyes de control de armas más estrictas en estados del norte; proporción de representantes y senadores que vota a favor de control de armas superior en el norte; diferencia en disposiciones legales sobre la reacción si un intruso entra al propio domicilio, incluso considerar adecuado matarlo; en el norte se castiga con cárcel la agresión de género, en el sur no; distinto peso de violencia doméstica cuando una corte decide sobre custodia de los hijos si los cónyuges se separan; diferencia en la tolerancia del castigo corporal

en las escuelas; gran diferencia sobre aceptación legal de la pena de muerte por ciertos delitos y sentenciados ejecutados.

- *Experimentos sociales* para probar hipótesis sobre diferencias *culturales* entre estados:
  - Cartas de solicitud de empleo a ciudades de todo el país diciendo que el firmante planeaba ir a vivir a esa ciudad y buscaba trabajo. En la mitad de las cartas añadía que acababa de salir de la cárcel, tras purgar una condena por un homicidio de honor; en las demás que se había debido al robo de un automóvil. Las respuestas del sur mostraron aceptación de quien decía haber delinquido por motivos de honor.
  - Nota a periódicos estudiantiles de universidades de todo el país, describiendo el homicidio cometido por un joven cuya novia había sido seriamente ofendida por la víctima. Se analizó la forma de informar el suceso en cada periódico, si destacaban detalles agravantes o atenuantes. Los periódicos del sur mostraron un tratamiento de mayor comprensión para el homicida que los periódicos de otras regiones.

FUENTE: NISBETT Y COHEN, (1996)

Investigaciones básicas documentales

Los diseños de este grupo coinciden con los del grupo anterior en que su carácter *básico*: su propósito no incluye producir cambios en la realidad estudiada, sino que directamente buscan conocerla mejor, sin descartar que en algún momento los hallazgos se puedan aprovechar para promover cambios. Difieren de los diseños anteriores en que no implican obtener información empírica nueva, sino que usan datos recabados previamente para otra(s) investigación(es), o bien por razones administrativas u otras, que se conservan en algún tipo de archivos o registros.

Vogt, Gardner y Haeffele se refieren de manera general a los diseños de este grupo como *diseños de archivo* (*Archival Designs*), y observan que otros se refieren a ellos con la expresión *análisis secundario de datos* (*Secondary Data Analysis*). Esos autores añaden que toda clasificación es discutible, pero que la diferencia de estos diseños respecto a los del apartado 2.1 es importante, y que captura diferencias reales en el oficio del investigador. (Vogt, Gardner y Haeffele, 2012: 86)

Compartiendo esta idea, en esta obra los diseños de este apartado se designan con la expresión *investigaciones documentales*, que se opone a la de *investigaciones vivas*, teniendo los dos grupos carácter básico. En seguida se describen cuatro diseños particulares de este grupo.

Investigación de archivo

El primer diseño de este grupo, para el que se retoma la expresión *investigación de archivo* en sentido más preciso, se distingue de otros en que la información que se utiliza en él

consiste en documentos de diversos tipos, conservados en repositorios más o menos organizados, a cargo de particulares o instituciones, de acceso abierto o restringido, y de carácter privado o público.

En México, los más conocidos, de especial importancia para investigación histórica, incluyen, desde luego, al Archivo General de la Nación y similares estatales y municipales, pero hay muchos más:

- Archivos de instituciones públicas como secretarías o ministerios; del poder legislativo y del poder judicial.
- Archivos parroquiales y de notarías, archivos de sindicatos o partidos, y de todo tipo de asociaciones.
- Archivos con documentos administrativos y legales de una fábrica, un comercio o una hacienda agrícola.
- Diarios personales o archivos en los que se conserva la correspondencia privada del propietario.

La información que se conserva en archivos suele ser principalmente de tipo textual (por ejemplo, actas, cartas y testamentos, entre otros), pero también puede incluir documentos con datos numéricos, como libros de cuentas, nóminas o cuentas de raya. Por ello en este tipo de diseños predominan los que no incluyen análisis estadísticos, aunque también pueden hacerse, si hay suficiente información.

Análisis secundario de datos

Este diseño usa también datos previamente obtenidos almacenados en repositorios, pero en este caso básicamente numéricos, en bancos de datos digitalizados.

Aunque en México estos bancos no son tan numerosos como los de países de larga tradición en este sentido, como Estados Unidos (*v.gr. National Center for Education Statistics*, NCES) o el Reino Unido (*Department for Education*) su número aumenta y su importancia es cada vez mayor. En forma destacada deben mencionarse las bases de datos del Instituto Nacional de Estadística y Geografía (INEGI), que no solo incluyen las derivadas de los censos nacionales de población que se levantan cada 10 años, y los conteos rápidos quinquenales, sino también las de censos especializados como los de los ámbitos agropecuario y económico, y además las bases de datos que generan, con distinta frecuencia, encuestas sobre temas especializados, como las de ingresos y gastos de los hogares, las de empleo, las de confianza de los consumidores, o las de percepción de la inseguridad.

En el campo educativo existen desde hace mucho tiempo bases de datos de la Secretaría de Educación Pública, como las derivadas de la aplicación del formato conocido como 911, con información básica sobre alumnos, maestros y otros datos, de todas las escuelas e instituciones educativas del país. A partir de 2004 hay un creciente número de bases de datos del Instituto Nacional para la Evaluación de la Educación, que no solo incluyen la información derivada de las pruebas nacionales e internacionales que el INEE aplica cada año a muestras de alumnos de diversos grados de la educación obligatoria, sino también bases de datos de un amplio conjunto de indicadores sobre otros aspectos del sistema educativo. Hasta hace poco el acceso a muchas bases de datos era difícil, lo que ha cambiado gracias a las disposiciones legales sobre acceso a la información pública gubernamental.

Aunque algunas bases de datos incluyen imágenes de documentos (por ejemplo de tareas o cuadernos de alumnos), o videograbaciones de clases, en la mayoría de los casos se trata de bases con información fundamentalmente numérica, y es frecuente que sea el resultado de operativos de campo de carácter censal o con muestras grandes, por lo que el tipo de estudios que se pueden hacer consiste sobre todo en análisis estadísticos de grandes números de sujetos, lo que se presta para que se hagan análisis complejos, sobre todo multivariados.

RECUADRO 2.3. ALGUNAS BASES DE DATOS

NCES: Estadísticas educativas https://nces.ed.gov/nationsreportcard/
UK https://www.gov.uk/government/organisations/department-for-education/about/statistics
IEA: Bases de TIMSS, PIRLS, Civics https://www.iea.nl/data
OCDE: PISA https://www.oecd.org/pisa/data/ Indicadores y TALIS https://stats.oecd.org/
INEGI: Censos y encuestas https://www.inegi.org.mx/datos/default.html#Microdatos
CONEVAL: Medición de pobreza https://datos.gob.mx/busca/organization/coneval
INEE: Indicadores y evaluaciones https://www.inee.edu.mx/evaluaciones/bases-de-datos/
La reforma constitucional que suprimió al INEE establece que sus bases de datos seguirán estando a disposición del público.

Análisis de evidencias

Un tercer tipo de diseño de este grupo se basa en información inferida a partir del análisis de algunos productos de la actividad de los sujetos a estudiar. El supuesto clave de estos estudios es que, como los sujetos no pretendían que los productos de su actuar se utilizaran para explorar su conducta, no hay riesgo de que intentaran engañar sesgando

intencionalmente los productos de las actividades que se quiere estudiar. Por ello, el análisis de evidencias es una tercera opción para la obtención de información, frente a las técnicas de interrogación y las de observación, y ha dado lugar incluso a este tipo de diseño.

Antecedente de este tipo de estudios es el análisis de huellas dactilares, usado hace tiempo en las investigaciones criminológicas, ya que al ser únicas las huellas de cada persona, encontrar las de alguien en un objeto del lugar en que se cometió un crimen, o en el arma con que se cometió, puede ser una prueba sólida de que esa persona estuvo allí, o manipuló esa arma, aunque nadie la haya visto, y aunque la persona sospechosa lo niegue.

Otro ejemplo más cercano al terreno de la investigación social es el que consiste en el análisis de restos en la basura. Si se interroga a una persona sobre su consumo de alcohol, o sus hábitos alimenticios, es posible que no responda en forma veraz, pero si en la basura que se genera en su domicilio hay muchas botellas de licor vacías o latas de cerveza, se pueden hacer inferencias más sólidas al respecto. Es sabido que las técnicas del trabajo de campo que utilizan hoy los arqueólogos pueden dar información sobre la dieta de hombres o animales que vivieron hace miles de años, analizando los restos de alimento acumulados en yacimientos prehistóricos, incluso en cantidades pequeñísimas.

En un campo afín al educativo, se han estudiado las preferencias del público que visita museos a partir del desgaste del piso frente al sitio en que se exhiben ciertas piezas, o las huellas de la nariz que dejan los niños al apoyarse en el cristal de un exhibidor, que incluso permiten estimar la edad de los niños que las dejaron, a partir de la altura de la huella en el cristal. Webb, Campbell, Schwartz y Sechrest (1966) denominan a estas mediciones *no invasivas o de baja interferencia* (*unobstrusive*), y distinguen los estudios de resultados por erosión, y los debidos a acumulación.

En el campo educativo hay un amplio espacio para el diseño de estudios basados en información inferida a partir del análisis de evidencias.

Para evaluar a los maestros se usan *portafolios* con planes de clase, cuadernos de trabajo de los alumnos o tareas que realizan por indicación del docente, registros de clases impartidas y otros materiales, como evidencias a partir de las que pueden inferirse rasgos del trabajo de los maestros mismos, sin necesidad de observarlos en el aula, y sin depender de su propio testimonio, no necesariamente confiable.

Otra variante son estudios de las prácticas de enseñanza con base en inferencias basadas en el *análisis de las tareas* que hacen los alumnos por indicaciones de los docentes, que revelan mucho de las prácticas de los segundos. Más reciente es el desarrollo de

protocolos para estudiar prácticas docentes como los que se describen en el Capítulo 3, sea de enfoque general, sea orientados a las prácticas en alguna de las áreas del currículo.

Síntesis de investigaciones

Cuando las comunidades de investigadores se consolidan, el número de trabajos que estudian un mismo tema se multiplican, y se hace posible preguntarse sobre la consistencia de cada uno, la comparabilidad de los hallazgos, los aspectos del tema que el conjunto de estudios consigue explicar y los que siguen obscuros.

Se acepta, además, que la ciencia es (o debería ser) acumulativa; que cada trabajo nuevo se basa (o debería basarse) en los anteriores, y no partir de cero. Por ello toda investigación comienza típicamente con la *revisión de literatura*, para sintetizar lo que otros han encontrado sobre el tema, y a partir de ello formular las preguntas de investigación o las hipótesis que orientarán el trabajo que se emprende.

En la forma tradicional, la revisión de literatura implicaba la localización de trabajos relevantes, su revisión, y la redacción de un texto que resumiera los hallazgos más destacados y las cuestiones pendientes más importantes, a juicio del revisor, sin que hubiera reglas especiales para elaborar ese tipo de *síntesis narrativa*. El aumento del número de estudios con aproximaciones estructuradas para hacer las observaciones de los aspectos relevantes (para medir ciertas variables), así como para analizar los resultados en forma descriptiva, correlacional o explicativa, univariada, bivariada o multivariada, permitió también que las revisiones de literatura dieran lugar a síntesis cuantitativas de los hallazgos de los estudios previos.

La versión más simple de este tipo de revisiones es la que se ha denominado *de conteo de votos*, que simplemente identifica los estudios cuya conclusión va en un sentido o en el opuesto, o los que no dan resultados concluyentes. Las versiones más elaboradas incluyen cuantificar los resultados obtenidos por trabajos que hayan utilizado un mismo tipo de mediciones de las mismas variables. Una idea clave es que, si bien la muestra empleada en cada estudio pudo ser relativamente pequeña, la muestra a la que se llegará agregando las de cada estudio considerado será mucho mayor, por lo que se podrá llegar a conclusiones mejor sustentadas.

Estas síntesis se designan con el término *meta-análisis*, que se puede entender *simplemente como una revisión sistemática que utiliza técnicas cuantitativas para resumir resultados cuantitativos*. (Vogt, 2007: 306)

Algún tipo de revisión de literatura es un paso necesario para emprender cualquier investigación, pero muchas veces no hay condiciones para hacer una revisión que utilice

técnicas de análisis más complejas, un meta-análisis. En lugares en que la comunidad de investigadores no es muy numerosa, y una proporción considerable de los trabajos no produce datos susceptibles de agregarse, es razonable limitarse a una revisión de literatura convencional, si bien los medios informáticos actuales permiten cubrir una cantidad de publicaciones mucho mayor que la que era posible con los medios biblio-hemerográficos usuales hasta hace no mucho tiempo.

Conviene hacer un *meta-análisis* cuando la cantidad de trabajos de investigación comparables y agregables es importante, sobre todo si los resultados de los trabajos previos no son coincidentes, y el *meta-análisis* puede identificar coincidencias y discrepancias que den lugar a buenas preguntas de investigación. Si los estudios previos no son suficientes no convendrá hacerlo, ni tampoco si recientemente se ha hecho ya uno sobre el tema. (Vogt, Gardner y Haeffele, 2012: 91)

Para diseñar un estudio de síntesis de investigaciones, y una vez que se ha reunido el conjunto de trabajos que se piensa incluir, es necesario responder antes algunas preguntas sobre cada uno de los estudios considerados, para lo que en el recuadro siguiente se propone una lista de cotejo, que se aplica en especial a las síntesis cuantitativas, los *meta-análisis*, pero que se puede adaptar a cualquier revisión de literatura con base en la cual se haga una *síntesis narrativa*.

RECUADRO 2.4. PREGUNTAS A RESPONDER PARA UNA SÍNTESIS DE INVESTIGACIONES

- Preguntas generales

¿Cuál es la hipótesis o pregunta de investigación que orienta el trabajo?
¿Por qué creen los autores que importa estudiar esa hipótesis o pregunta?
¿Qué diseño y qué técnicas de obtención de información usaron los autores?
¿Eran apropiados para responder la pregunta?
¿Cuáles son los principales hallazgos o conclusiones?
¿Son convincentes?

- Preguntas sobre las variables

¿Cuál es la variable dependiente, o resultado?
¿Cuáles son las variables independientes, o predictores?
¿Se identifican variables intervinientes o mediadoras?
¿Deberían estudiarse esas variables intervinientes?
¿Se consideran variables de control?
¿Deberían examinarse otras variables de control?
¿Se discuten posibles variables moderadoras o efectos de interacción?

¿Cómo se operacionalizan (cómo se definen y miden) las variables?
¿Son apropiadas las definiciones y mediciones de las variables?

- Preguntas sobre los sujetos y la muestra

¿Qué sujetos se estudian? ¿Son apropiados dados los propósitos?
¿Cuántos se estudian? ¿Son suficientes dados los propósitos?
¿Es representativa la muestra? ¿Qué tanto se puede generalizar?

- Preguntas sobre las conclusiones

¿Son significativos estadísticamente los hallazgos?
¿Son significativos científicamente los hallazgos?
¿Qué tan grandes son los efectos encontrados?
¿Son significativos prácticamente los hallazgos?
¿Están bien sustentadas las conclusiones por la evidencia que se cita?

- Preguntas finales, a partir de las respuestas a las anteriores

¿Cómo se podría mejorar la investigación?
¿Qué preguntas o problemas deja sin responder el artículo?
¿Cómo podría Usted hacer un trabajo mejor?

FUENTE: VOGT, 2007: 300, FIGURA 17.1.

Para terminar este apartado, se presentan algunas consideraciones generales sobre investigación documental.

Uno de los primeros trabajos en abordar el tema de la comparación e integración de resultados de investigaciones, en la perspectiva hoy conocida como *meta-análisis*, comienza con una cita de la autobiografía de Mark Twain: "La 13ª campanada de un reloj no solo es falsa en sí misma; además suscita graves dudas sobre la credibilidad de las doce campanadas anteriores". (En Light y Smith, 1971: 67)

Los autores comentan:

> La frase de Mark Twain capta una llamativa parte de la experiencia de hacer investigación educativa. Parece que por cada docena de estudios que llegan a cierta conclusión siempre es posible encontrar un decimotercer trabajo que no está de acuerdo. (Light y Smith, 1971: 67-68)

Como seguramente los trabajos previos no son todos de igual calidad, en lugar de desecharlos todos parece preferible revisarlos para identificar los más sólidos, así como los hallazgos en que un mayor número coincida, como se intenta en toda revisión de literatura, y en la forma que emplea técnicas cuantitativas, un *meta-análisis*.

Las razones fundamentales por las que puede ser preferible hacer una investigación documental que una viva son la existencia de fuentes adecuadas, y el que en ellas sea posible encontrar información sobre el objeto de estudio más amplia que la que el investigador podría obtener como resultado de su propio trabajo de campo.

La contrapartida es que la investigación documental no podrá dar más información que la que obtuvieron en su momento quienes la recabaron; si las preguntas de investigación requieren de cierta información diferente, la única forma de conseguirla es hacerlo mediante algún tipo de investigación viva. Por otra parte, la posible complementariedad de los dos grandes grupos de diseños es evidente.

En todos los diseños de este grupo los investigadores deberán atender cuestiones similares, relativas a la identificación de las fuentes de información que se usarán en cada caso; a la selección de los documentos de cada fuente a utilizar en un proyecto particular; a la manera de codificar y registrar la información relevante; y a la forma de analizarla. En todos los casos, y como en cualquier otro diseño, esas cuestiones deberán atenderse en función de las preguntas que guíen la investigación. (Vogt, Gardner y Haeffele, 2012: 86-102)

Para ampliar la información anterior se puede consultar Cooper, 1984; y Cooper y Hedges, 1994.

Investigaciones aplicadas

A diferencia de los dos grupos anteriores, los diseños de este se caracterizan por tener un propósito que no se limita a la ampliación del conocimiento sobre ciertos fenómenos, sino que incluye buscar algún cambio de la realidad estudiada. Este tipo de propósito tiene dos consecuencias que definen lo que tiene de especial la investigación aplicada: por una parte, debe prestar atención a la identificación de causas, ya que debe tener bases para concluir que la modificación que se haya producido se deba efectivamente a lo que se postulaba; por otra, el objeto de estudio no se reduce a una o más preguntas, sino que incluye la precisión de un aspecto de la realidad que no se considera adecuado y se pretende modificar, o sea incluye la precisión de un problema a enfrentar. La investigación aplicada incluye los estudios llamados de intervención, pero no se reduce a ellos. En esta obra se consideran otros dos tipos, que son la investigación evaluativa y la metodológica.

Estudios de intervención

En educación, la investigación aplicada de intervención se refiere en general al campo pedagógico, ya que frecuentemente se propone desarrollar o mejorar el currículo y/o los materiales educativos, las prácticas de enseñanza, entre otros.

La obra editada por Bickman y Rog (1998) presenta elementos que comparten, en general, las investigaciones de este tipo, y también algunas de sus variantes. En el primer capítulo de dicha obra se presenta una visión de conjunto del proceso de planeación de una investigación aplicada, en la que se distinguen dos etapas:

- Definición del problema de investigación, variante de la construcción del objeto de estudio (Cap. 1): se debe explicitar la comprensión que se tiene del problema a enfrentar, lo que incluye el desarrollo de un marco conceptual; luego se deberán identificar preguntas de investigación, en el caso referidas al problema; y finalmente se deberán refinar o revisar dichas preguntas.
- Elaboración de un plan de investigación: elección de un diseño y un enfoque para obtener información; inventario de los recursos disponibles; valoración de la factibilidad del estudio; y determinación de los aspectos que se podrán conservar o se deberán sacrificar. (Bickman, Rog y Hedrick, 1998:6)

Intervenciones controladas

Por centrarse en la búsqueda de sustento para hacer atribuciones causales, los diseños experimentales y los cuasiexperimentales se prestan especialmente para ser utilizados en estudios que no busquen solo incrementar los conocimientos sobre ciertos temas, sino producir algún tipo de cambio en la realidad estudiada, o sea en investigaciones aplicadas. Se entiende así que los primeros diseños de intervención fueran de tipo experimental o cuasiexperimental, con la diferencia de que el típico experimento de laboratorio se refería muchas veces a fenómenos de escasa relevancia práctica, como identificar el efecto causal de cierto tipo de estímulo sobre la cantidad de cifras o de palabras sueltas que un sujeto conseguía memorizar. En el caso de las intervenciones, en cambio, lo que se ponía a prueba eran cuestiones de importancia práctica, como el impacto sobre el aprendizaje de los alumnos de un nuevo método para enseñar lectoescritura o matemáticas.

A este tipo de trabajos se refieren dos capítulos del *Handbook of Applied Social Research Methods*: uno sobre experimentos controlados con aleatorización para evaluación y planeación (Boruch, 1998); otro sobre cuasiexperimentos (Reichardt y Mark, 2012).

Otros diseños también se pueden utilizar en investigación aplicada, como muestra el capítulo de Yin (1998) en la misma obra, que presenta una versión simplificada de los estudios de caso que puede usarse en este sentido.

En un texto referido en particular a los estudios de intervención, Tymms señala:

> Una intervención es un intento deliberado por cambiar de alguna manera la realidad, con la intención de valorar el impacto de tal intervención. Esta se arregla (se diseña) de tal manera que el investigador pueda interpretar los resultados de la intervención en términos causales. La manera más simple de hacer esto consiste en formar dos grupos equivalentes gracias a la asignación aleatoria de sujetos a esos grupos [...] Hay muchas variantes de este tema [...] (2012: 137)

El resto del texto de Tymms describe algunas posibilidades de diseños de tipo experimental, como los que se han presentado antes.

Investigación acción y practitioner research

La primera parte de este inciso se refiere a las que seguramente fueron las primeras intervenciones que no se basaban en diseño experimental, la investigación acción; la segunda parte describe modalidades más recientes de estos trabajos.

La tradición de la investigación acción se remonta a trabajos hechos en el contexto de la escasez de carne durante la Segunda Guerra Mundial. Kurt Lewin estudiaba los cambios de actitudes de las personas, pero a la vez pretendía modificarlas, induciendo a las amas de casa a aprovechar las vísceras de los animales para su alimentación cotidiana, modificando sus hábitos tradicionales.

Poco después de la guerra, y también en el Reino Unido, se conocieron como investigación acción los trabajos realizados por Albert H. Halsey, en relación con las zonas prioritarias para actividades educativas (*Educational Priority Areas*). (Corey, 1953)

En la década de 1970 se desarrolló en América Latina una versión de investigación acción inspirada en trabajos de educación popular de Paulo Freire, con influencia de ideas marxistas radicales, según las cuales lo que determinaría la calidad del conocimiento generado por una investigación no serían elementos metodológicos o teóricos, sino la posición de clase de quienes busquen comprender un fenómeno.

A partir de la idea de que en toda sociedad habría básicamente dos clases sociales, lo que haría la diferencia entre unas investigaciones y otras no sería uno u otro enfoque metodológico, o la utilización de ciertas técnicas de obtención o análisis de la información, sino la adopción del punto de vista de la clase dominante o de la dominada, con la oposición de ciencia burguesa versus ciencia proletaria.

Un reporte sobre proyectos de investigación acción participativa publicado entonces por el *International Council for Adult Education* señala rasgos que los distinguirían:

- El problema nace en la comunidad que lo define, lo analiza y resuelve.
- El objetivo último es la transformación radical de la realidad social y la mejora de la vida de los implicados; los beneficiarios son los miembros de la comunidad.
- La investigación participante exige la participación plena y entera de la comunidad durante todo el proceso de investigación.
- La investigación participante implica un conjunto de grupos y de personas que no poseen el poder: explotados, oprimidos, pobres, marginales, entre otros.
- El proceso de la investigación participante puede suscitar entre los participantes una mejor toma de conciencia de sus propios recursos, y movilizarlos con miras a un desarrollo endógeno.
- Se trata de un método de investigación más científico que la investigación tradicional, en el sentido de que la participación de la comunidad facilita un análisis más preciso y más auténtico de la realidad social.
- El investigador es aquí un participante comprometido; aprende a medida que investiga. Milita en lugar de buscar el distanciamiento. (ICAE, 1977)

En la década de 1990 se identificaban dos maneras de entender la investigación acción, una con autores como John Elliott, que seguía la propuesta de Kurt Lewin y retomaba ideas de Lawrence Stenhouse sobre currículo, y Donald Schön sobre el profesional reflexivo, y otra representada, por ejemplo, por Bud Hall, que adoptaba las ideas de la llamada ciencia social crítica de la Escuela de Frankfurt, cercana a la versión radical latinoamericana. (Kemmis, 1997: 177)

Según la primera versión, la investigación acción se distingue por estos rasgos:

- Busca mejorar la educación transformándolay aprender de los cambios [...]
- Se desarrolla en una espiral auto-reflexiva de planeación, implementación de los planes, observación sistemática y reflexión [...]
- Es participativa [...]
- Es colaborativa [...]
- Involucra a la gente para que teorice su práctica [...]
- Requiere que la gente ponga a prueba sus prácticas, ideas y supuestos sobre las instituciones [...]
- Es abierta en cuanto a lo que cuenta como evidencia o datos [...]
- Permite que los participantes registren sus avances [...]

- Inicia en pequeña escala, con cambios limitados, ciclos cortos y grupos de tamaño reducido, y va creciendo [...]
- Involucra a la gente en la elaboración de análisis críticos de las situaciones en que trabaja, aulas, escuelas, sistemas [...]
- Es un proceso político [...] (Kemmis, 1977:175-176)

Según un representante de la segunda perspectiva, "la investigación participativa es un proceso de acción social inclinado a favor de personas dominadas, explotadas, pobres, o marginadas de otras formas". (Hall, 1997: 198)

El texto de este autor señala expresamente que el origen de esta versión se remonta a los trabajos de campo de Engels, y recientemente a los de científicos sociales radicales del llamado Tercer Mundo, como Gunter Frank y Samir Amin, de Paulo Freire y Orlando Fals Borda, y la relaciona expresamente con las teorías feministas, definiéndola como una práctica contrahegemónica. (Hall, 1997: 198-201)

Aunque hay planteamientos parecidos a la dicotomía ciencia burguesa-ciencia proletaria, como las que oponen una supuesta ciencia islámica a la occidental, o la ciencia afroamericana o feminista a su contraparte, esta versión de la investigación acción parece ir quedando en el pasado. En relación con las pretensiones de la investigación participativa, de constituir un nuevo paradigma para la metodología de la investigación social, Pablo Latapí escribió en la época de apogeo de la corriente:

> Hay pocas objeciones fuertes que se opongan a la aceptación de la investigación participativa (IP) como una metodología educativa; es evidente que los adultos pobres aumentan su capacidad de autoanálisis y organización reflexionando sistemáticamente sobre su propia práctica. Pero es contra la pretensión de la IP de ser considerada un procedimiento riguroso de investigación contra la que surgen preguntas incómodas. Para algunos de sus defensores, se trata de una alternativa a la investigación prevaleciente, con pretensiones estrictamente académicas. Algunos autores pretenden que se trata de un paradigma emergente que revolucionará las ciencias sociales. Hay inclusive algunos que hablan de una ciencia popular que substituirá a la ciencia social convencional. Estas pretensiones identifican la cuestión central: ¿es válido considerar a la IP como un paradigma alternativo capaz de reorientar la investigación social en el futuro?, ¿o es sólo una moda que debe rechazarse, más retórica que sólida en sus pretensiones? [...] En mi opinión, las insuficiencias de la IP que hemos analizado con respecto al concepto de ciencia, la relación entre acción y conocimiento y la relación entre teoría y práctica tienden a descartar la idea de que la IP es un nuevo

paradigma científico. Esto no significa, sin embargo, que todos sus elementos deban ser rechazados. La investigación social establecida puede aprender algunas lecciones importantes de la IP [...] (1988: 317-318)

La visión actual de la investigación acción, como la presenta Munn-Giddings:

- Es una forma de indagación que pueden emprender profesionales, como los maestros (*insiders*), en contraposición a la *investigación de torre de marfil*, una actividad técnica realizada por expertos (*outsiders*).
- No se orienta a la descripción (cuantitativa o cualitativa) de un fenómeno *tal como es*, sino que pretende modificarlo.
- Suele concebirse como una serie de etapas, espirales o ciclos de planeación, acción, observación y reflexión.
- Utiliza cualquier tipo de técnicas, sobre todo con un enfoque de métodos mixtos que combina datos cuantitativos y cualitativos. (2012: 71-72)

Hoy se acepta en general un rasgo de la investigación acción: que los profesionales de un campo (*v.gr.* maestros) suelen tener un conocimiento de su propio trabajo más amplio que otras personas, aunque no sea explícito, y tenga las limitaciones de cualquier *insider*. Muchas personas reconocen también que los investigadores manejan herramientas metodológicas y técnicas útiles, y que la posición de *outsider* no solo implica límites en cuanto al conocimiento del objeto, sino también ventajas.

Por ello hoy se reconocen cada vez más las ventajas de una colaboración entre profesionales e investigadores, no solo en cuanto a la recolección de información, sino en todas las etapas de un estudio, desde diseño y formulación de preguntas de investigación, hasta el análisis de resultados y las conclusiones. Esto es lo que plantean las concepciones actuales sobre investigación acción y su versión más reciente, la de los estudios conocidos como *practitioner research*.

Ningún capítulo de la obra más reciente de AERA sobre *métodos complementarios en investigación educativa* lleva como título *investigación acción*, pero sí hay uno denominado *Practitioner Inquiry*, con el subtítulo: *Borrando las fronteras entre la investigación y la práctica.* (Cochran-Smith y Donnell, 2006)

Juzgando equivalentes las expresiones *practitioner inquiry y practitioner research,* el texto las entiende como:

> [...] una gama de tipos de investigación educativa en los que el profesional es el investigador, el contexto del ejercicio profesional es el sitio de investigación y el foco del estudio es la práctica profesional misma [...] (Cochran-Smith y Donnell, 2006: 503)

El texto considera que la investigación acción es solo una de las variantes incluidas en la gama de tipos de investigación que comprende la definición anterior:

> Las formas más usuales y mejor conceptualizadas de practitioner inquiry incluyen la investigación acción, la investigación hecha por los docentes, los auto estudios, los trabajos sobre enseñanza y aprendizaje, y los que definen la práctica profesional como objeto de investigación. Aunque estos tipos no agotan la totalidad de la practitioner inquiry, ofrecen una visión de conjunto del campo [...] (Cochran-Smith y Donnell, 2006: 504)

Después de describir los cinco tipos particulares incluidos en la enumeración del párrafo anterior, el texto citado enumera los rasgos que tienen en común, que incluyen entender que:

- El profesional juega simultáneamente el papel de investigador.
- Quienes viven y trabajan en contextos particulares tienen una perspectiva cognitiva significativa sobre esa situación.
- El contexto profesional es a la vez el sitio de investigación.
- Las fronteras entre la investigación y la práctica se desdibujan.
- Las nociones de validez y generalizabilidad se definen legítimamente en formas distintas a los criterios tradicionales de transferibilidad y aplicación.
- En todos los casos se cuida la sistematicidad y la intencionalidad.
- Es importante abrir el trabajo de enseñanza, aprendizaje y, en general, el trabajo de la escuela, a la crítica de la comunidad más amplia. (Cochran-Smith y Donnell, 2006: 507-512)

Design research

Un nuevo tipo de estudios de intervención que ha cobrado fuerza es el que denotan las expresiones *design research* o *design experiments*, entre otras, ya que aún no hay una terminología generalmente aceptada. Con base en lo que sigue, este tipo de trabajos se puede conceptualizar como *investigaciones para el desarrollo de innovaciones mediante procesos iterativos de diseño y prueba en contextos reales*.

Esta tendencia ha recibido la influencia de posturas críticas desarrolladas desde los años 1950: revolución cognitiva/conductismo; enfoques *cuali/cuanti*); papel de la teoría en la investigación (teorías de alcance medio y *grounded theory* vs. grandes teorías); papel de los profesionales vs los investigadores (investigación acción vs. investigación convencional). Los estudios a que se refiere este inciso se distinguen porque combinan, en distintas formas, elementos como los siguientes:

- Orientación aplicada, a partir de la idea de que en educación hay problemas, evidenciados por los bajos resultados de aprendizaje de muchos alumnos, y debidos a la prevalencia de tecnologías (métodos de enseñanza, currículos, materiales, etc.) poco eficaces, que la investigación educativa convencional no ha conseguido modificar significativamente durante muchas décadas.
- Sin detrimento de lo anterior, intentos por hacer avanzar el conocimiento, con aportaciones modestas a la construcción de teorías sobre los aspectos de la realidad que interesan, de los que las teorías existentes saben poco.
- Insatisfacción con métodos de investigación tradicionales, en particular los diseños experimentales, que se cree no tienen en cuenta la complejidad de los fenómenos educativos, en que no procederían esfuerzos por identificar variables a controlar, para aplicar los criterios a fin de atribuir causalidad.
- En consecuencia, preferencia por procesos que combinan ideas concretas de innovación y pruebas de las mismas en contextos reales, para desarrollar ideas más completas que serán a su vez sometidas a prueba, iterativamente.
- Valoración de la opinión de los profesionales del campo, junto con los investigadores, para el planteamiento de las ideas iniciales y su desarrollo

Esto coincide con una obra según la cual, aunque no haya consenso en términos, sí lo hay en cuanto a los rasgos que distinguen estos trabajos:

- Propósito de intervención: la investigación tiene como propósito el diseño de intervenciones en el mundo real.
- Iteración: se incorpora acercamiento cíclico de diseño, evaluación y revisión.
- Orientación a procesos: se evita modelo de caja negra y medición de insumos y productos; se centra atención en comprender y mejorar las intervenciones.
- Orientación utilitaria: el mérito de un diseño se mide, en parte, por su utilidad práctica para usuarios en contextos reales.

- Orientación teórica: el diseño, al menos en parte, se base en proposiciones teóricas, y las pruebas de campo del diseño contribuyen a la construcción de teoría. (Van Den Akker, Graveneijer, McKenney y Nieveen, 2006: 5)

Un pionero recuerda que, desde la década de 1980, varios estudiosos emprendieron trabajos que se conocerían como *design experiments*, sin emplear la expresión, que comenzó a usarse en 1992, con la aparición de artículos de Ann Brown y Allan Collins (Schoenfeld, 2006: 194). A partir del trabajo de los hermanos Wright en los inicios de la aviación, este autor comenta que, cuando se sabe poco de cierto tema:

> La teoría y el diseño crecen en una relación dialéctica, en la que la naciente teoría sugiere un diseño mejorado de alguna manera, y algunos aspectos del diseño sugieren nuevas dimensiones de la teoría, incluso por bricolage [...] a veces hay que crear algo para poder explorar sus propiedades. El acto de crear es un acto de diseño. Y si la creación se hace con un ojo dirigido a la generación y examen sistemático de datos, y al refinamiento de la teoría, el resultado se puede considerar design experiment. (Schoenfeld, 2006: 193)

Brown (1992) ilustra el abismo que hay entre estudiar el aprendizaje en laboratorio o *en la confusión floreciente y vibrante de las aulas de escuelas urbanas empobrecidas*. (*Cfr.* Schoenfeld, 2006: 196)

Entre los partidarios de *design experiments* hay posturas que van del rechazo total de enfoques tradicionales, a la idea de que serían complementarios. Phillips señala que, para una instancia financiadora, es difícil valorar la solidez de propuestas que, por su propio enfoque, no pueden precisar los resultados que podrán ofrecer, que se irán clarificando a medida que se desarrollen, pero sostiene que reducir la investigación científica a trabajos experimentales es inadecuado. Lo justifica con reflexiones sobre la diferencia de las etapas iniciales de la investigación sobre un tema, en las que se precisan las ideas básicas, y las etapas terminales, en las que ya hay definiciones precisas de aspectos que es posible controlar, con estudios que evalúen rigurosamente su impacto. En perspectiva filosófica, se apoya en John Dewey y Karl Popper, e incluso el positivista lógico Reinchenbach y su noción de contexto de descubrimiento vs contexto de justificación. (Phillips, 2006: 94-95)

Dada la variedad de las formas que los trabajos de *design research* pueden adoptar, Kelly opina que no es plausible ni deseable una sola lista de criterios de calidad, y propone dejar abierta la cuestión, regresando al artículo de Brown, en el que:

- Abogaba por un enfoque de métodos mixtos, cualitativos y cuantitativos.
- No se oponía a las mediciones cuantitativas, observando que se puede perfectamente hacer análisis estadísticos en estudios de caso.
- Consideraba que el diseño incluye el aula y el laboratorio bidireccionalmente.
- Llamaba la atención sobre el error de escoger episodios que apoyen las hipótesis favoritas del investigador (*Efecto Bartlett*).
- Proponía aumentar escala de los estudios, pasando de la fase alfa (desarrollo con fuerte control) a la beta (pruebas de campo en sitios selectos con menos control) y a la crítica fase gamma (adopción masiva con control mínimo).
- No rechazaba el propósito de aislar variables y atribuir impacto causal, reconociendo que para pasar a una escala mayor hay que separar variables que originalmente estaban mezcladas [...] (Kelly, 2006: 108-110)

Para un tratamiento amplio del tema véase el manual de Kelly, Lesh y Baek (2008).

Investigación evaluativa

En esta obra se considera que la investigación aplicada comprende los estudios de intervención (intervenciones controladas, investigación acción, *practitioner research* y *design research*), pero no se reduce a ellos. Se propone que hay otros dos tipos de investigación cuyo propósito tampoco se limita a conocer mejor el objeto de estudio, sino que incluye algo más: la investigación evaluativa y la metodológica.

En investigación evaluativa ese propósito adicional es llegar a formular un juicio de valor sobre el objeto de estudio, que puede ser el desempeño de una persona o un grupo (*v.gr.* conocimientos de lenguaje o matemáticas de algunos estudiantes), pero también los resultados de actividades de ciertas instituciones, programas o políticas.

La evaluación está presente en muchos ámbitos de la actividad humana, como economía, salud y educación. Se distinguen muchos tipos de evaluación, según varios criterios: atendiendo a la función que cumpla, hay evaluación diagnóstica, sumativa y formativa; según quien la realiza hay autoevaluación, coevaluación y heteroevaluación; hay evaluación en pequeña o gran escala, de alto o bajo impacto; con referentes normativos o criteriales; de nivel individual o colectivo, etc. Según otro criterio, las evaluaciones se pueden distinguir según su objeto, según el *evaluando*. En educación se pueden distinguir evaluaciones de personas, procesos y otros objetos. De especial interés resultan las evaluaciones de los aprendizajes que alcanzan los alumnos; las del desempeño de maestros y directivos y de las prácticas docentes y de gestión; las de

planteles e instituciones; las del currículo y los materiales educativos; y las de programas y políticas particulares.

Sabiendo que el concepto es complejo, se propone la siguiente definición:

> *Evaluación es la acción y efecto de formular un juicio de valor sobre algún(os) aspecto(s) de cierto evaluando, contrastando el resultado de su medición con un referente (estándar, criterio).*

Se pretende que esta definición sea aplicable a cualquier tipo de evaluación y a las que se refieren a cualquier objeto, destacando además que siempre subyace otra noción igualmente compleja, la de *calidad,* ya que el juicio de valor en que consiste centralmente la evaluación finalmente lo que dice es si el aspecto evaluado es de *buena o mala calidad*.

Esta definición implica que para evaluar algo primero hay que *medirlo* bien. Por ello, para ser de buena calidad, una evaluación debe satisfacer los criterios aplicables a cualquier trabajo riguroso, incluyendo la conciencia de las implicaciones y límites que se derivan del nivel de medición de las variables implicadas.

Pero medir algo no basta para evaluarlo; es necesario además tener referentes precisos de la situación que debería presentar lo medido para ser adecuado. Si no es claro, por ejemplo, qué deben saber de ciertos temas los alumnos que terminan un grado o nivel educativo, será imposible saber si su nivel de aprendizaje es el adecuado, aunque se haya medido con mucha precisión lo que saben.

Para llegar al juicio de valor en que consiste la evaluación es indispensable tener parámetros con los que se pueda contrastar el resultado de la medición, y estos referentes no se pueden derivar de ningún trabajo empírico.

El elemento *medición* es del ámbito empírico, *de lo que es;* pertenece al campo de la investigación. Los referentes pertenecen al ámbito de lo normativo, del *deber ser*, y se ubican en el terreno valoral, ético y jurídico-político. El nivel de conocimientos que tienen ciertos alumnos pertenece al ámbito de lo empírico, y puede medirse con razonable precisión si se utilizan las técnicas adecuadas; pero el nivel que deberían tener los estudiantes del nivel de que se trate es de otro ámbito, y no puede resultar de ninguna medición, sino que lo debe fijar quien tenga facultades legales para ello. En este caso la evaluación se derivará del contraste entre lo que los alumnos saben (medido empíricamente) y lo que deberían saber, según la normatividad aplicable.

En lo que se refiere a medición, se deberán utilizar las mejores técnicas y respetar los cuidados de toda investigación, para asegurar altos niveles de confiabilidad y validez.

En cuanto a referentes, cuando no hay uno preciso definido legalmente se debe acudir al juicio de expertos en medición y profesionales del campo (docentes).

Los juicios de valor a que se llega gracias a una evaluación sustentan la toma de decisiones, de alto o bajo impacto, como detectar puntos débiles de un estudiante, o un grupo para que el docente puede orientar mejor la enseñanza; decidir si un alumno debe pasar al grado siguiente o repetir el que acaba de cursar; identificar a los mejores candidatos para iniciar una carrera y aceptarlos, dejando fuera a otros; dar estímulos o aplicar medidas correctivas a los docentes, según su desempeño; decidir si un programa o una política debe continuar o suspenderse, etc.

Por lo anterior, es claro que las evaluaciones son investigaciones a las que se deben aplicar criterios de calidad similares a los que proceden en una investigación básica, pero cuyo propósito no se limita a incrementar lo que se sabe del objeto de estudio, sino que incluye el emitir un juicio de valor sobre el mismo, con base en el cual se tomarán decisiones. Por ello, una buena evaluación se distingue por el nivel técnico de las mediciones en que se base, para garantizar su confiabilidad y validez; por un diseño sólido; enfoques diversos y complementarios en modelos e instrumentos; selección cuidadosa de muestras representativas, si procede; procesos rigurosos de recolección de datos; y análisis cuidadosos de los datos obtenidos. Pero como evaluar va más allá de medir, una buena evaluación implica también:

- *Pertinencia de los referentes que se definan como parámetros* para contrastar con ellos los resultados de la medición, de manera que las comparaciones tengan sentido. Los referentes se definen normativamente, no se derivan de los datos empíricos. Pueden ser *óptimos*: ideales con que se compara una situación; *promedios* de los individuos que se evalúan; y *mínimos*, con los menores valores aceptables. Cada uno arroja cierta luz sobre lo evaluado y ninguno es suficiente. Convendrá usar los tres tipos de parámetro, para una mejor apreciación. Puede también utilizarse como referente la situación del evaluando mismo en el pasado, con lo que se podrá apreciar si mejora, empeora o se mantiene estable.
- *Mesura de los juicios de valor* derivados de contrastar mediciones y parámetros, evitando excesos triunfalistas o derrotistas y teniendo en cuenta la equidad, considerando el contexto de alumnos, docentes, escuelas u otros evaluandos.
- *Amplitud, oportunidad y transparencia de la difusión de resultados*, para llegar a los sectores involucrados en versiones adecuadas a sus características.

El reconocimiento de la evaluación como investigación aplicada es el resultado de las reflexiones de más de medio siglo. En la década de 1960, Cronbach y Suppes distinguían *investigación pura*, enfocada a establecer con la mayor solidez la verdad de sus conclusiones, e *investigación aplicada*, orientada a sustentar decisiones sobre las acciones a emprender en un contexto práctico. El campo de la evaluación "emergió como una especialidad semiindependiente dentro de los límites de la comunidad de investigación educativa [...]" (Phillips, 2018: 18-19)

Cronbach y colaboradores (1980) hicieron un llamado a emprender una profunda reforma de la evaluación. En referencia expresa a la Reforma promovida por Lutero en 1517, por considerar que la iglesia de Roma se había corrompido, decían:

> Sus sacerdotes y patronos, y quienes desean sus beneficios, han buscado en la evaluación lo que no puede dar, y probablemente no debe [...] La evaluación tiene un trabajo vital por hacer, pero sus instituciones y las concepciones dominantes son inadecuadas. Enamorados de una visión de que las decisiones "correctas" pueden sustituir a los acuerdos políticos, algunos de los que encargan evaluaciones plantean demandas no realistas a los evaluadores [...] Los evaluadores, ansiosos por servir, e incluso por manipular a los que detentan el poder, pierden de vista lo que deberían hacer. Más aún, lo evaluadores se quedan cautivados por las técnicas. Mucho de lo que pretende ser teoría de evaluación es escolástica [...] Estos teólogos de última hora discuten [...] sobre las derivaciones numéricas de modelos artificiales —"¿Cuántos ángeles caben [...]?" —una y otra vez. Es demasiado raro que la discusión baje a tratar cuestiones terrestres como ¿vale la pena la información que se ha recolectado? (Cronbach *et al.*, 1980: 1-2)

Phillips (2018: 29-30) simpatizaba con la idea de que habría diferencia importante entre investigación y evaluación, en particular en el plano operativo: la investigación aplicada, como la evaluación, debe dar resultados en plazos improrrogables y cortos, con recursos limitados, mientras en investigación pura en principio se pueden ampliar los plazos planteando nuevos proyectos de una misma línea cuando los resultados del primero no son concluyentes. Hoy ya no sostiene ese punto de vista, *al menos no tan fuertemente*, por considerar que la diferencia operativa no es absoluta, y que se sitúa más bien en el papel del investigador, distinto en el caso de la investigación orientada a conclusiones y en la enfocada a sustentar decisiones.

Los estudiosos de políticas públicas advierten lo insuficiente de concebir de manera simple la relación evaluador-clientes, como si las decisiones de políticas fueran cuestión

meramente técnica sin consideraciones ideológicas, políticas y éticas, y como si los *técnicos* solo tuvieran que analizar la eficiencia de unos medios y otros, prescindiendo de la discusión sobre fines perseguidos y valores implicados. Estos especialistas plantean que el papel de un analista de políticas es parecido al de un artesano, y prestan atención al papel de la argumentación y la persuasión en los procesos de deliberación pública que implica definir políticas (*Cfr*. Maione, 1997).

Por otra parte, las evaluaciones se distinguen por rasgos particulares, que tienen implicaciones por lo que se refiere a su planeación y desarrollo. Las más conocidas son las de los aprendizajes que alcanzan los alumnos. Los docentes evalúan regularmente a los alumnos del grupo a su cargo, lo que pueden hacer rutinaria o cuidadosamente, pero los resultados de los alumnos de un maestro no son estrictamente comparables con los de otros maestros, incluso del mismo plantel.

Para tener resultados comparables hay que utilizar instrumentos estandarizados. Las pruebas de este tipo surgieron desde inicios del s. xx en los Estados Unidos, y se han extendido a muchos sistemas educativos, sin que autoridades y público en general tengan siempre ideas claras de sus alcances y límites, como el que las formas usuales solo miden competencias de nivel de demanda cognitiva bajo, dejando fuera áreas completas del currículo y lo no cognitivo; que los resultados tienen un margen de error y son inestables; que no es fácil saber con precisión en qué medida influyen en los resultados los factores de la escuela y los del hogar, etc.

De la extensa literatura sobre el tema se pueden destacar la versión más reciente de los *Estándares para pruebas educativas y psicológicas* (AERA, APA, NCME, 2014), y la obra de referencia *Educational Measurement* (Brennan, 2006).

Las evaluaciones del desempeño de los maestros y las prácticas docentes son otra variante de gran complejidad, desarrollada a partir de 1986, con el trabajo de Lee Shulman. La forma tradicional de evaluar a los maestros, con base en las observaciones del trabajo en el aula hechas por un supervisor una o dos veces al año, no eran confiables, y la insatisfacción al respecto llevó al desarrollo de otras formas:

- Unas basadas en pruebas de habilidades básicas, de las materias a enseñar y de conocimiento pedagógico, o sistemas observación, centrados en puntos superficiales, como comenzar a tiempo, manejar rutinas, redactar objetivos.
- Otras retomando la idea del pago por mérito con Modelos de Valor Agregado (*Value-Added Models*), que permitirían controlar diferencias en la situación inicial de los alumnos, evitando comparar solo resultados a fin del curso, para

poderlos atribuir al trabajo del maestro, pero que en la práctica son también poco confiables e inestables, además de difíciles de implementar.
- Y sistemas de indicadores múltiples, con pruebas, portafolios de evidencias y observaciones del trabajo en aula videograbadas. (Martínez Rizo, 2016).

De la literatura sobre el tema puede mencionarse Darling-Hammond, 2013; Gitomer, 2009; Ingvarson y Hattie, 2008; Kane, Kerr y Pianta, 2014; Peterson, 2000.

En evaluación de programas y políticas las hay de diseño, de implementación, para revisar, por ejemplo, si se cumplieron los lineamientos establecidos, si los recursos se aplicaron a los propósitos, y de impacto, que valoran el grado en que se alcanzan los efectos planeados. En este caso hay que determinar si los efectos observados se debieron realmente al programa o política, y no a factores ajenos, es necesario el uso de experimentos estrictos, o al menos de diseños cuasiexperimentales, para sustentar las atribuciones de causalidad. De la literatura de este campo se pueden mencionar Fernández-Ballesteros, 2001; McDavid y Hawthorn, 2006; Sykes, Schneider y Plank, 2009; Wholey, Hatry y Newcomer, 2004.

Sobre los temas anteriores, así como sobre otros tipos particulares de evaluación, como de planteles e instituciones escolares, del currículo y otras se puede ver el *Manual Internacional* de Kellaghan y Stufflebeam (2003).

Investigación metodológica

Este tipo de investigación tampoco pretende mejorar el conocimiento de un objeto de estudio, sino solo desarrollar herramientas que permitan perseguir ese propósito sustantivo en trabajos posteriores. Aunque el desarrollo de instrumentos de obtención y análisis de información ha sido siempre objeto de atención por los investigadores, en general no lo ha sido suficiente para reconocer que estos trabajos tienen especificidad e importancia suficientes para constituir un tipo investigación como aquí se hace. En los estudios de tipo encuesta, por ejemplo, era usual considerar como un paso particular en el desarrollo de un proyecto la elaboración de un cuestionario u otro instrumento similar, o bien la elección de uno que hubiera sido elaborado antes por otros investigadores, en el marco de otro proyecto.

La importancia práctica y la complejidad de algunas investigaciones educativas ha llevado a que la etapa de desarrollo de instrumentos exija cada vez más tiempo y cuidados más exigentes, lo que ha hecho que alcancen un reconocimiento como investigaciones de pleno derecho por sí mismas. Por otra parte, los diferentes tipos de investigación pueden requerir la aplicación de varios instrumentos, y el desarrollo

de cada uno implica pasos y cuidados particulares. Por ello hay también variantes de la investigación metodológica o instrumental:

- La metodología para elaborar pruebas estandarizadas se ha desarrollado a lo largo de más de un siglo con la Teoría Clásica de las Pruebas, las pruebas con referencia a un dominio o criterio, modelos de respuesta al ítem, teoría de generalizabilidad, técnicas de equiparación, diseños matriciales, pruebas adaptativas por computadora, preguntas de respuesta construida, etc.
- Las técnicas de entrevista han tenido una evolución importante, incluyendo las entrevistas cognitivas y de pensamiento en voz alta. Los diseños basados en observación visual han visto el desarrollo de protocolos de observación desde mediados del siglo XX hasta la fecha, y algo similar ha ocurrido con los protocolos para la obtención y análisis de evidencias o artefactos.
- Otros instrumentos de evaluación son los indicadores, a veces tratados de manera superficial, como resultado de un trabajo meramente administrativo, cuando los especialistas consideran que el tiempo que se requiere para el desarrollo y la maduración de un solo indicador puede llevar décadas.
- Después de que un instrumento ha sido desarrollado hay que validarlo y, en ocasiones, adaptarlo, lo cual implica también investigaciones metodológicas que pueden requerir el trabajo de equipos completos durante varios años.

Cuidar la calidad de la información que se obtiene con un instrumento implica, pues, investigaciones que consideren el proceso de medición en todas sus etapas. En este sentido es útil el enfoque *paso-por paso* de Crooks, Kane y Cohen:

> Una evaluación educativa es una cadena de etapas eslabonadas entre sí [...] El modelo de la cadena sugiere que la validez del conjunto se ve limitada por el eslabón más débil, y que los esfuerzos por hacer particularmente fuertes sólo algunos eslabones pueden ser estériles e incluso dañinos. (1996)

Conclusión

En la Introducción de la obra se comentó la tendencia a reducir la investigación educativa a estudios experimentales y cuasiexperimentales. Se comentó también la reacción de la comunidad de investigadores, en la forma de la definición más amplia de la AERA, y en particular en unas palabras de Shavelson y Towne:

> [...] El diseño de un estudio no lo hace científico por sí mismo. Hay una amplia gama de diseños legítimos que se pueden usar en la investigación educativa que van de un experimento con asignación aleatoria para estudiar un programa de bonos educativos, a un estudio de caso etnográfico en profundidad de unos maestros, o a un estudio neurocognitivo de cómo se aprenden los números, utilizando tomografía por emisión de positrones para formar imágenes del cerebro [...] (Shavelson y Towne, 2002: 6)

Esta es la idea que orienta este capítulo, coincidiendo con Vogt, Gardner y Haeffele, de quienes se han tomado muchas ideas. Estos autores advierten que los señalamientos sobre los límites de los estudios experimentales no se deben entender como una crítica de ese tipo de diseño, pero sí de una forma de pensar sobre los métodos de investigación que parta de la idea de que hay un método superior, cuyas bondades deberían tratar de imitar los demás (*gold-standard thinking*), y añaden:

> No sería un buen consejo para investigadores ni para tomadores de decisión elevar los modelos multinivel, las entrevistas semi-estructuradas, observación participante o cualquier otro buen método, al nivel de estándar único contra el que todos los demás deberían compararse [...] Partimos del supuesto de que todas las preguntas de investigación pueden abordarse de múltiples maneras, cada una de las cuales tiene ventajas y límites [...] la elección que uno haga de un diseño debe ser dirigida por la pregunta de investigación, por el contexto en que uno trata de responderla, y por los propósitos del estudio [...] (2012: 49)

Referencias

Introducción

Arthur, J., Waring, M., Coe, R. y Hedges, L. V. (Eds.). (2012). *Research Methods & Methodologies in Education*. Thousand Oaks: Sage.

Green, J. L., G. Camilli y P. B. Elmore (Eds.). (2006). *Handbook of Complementary Methods in Education Research*. New York: Routledge.

Jaeger, R. M. (1997). *Complementary methods for research in education, 2nd ed.* Washington: American Educational Research Association. (1rst ed. 1988)

Moss, P. A. y Haertel, E. H. (2016). *Engaging Methodological Pluralism*. En Gitomer, Drew H. y Bell, Courtney A. *Handbook of Research on Teaching, 5th Ed.* [Cap. 3, pp. 127-247). Washington: American Educ. Research Assoc.

Spector, P. E. (1981). *Research Designs.* Sage, QASS N° 23.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design***.** New York: The Guilford Press.

Encuestas

Bradburn, N. M. y S. Sudman (1988). *Polls and Surveys. Understanding What They Tell Us.* San Francisco: Jossey Bass.

Fowler, Floyd J. Jr. (1995). *Improving Survey Questions. Design and Evaluation*. Thousand Oaks: Sage.

Groves, R. M., Fowler, Floyd J., Couper, M. P., Lepkowski, J. M., Singer, E. y Tourangeau, R. (2009). *Survey Methodology, 2nd Ed.* Hoboken, NJ: John Wiley & Sons, Inc.

Harkness, J. A., Fons J. R. van de Vijver y P. Ph. Mohler. (2002). *Cross-Cultural Survey Methods.* Wiley Series in Survey Methodology. Wiley-Interscience.

Rea, L. M. y R. A. Parker. (1992). *Designing and conducting survey research. A comprehensive guide.* San Francisco: Jossey Bass.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design,* [Surveys, Cap. 1, 7 y 13]. New York: The Guilford Press.

Estudios de casos

Ashley, L. D. (2012). Case study research. En A., James, Waring, M., Coe, R. y Hedges, L. V. (Eds.). *Research Methods and Methodologies in Education* [13, pp. 102-107]. Thousand Oaks: Sage.

Delgado, J. M. y Gutiérrez, J. (Eds.). (1999). *Métodos y técnicas cualitativas de investigación en ciencias sociales.* Madrid: Ed. Síntesis.

Hine, Ch. (2011). Virtual Ethnography: Modes, Varieties, Affordance. En Fielding, N., Lee, R. M. y Blank, G. (Eds.). *The SAGE Handbook of Online Research Methods* (257-270)**.** Los Angeles: Sage (1st. ed. 2008).

LeCompte, M. D., Millroy, W. L., y Preissle, J. (Eds.). (1992). *The Handbook of Qualitative Research in Education.* San Diego: Academic Press.

Lincoln, Y. S. y Guba, E. G. (1985). *Naturalistic Inquiry.* Beverly Hills: Sage.

Miles, M. B. y Huberman, A. M. (1984). *Qualitative Data Analysis. A Sourcebook of New Methods.* Beverly Hills: Sage.

Merriam, S. B. (2009). *Qualitative Research. A Guide to Design and Implementation. (Revised and Expanded from Qualitative Research and Case Study Applications in Education***.** San Francisco: Jossey Bass.

Stenhouse, L. (1985). Case Study Methods. En Husén, T. y Postlethwaite, T. S. N. (Eds.). *The International Encyclopedia of Education* [pp. 645-650]. Oxfor-New York: Pergamon.

Sturman, A. (1997). Case Study Methods. En Keeves, J. P. (Ed.). *Educational Research, Methodology and Measurement. An International Handbook* [pp. 61-67]. Oxford-New York: Pergamon.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design,* [Interviews, Cap. 2, 8 y 14]. New York: The Guilford Press.

Weiss, E. (2017). Hermenéutica y descripción densa versus teoría fundamentada. *Revista Mexicana de Investigación Educativa,* Vol. 22 (73): 637-654.

Yin, R. K. (1984). *Case Study Research: Design and Methods.* Thousand Oaks: Sage.

Yin, R. K. (1993). *Applications of Case Study Research.* Thousand Oaks: Sage.

Observaciones

Everston, C.M. y Green, J. L. (1986). Observation as Inquiry and Method. En Wittrock, M. C. (Ed.). *Handbook of Research on Teaching. Third Ed.* (pp. 162-213). New York: Macmillan Publ. Co.

Floden, R. E. (2001). Research on Effects of Teaching: A Continuing Model for Reseach on Teaching. En Richardson, V. (Ed.). *Handbook of Research on Teaching, Fourth Edition.* (Págs. 3-16). Washington: AERA.

Goe, L., C. Bell y O. Little (2008). *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. Washington: NCCTQ.

Good, Th. L. y Brophy, J. E. (2010). *Looking in Classrooms,* 10thed.Boston: Pearson. (1st Ed. 2000).

Medley, D. M. y Mitzel, H. E. (1963). Measuring Classroom Behavior by Systematic Observation**.** En Gage, N. L. (Ed.). *Handbook of research on teaching.* (pp. 247-328). Chicago: Rand McNally.

Rosenshine, B. y N. Furst. (1973). The use of direct observation to study teaching. En Travers, Robert M. W. (Ed.). *Second Handbook of Research on Teaching* (pp. 122-183). Chicago: Rand McNally College Publ. Co.

Stallings, J. A. (1977). *Learning to Look. A Handbook on Classroom Observation and Teaching Models.* Belmont, CA: Wadsworth Publiching Co., Inc.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design,* [Observations, Cap. 4, 10 y 16]. New York: The Guilford Press.

Experimentos y cuasiexperimentos

Campbell, D. T., y Stanley, J. C. (1963). Experimental and Quasi-experimental Designs for Research on Teaching. En Gage, N. L. (Ed.). *Handbook of Research on Teaching.* [5, pp. 171-246]. Chicago: Rand McNally & Co. (En español, Buenos Aires: Amorrortu, 1973).

Murnane, R. J. y Willett, J. B. (2011). *Methods Matter. Improving Causal Inference in Educational and Social Research.* Oxford: Oxford Univ. Press.

Schneider, B. ***et al.*** (2007). *Estimating Causal Effects Using Experimental. And Observational Designs. A Think Tank White Paper.* AERA.

Shadish, W. R., Cook, T. D., y Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston-New York: Houghton Mifflin Co.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design,* [Experiments, Cap. 3, 9 y 15]. New York: The Guilford Press.

Diseños longitudinales

Keeves, J. P. (1997). Longitudinal Research Methods. En Keeves, J. P. (Ed.). *Educational Research, Methodology and Measurement. An International Handbook* [pp. 138-149]. Oxford-New York: Pergamon.

Lietz, P., y Keeves, J. P. (1997). Cross-sectional Research Methods. En Keeves, J. P. (Ed.). *Educational Research, Methodology and Measurement. An International Handbook* [pp. 119-126]. Oxford-New York: Pergamon.

Ryder, N. B. (1965). The Cohort as a Concept in the Study of Social Change. *American Sociological Review,* Vol. 30, N° 6, pp. 843-861.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin,* 64: 92-107.

Willet, J. B., y Singer, J. D. (1997). Event History Analysis. En Keeves, J. P. (Ed.). *Educational Research, Methodology and Measurement. An International Handbook* [pp. 513-519]. Oxford-New York: Pergamon.

Estudios múltiples

Bericat, E. (1998). *La integración de los métodos cuantitativo y cualitativo en la investigación social. Significado y medida*. Barcelona: Ariel.

Gorard, S. y Taylor, Ch. (2004). *Combining Methods in Educational and Social Research*. Berkshire: Open University Press.

Tashakkori, A. y Teddlie, Ch. (1998). *Mixed Methodology. Combining Qualitative and Quantitative Approaches*. Thousand Oaks: Sage.

Nisbett, R. y Cohen, D. (1996). *Culture of Honor. The Psychology of Violence in the South*. Boulder: Westview Press.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design,* [Combined, Cap. 6, 12 y 18]. New York: The Guilford Press.

Investigación de archivo y análisis secundario de datos

Bruns, A. y J. Burgess. (2012). Doing blog research. En A., James, M. Waring, R. Coe y L. V. Hedges (Eds.). (2012). *Research Methods & Methodologies in Education*. Thousand Oaks: Sage. Cap. 28, pp. 202-209.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design,* [Archival Designs, Cap. 5, 11 y 17]. New York: The Guilford Press.

Análisis de evidencias

Martínez, J. F., Borko, H. y Stecher, B. M. (2012). Measuring Instructional Practice in Science using Classroom Artifacts: Lessons Learned from Two Validation Studies. *Journal of Research in Science Teaching,* 49 (1): 38-67.

Porter, A. C., Youngs, P. y Odden, A. (2001). Advances in teacher assessments and their uses. En Richardson, Vi. (Ed.). *Handbook of Research on Teaching*. Washington: AERA, pp. 259-297.

Webb, E. J., Campbell, D. T., Schwartz R. D., y Sechrest, L. (1966). *Unobstrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally College Publishing Co.

Síntesis de investigaciones

Cooper, H. M. (1984). *The Integrative Research Review: A Systematic Approach.* Beverly Hills: Sage.

Cooper, H. y Hedges, L. V. (Eds.). (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Light, R. J. y Smith, P. V. (1971). Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies. *Harvard Educational Review.* Vol. 41 (4): 429-471. Reprint N° 16, pp. 67-104.

Investigación aplicada

En general

Bickman, L. y Rog, D. J. (Eds.). (1998). *Handbook of Applied Social Research Methods*. Thousand Oaks: Sage.

Bickman, L., Rog, D. J. y Hedrick, T. (1998). Applied Research Design: A Practical Approach. En Bickman, L. y Rog, D. J. (Eds.). *Handbook of Applied Social Research Methods* [1, pp. 5-37]. Thousand Oaks: Sage.

Intervenciones controladas

Boruch, R. F. (1998). Randomized Controlled Experiments for Evaluation and Planning. En Bickman, L. y Rog, D. J. (Eds.). *Handbook of Applied Social Research Methods* [6, pp. 161-191]. Thousand Oaks: Sage.

Reichardt, Ch. S. y Mark, M. M. (2012). Quasi-experimentation. En Arthur, J., Waring, M., Coe, R. y Hedges, L. V. (Eds.). *Research Methods & Methodologies in Education* [7, pp. 193-228]. Thousand Oaks: Sage.

Tymms, P. (2012). Interventions: experiments. En Arthur, J., Waring, M., Coe, R. y Hedges, L. V. (Eds.) *Research Methods & Methodologies in Education* [19, pp. 137-140]. Thousand Oaks: Sage.

Yin, R. K. (1998). The Abridged Version of Case Study Research. En Bickman, L. y Rog, D. J. (Eds.). *Handbook of Applied Social Research Methods* [8, pp. 229-260. Thousand Oaks: Sage.

Investigación acción y practitioner research

Cochran-Smith, M. y Donnell, K. (2006). Practitioner Inquiry: Blurring the Boundaries of Research and Practice. En Green, Judith L., G. Camilli y P. B. Elmore (Eds.). *Handbook of Complementary Methods in Education Research*. [30, pp. 503-518]. New York: Routledge.

Corey, S. M. (1953). *Action Research to Improve School Practices*. New York: Teacher College.

Kemmis, S. (1997). Action Research. En Keeves, J. P. (Ed.). *Educational Research, Methodology and Measurement. An International Handbook* [pp. 173-179]. Oxford-New York: Pergamon.

Hall, B. L. (1997). Participatory Research. En Keeves, J. P. (Ed.) *Educational Research, Methodology and Measurement. An International Handbook* [pp. 198-205]. Oxford-New York: Pergamon.

International Council for Adult Education (1977). *Status Report on the Participation Research Projects*. Toronto: ICAE.

Latapí, P. (1988). Participatory Research: A New Research Paradigm? *The Alberta Journal of Educational Research*. Vol. XXXIV (3): 310-319.

Munn-Giddings, C. (2012). Action research. En Arthur, J., Waring, M., Coe, R. y Hedges, L. V. (Eds.) *Research Methods & Methodologies in Education* [8, pp. 71-75]. Thousand Oaks: Sage.

Design research

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal of the Learning Sciences,* 2 (2), 141-178.

Kelly, A. E. (2006). Quality criteria for design research: evidence and commitments. En Van Den Akker, J., Graveneijer, K., McKenney, S. y Nieveen, N. (Eds.). *Educational Design Research* [pp. 107-118]**.** New York-London: Routledge.

Kelly, A. E., Lesh, R. A. y Baek, J. Y. (Eds.). (2008). *Handbook of Design Research Methods in Education. Innovations in STEM Learning and Teaching.* New York-London: Routledge.

Phillips, D. C. (2006). Assessing the quality of design research proposals: some philosophical perspectives. En Van Den Akker, J., Graveneijer, K., McKenney, S. y Nieveen, N. (Eds.). *Educational Design Research* [6, pp. 93-99]**.** New York-London: Routledge.

Schoenfeld, A. H. (2006). Design Experiments. En Green, J. L., G. Camilli y P. B. Elmore. (Eds.). *Handbook of Complementary Methods in Education Research.* [11, pp. 193-205]. New York: Routledge.

Van Den Akker, J., Graveneijer, K., McKenney, S. y Nieveen, N. (2006). Introducing educational design research. En Van Den Akker, J., Graveneijer, K., McKenney, S. y Nieveen, N. (Eds.). *Educational Design Research* [pp. 3-7]**.** New York-London: Routledge.

Investigación evaluativa

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, Authors.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Westport, American Council on Education & Praeger.

Cronbach, Lee J. ***et al.*** (1980). *Toward Reform of Program Evaluation. Aims, Methods, and Institutional Arrangements.* San Francisco: Jossey Bass Publ.

Darling-Hammond, L. (2013). *Getting Teacher Evaluation Right. What Really Matters for Effectiveness and Improvement.* New York: Teachers College**.**

Fernández-Ballesteros, R. (2001). (Ed.). *Evaluación de programas. Una guía práctica en ámbitos sociales, educativos y de salud.* Madrid: Síntesis.

Gitomer, D. H. (2009). (Ed.). *Measurement Issues and Assessment of Teaching Quality.* Thousand Oakes: Sage.

Ingvarson, L. y J. Hattie (Eds.). (2008). *Assessing Teachers for Professional Certification: The First Decade of the National Board for Professional Teacher Standards.* Amsterdam: Elsevier-JAI.

Instituto Nacional para la Evaluación de la Educación. (2006). *Plan maestro de desarrollo 2007-2014*. México: Autor.

Kane, T. J., Kerr, K. A., y Pianta, R. C. (2014). *Designing Teacher Evaluation Systems.* San Francisco: Jossey-Bass.

Kellaghan, T. y Stufflebeam, D. L. (Eds.). (2003). *International Handbook ef Educational Evaluation* [2 vols.]. Dordrecht: Kluwer Academic Publishers.

Maione, G. (1997). *Evidencia, argumentación y persuasión en la formulación de políticas.* México: Fondo de Cultura Económica y CNCPAP.

Martínez Rizo, F. (2016). *La evaluación de docentes de educación básica. Una revisión de la experiencia internacional.* México: INEE.

McDavid, J. C. y Hawthorn L. R. (2006). *Program Evaluation & Performance Measurement. An Introduction to Practice.* Thousand Oaks: Sage.

Peterson, K. D. (2000). *Teacher Evaluation. A Comprehensive Guide to New Directions and Practices, 2nd Ed.* Thousand Oaks: Corwin Press.

Phillips, D. C. (2018). Las muchas funciones de la evaluación en educación. En de Ibarrola Nicolín, M. (Coord.). *Temsa clave de la evaluación de la educación básica* [17-31]. México: FCE-INEE.

Sykes, G., Schneider, B. y Plank, D. N. (Eds.). (2009). *Handbook of Education Policy Research.* New York-London: Routledge-AERA.

Wholey, J. S., Hatry, H. P. y Newcomer, K. E. (Eds.). (2004). *Handbook of Practical Program Evaluation.* Thousand Oakes: Sage.

Investigación metodológica

Crooks, T. J., M. T. Kane y A. S. Cohen (1996). Threats to the Valid Use of Assessments. *Assessment in Education,* Vol. 3 (3): 265-285.

Conclusión

Shavelson, R. J. y L. Towne (Eds.). (2002). *Scientific Research in Education*. Washington. National Research Council: National Academy Press.

Vogt, W. P., Gardner, D. C. y Haeffele, L. M. (2012). *When to Use What Research Design.* New York: The Guilford Press.

CAPÍTULO 3
OBTENCIÓN DE INFORMACIÓN EMPÍRICA

CONTENIDO

Introducción

Una vez precisado un objeto de estudio y formuladas unas preguntas investigables hay que planear el trabajo, seleccionando un diseño entre varios posibles, los que se han presentado en el Capítulo 2. En seguida hay que recoger la información empírica que servirá para darles respuesta. En este capítulo, después de revisar las nociones de observación y medición, para ver su relación y lo que implica medir, se presentan tres grupos de acercamientos que se pueden emplear para obtener información en estudios empíricos en ciencias sociales y de la conducta: los que se basan en información ofrecida por los sujetos dando respuesta a preguntas que se les dirigen o reaccionando a estímulos orales o escritos que se les presentan; los basados en observación visual, en vivo o con videograbaciones; y los que se basan en el análisis de productos de la actividad de los sujetos.

Observación y medición

Tanto el término observación como el de medición implican captar, percibir o registrar de alguna forma una realidad. En el sentido usual, *observación* alude a captar información mediante la vista. En un sentido más amplio se puede usar para referirse a la obtención de información por medio del oído o cualquier sentido, pero en general entendiendo que se trata de formas poco estructuradas o sistemáticas de captar información. En el sentido usual *medición*, por su parte, se refiere a formas altamente

estructuradas de captar información, en concreto de carácter numérico o, si se prefiere, cuantitativo. Ahora bien, la observación, en el sentido de información captada por cualquier sentido, puede hacerse en forma más o menos estructurada, y por ende la medición puede verse como un tipo particular de observación muy estructurada. Y como técnicamente hay varios niveles de medición, la diferencia entre observación y medición es menos clara.

La actividad de medir es antiquísima respecto a ciertos aspectos de la realidad, como la medición de longitudes o distancias, con pulgadas, pies, codos o brazas, o con piedras millares en las calzadas romanas. Por lo que se refiere a variables relevantes en ciencias sociales, los primeros esfuerzos por medirlas se remontan a las pruebas de inteligencia elaboradas por Binet y Simon o Thurstone, a principios del siglo XX, aunque ya en 1860 se había desarrollado la "psicofísica" para tratar de medir las reacciones fisiológicas a estímulos sensoriales. (Stevens, 1976)

Sin embargo, la medición era una actividad sin fundamentos teóricos rigurosos. No fue hasta 1917 cuando Campbell elaboró la teoría de la medición física (Keats, 1988). Los test de inteligencia, por su parte, se basaron durante medio siglo o más en la teorización de Spearman, no relacionada con la de Campbell. Una comisión que analizó el tema (Ferguson *et al.*, 1940) concluyó que los principios de la medición física no eran plenamente aplicables en ciencias del hombre (Keats, 1988). Los trabajos de Gulliksen, y sobre todo Stanley Stevens, llevaron a una conclusión más favorable, con la distinción fundamental de los niveles de medición nominal, ordinal, de intervalo y de razón (1946).

Campbell definía medición como "asignación de números para representar propiedades", y Stevens precisó que medir, hablando en general, es "la asignación de numerales a objetos o eventos según ciertas reglas". (Cicourel, 1964: 10)

Pero la forma en que se haga esa asignación puede variar mucho. Veamos tres situaciones muy diferentes:

Primera: Tenemos un grupo de niños y niñas, a los que separamos por sexo y asignamos convencionalmente el numeral 1 a varones y el 2 a mujeres.

Segunda: Pedimos al grupo que forme una fila por orden de estatura, sin tener en cuenta el sexo, y asignamos el símbolo 1º al más alto, 2º al que le sigue, etc.

Tercera: Mediante una balanza pesamos a los miembros del grupo y a cada uno asignamos el número que corresponde a su peso en kilogramos o en libras.

El primer caso es el de una medición a nivel nominal de la propiedad sexo; el segundo, una medición ordinal de la estatura; y el tercero de la medición *cardinal*

del peso. Pero lo que hemos hecho en cada caso es diferente, de suerte que usar el verbo *medir* o el sustantivo medición en los tres casos es engañoso. En el lenguaje ordinario no diríamos *medir* más que en el tercer caso.

Analicemos la forma de *asignar numerales*.

Medición a nivel nominal (clasificación)

En el nivel de medición llamado nominal, la asignación de numerales es convencional o arbitraria. Así como se asignó 1 al *valor masculino* de la característica (o variable) sexo, y 2 al *femenino*, se pudo hacer al revés, 1 a F y 2 a M. Entre los *valores* masculino y femenino no hay relación de orden, mucho menos posibilidad de precisar intervalos entre valores. Se trata, como indica el término mismo, solamente de *nombres* distintos.

Si consideramos otros ejemplos de variables de este nivel, como el lugar de nacimiento o el nombre propio de las personas de un grupo, podemos clasificarlas en nacidas en Aguascalientes, Zacatecas o Jalisco, o en subgrupos de los que se llamen José, Juan o Francisco y asignar convencionalmente números a dichos *valores* (en este caso *nombres*, medición *nominal*). Se puede asignar el 1 a los nacidos en Aguascalientes, 2 en Jalisco, o bien 1 los que se llaman José, 2 los Juanes y así sucesivamente. Medir a nivel nominal, en el sentido usual de las palabras no es medir sino *clasificar*.

Medición a nivel ordinal (ordenación)

¿Qué se hace para asignar números cuando se *mide* a este nivel? En este caso la operación fundamental es *comparar entre* sí los objetos que se van a medir (según cierto aspecto), unos con otros, ordenarlos y asignar números ordinales progresivamente de mayor a menor o viceversa, sin que importe la diferencia que haya entre unos y otros.

En principio para ordenar así un conjunto de objetos habría que compararlos todos entre sí, de dos en dos. En realidad, dada la propiedad de transitividad (si A> B y B> C se sigue que A> C) no es indispensable hacer todos los pares de comparaciones posibles, sino sólo un número menor. Por otra parte, no siempre será tan fácil hacer la comparación de los objetos de dos en dos como en el caso de la estatura, en el que puede procederse *a simple vista*.

Esta situación es excepcional. En general habrá que ingeniárselas para idear una forma de hacer esa comparación. Por ejemplo, si tratáramos de ordenar a los niños y niñas de un grupo según su peso tendríamos que pensar en un artefacto, aunque sea tan sencillo como una balanza simple bien equilibrada y sensible para comparar a los niños de dos en dos y poderlos ordenar del más pesado al más liviano o viceversa.

En sentido estricto, en el caso de variables de nivel ordinal tampoco hablaríamos de medir sino de *ordenar*.

Medición a nivel cardinal (medición propiamente dicha)

En este caso, el criterio de asignación de números implica la comparación de lo que se quiere medir con una unidad estándar, que pueda tener múltiplos y submúltiplos, para ver cuántas veces la contiene la realidad que se desea medir.

Este es el sentido usual de la palabra, pero ¿de dónde salen las unidades estándar de comparación? Hoy podemos medir la estatura de una persona, o cualquier distancia, en metros, centímetros o kilómetros, o en pies, yardas y millas, pero hay que pensar cómo se hacía antes de que existiera el sistema métrico decimal o se estandarizara el sistema inglés. La Revolución Francesa implantó el primero ante el caos que representaba que *no solo cada provincia*, *sino cada distrito*, y *casi cada pueblo*, tuviera sus propias medidas, lo que provocaba una inmensa perplejidad, en palabras de un agrónomo inglés que visitó Francia poco antes (*Cfr*. Hand, 2016: 5).

Las medidas de longitud antiguas se referían a partes del cuerpo humano: pie, brazo, codo, cuarta, geme, dedo, pulgada, y otras. A nadie escapa que, dada la variedad de tamaños de manos o pies de diferentes personas, se necesita unificar, *estandarizar* una medida así, para que sea uniforme. Así, se dice que para los griegos la unidad no era cualquier *pie*, sino precisamente el pie de la estatua de Apolo que se veneraba en el santuario de Delfos. Esto es tan arbitrario —y tan válido— como la definición de metro como diezmillonésima parte del cuadrante del meridiano terrestre, como la longitud de una barra de platino iridiado que se guarda en la Oficina Internacional de Pesas y Medidas de Sevres, o como las definiciones basadas en la longitud de onda de cierta sustancia o en la velocidad de la luz (Robinson, 2007: 30). Un curioso texto del siglo XVI nos muestra otra manera de *estandarizar* una de estas medidas naturales:

> Un domingo, a la puerta de una iglesia, pida a 16 hombres, de estatura alta y baja, como vayan pasando, que se detengan al salir al fin del servicio religioso; hágalos que pongan su pie izquierdo uno tras el otro, y la longitud total que se obtenga será una vara correcta y legítima para medir e inspeccionar la tierra, y la dieciseisava parte de ella será un pie correcto y legítimo. (*Cfr*. Hand, 2016: 3-4)

Habiendo definido una unidad estándar, con múltiplos y submúltiplos, se puede no sólo ordenas objetos de mayor a menor, sino también precisar la diferencia, el

intervalo que separa a uno de otro, precisar qué tanto mayor o menor es un objeto en comparación con otro, ya que ambos son comparados con la unidad estándar.

Si consideramos el peso, con una balanza simple podemos comparar el peso de unos objetos con otros, poniéndolos de dos en dos en la balanza y estableciendo un orden del más pesado al más ligero de ellos. Pero podemos hacer algo más: escoger un objeto de cierto peso como unidad de medida, digamos la piedra A; identificar otras que pesen lo mismo, o bien múltiplos (*v.gr.* la piedra B equivale a 2A, la C a 5A) y submúltiplos (*v.gr.* las dos piedras D, que pesan lo mismo, equivalen juntas al peso de la piedra A, por lo que su peso es la mitad de A). Una vez hecho lo anterior podemos comparar cada objeto con la unidad estándar, y no sólo establecer un orden, sino decir: el objeto 1 pesa 10 piedras A; el objeto 2 es ligeramente más pesado, pues pesa 10½ piedras A; en tanto que el objeto 3 es mucho más ligero puesto que pesa sólo la mitad que el objeto 1, 5 piedras A. En este ejemplo y en el anterior hemos logrado *cuantificar* o, si se prefiere, *medir en sentido estricto* la variable en cuestión.

- *Medir a nivel nominal* es solamente *clasificar* ciertos objetos y asignarles números de manera arbitraria, como cuando se asigna a los hombres el número 1 y a las mujeres el 0 (o viceversa) para procesar los datos recabados en una encuesta.
- *Medir a nivel ordinal* implica *comparar entre sí* unos objetos para establecer un orden entre ellos en lo relativo a alguna propiedad (*v. gr.* estatura o peso de unas personas), y finalmente asignar números, que se llaman precisamente *ordinales*.
- *Medir a nivel cardinal (de intervalo o de razón)*, cuantificar o medir a secas a ciertos atributos, quiere decir, primero, definir una unidad estándar; luego comparar con esa unidad los objetos para ver cuántas unidades o partes de unidad tiene cada uno, lo que permite ordenar, y además precisar el intervalo o distancia que separa un objeto de otro; y por fin asignar números cardinales según el resultado.

En sentido ordinario sólo en el tercero de estos casos se puede hablar de medición; en los otros dos se trata de clasificación y de ordenamiento. El autor que propuso distinguir estos niveles de medición, Stanley S. Stevens, en el texto en que lo hizo, advertía ya:

> La escala nominal es una forma primitiva, y en forma muy natural hay muchas personas que dirán que es absurdo atribuir la dignidad que implica el término medición a este proceso de asignar numerales [...] (1946: 679)

La medición en ciencias sociales

En 1940, con fuerte influencia de Norman Campbell, el Comité Ferguson había concluido que no era posible aplicar los principios de la medición de objetos físicos a variables psicológicas, que no eran cantidades. La propuesta de Stevens que distinguía los niveles de medición nominal, ordinal, de intervalo y de razón, iniciando la *Teoría Operacional de la Medición*, fue precisamente un esfuerzo por extender la noción a otro tipo de variables que, aunque no sean como las de carácter físico, pueden cuantificarse de alguna manera, lo que las hace *cantidades*, en el sentido de la vieja definición según la cual cantidad es todo aquello que puede aumentar o disminuir. Ya en el libro V de los Elementos de Euclides se distinguía magnitud y multitud, que corresponden a la distinción entre cantidades continuas y discretas. (Michell, 2001)

La reflexión teórica en este campo continúa dando lugar a la *Teoría Representacional* (*Representational Theory of Measurement, cfr.* Suck, 2001) y, en lo relativo a la medición de los atributos psicológicos, a la *Teoría de la Medición Conjunta* (*Theory of Conjoint Measurement, cfr*. Fishburn, 2001).

Gracias a estos avances hoy es claro que, si la medición a nivel nominal o clasificación no es un problema en ciencias sociales, como tampoco lo es la ordenación, tampoco la cuantificación estricta debe verse como un obstáculo insuperable. Obviamente hay aspectos para los que es más fácil realizar operaciones que requieren los niveles de medición ordinal y, sobre todo, cardinal, mientras en otros casos esto será más difícil.

Si consideramos el nivel de ingresos de las personas, se pueden hacer comparaciones entre unas y otras, y ordenarlas en ese aspecto, y se pueden establecer unidades más o menos arbitrarias, que se pueden estandarizar y subdividir, sobre todo en una economía con moneda, y tratándose de los ingresos de trabajadores asalariados. El paso de la apreciación *tan rico como Creso*, a la clasificación de los multimillonarios en dólares de la revista *Forbes* ilustra la diferencia. Con una variable más escurridiza, como el racismo, la comparación de los sujetos entre sí para establecer un orden, o la definición de una unidad estándar, es más difícil, pero puede intentarse, sobre todo si la variable se entiende como *manifestaciones* de racismo. Hay escalas ordinales de *actitudes racistas*, pero en casos como este hay obviamente problemas serios para medir a nivel cardinal, para definir una unidad estándar y, peor aún, sus múltiplos y submúltiplos.

Ninguna realidad es en sí misma nominal, ordinal o cardinal. Nosotros podemos —con mayor o menor dificultad— construir los datos y medir. Ni siquiera la propiedad física más medible, la longitud, está ya medida en sí misma. Es el hombre quien construye medidas, efectuando operaciones análogas a las que se deben realizar para

medir realidades más huidizas como la presión atmosférica, el *spin* de los electrones, o ciertas actitudes.

Según Cicourel, para medir en sociología se necesitan "sistemas teóricos que pudieran ser axiomatizados significativamente en forma tal que generen propiedades numéricas que correspondan a (y presumiblemente sean isomorfas con respecto a), por ejemplo, el conjunto de los enteros o de los números reales. En ausencia de ese tipo de sistemas teóricos raras veces podemos medir rigurosamente los eventos sociales". (1964: 4).

Si no podemos formalizar las teorías de que disponemos para analizar la realidad social, sí podemos comenzar por precisar los términos que vamos a utilizar a fin de evitar su uso en forma equívoca. Y después de ese comienzo imprescindible de la ciencia y su lenguaje que es la definición de sus términos, tan precisa y unívoca como sea posible, la búsqueda de niveles superiores de medición es no sólo legítima y posible, sino también deseable e importante, sin que esto quiera decir que los aspectos menos susceptibles de tal tratamiento deban olvidarse, o que no tengan cabida procedimientos cualitativos.

Ralph Sleeper recuerda una anécdota del gran educador norteamericano John Dewey, al dirigirse a un congreso de investigadores educativos que habían estado discutiendo durante algunos días el tema de la medición:

> Dewey comenzó su intervención diciendo que lo que había estado observando durante la semana le recordaba mucho la forma en que se acostumbraba pesar a los cerdos el día de mercado en la granja de su abuelo en Vermont: tomaban una tabla y la equilibraban bien sobre una sólida cerca; luego sujetaban el cerdo en uno de los lados y amontonaban piedras del otro lado hasta que la tabla quedaba equilibrada. Entonces, dijo Dewey, la gente adivinaba a ojo cuánto pesaban las piedras. (Sleeper, 1989)

Si se entiende que, más allá de muchas diferencias entre ciencias de la naturaleza y del hombre, hay una lógica básica común, que incluye la posibilidad de medir, la lección que deberemos sacar de las palabras de Dewey no es que hay que abandonar toda pretensión de medir en ciencias sociales y educación, sino que nos queda mucho camino por andar si queremos perfeccionar nuestras medidas, y que hay que andarlo.

Si en el mercado de Vermont existen hoy sofisticadas balanzas electrónicas, con las cuales los cerdos son pesados con gran precisión, es porque los contemporáneos del abuelo de Dewey, hace más de 100 años, sacaron precisamente esa lección positiva.

Observación —en un sentido amplio, no limitado a la que se hace con la vista— y medición —en sentido amplio también, incluyendo el nivel nominal, el ordinal, y los

cardinales de intervalo y razón— son, pues, nociones equivalentes, y ambas expresan lo mismo que la frase que sirve de título al capítulo: la obtención de información.

En los tres incisos siguientes se presentan tres grupos de acercamientos para obtener información empírica: unos basados en la respuesta a preguntas, otros en observación en el sentido de la que se hace principalmente mediante la vista, y unos más en el análisis de evidencias derivadas de la actividad de los sujetos.

Acercamientos basados en interrogación

Cuestionarios

El cuestionario es la herramienta más conocida del grupo, porque se considera que es fácil de elaborar y aplicar, eficiente para obtener gran cantidad de información a bajo costo. En realidad, elaborar y aplicar este tipo de instrumentos reviste una complejidad importante que el investigador debe reconocer.

Usar cuestionarios en una investigación supone que las personas a las que se aplicarán conozcan la información sobre la que se les interrogará, y que estén dispuestos a darla; un supuesto adicional, prerrequisito de los anteriores, es que los informantes entiendan las preguntas que se les hagan. Cada uno de estos supuestos implica varias cosas.

- Que los informantes entiendan las preguntas

El supuesto inicial, que afecta a los otros dos, es el relativo a la comprensión misma de las preguntas por parte de los informantes, que enfrenta varias dificultades: unas tienen que ver con el vocabulario que se use y su equivocidad, ya que los universos semánticos de quienes responden pueden ser diferentes por razones culturales o de edad y género; otros problemas se derivan de la sintaxis y la claridad de la redacción, amenazada por el uso de frases subordinadas y dobles negativas; influye también la extensión y el fraseo de las preguntas, que conviene sean cortas, pero no demasiado escuetas, y centradas en una sola idea. Con preguntas cerradas es fundamental la calidad y exhaustividad de las opciones de respuesta, y con preguntas abiertas se debe evitar la vaguedad, como la que se da muchas veces con la pregunta *por qué*, o cuando se indaga la opinión sobre si algo es *mucho o poco,* sin que los puntos de referencia sean claros.

- Que conozcan la información sobre la que se les pregunta

No es evidente lo que alguien sabe o ignora sobre lo que se le pregunta. Se puede carecer de cierta información porque realmente se la ignora, pero también porque es irrelevante para el sujeto. Se suele subestimar la complejidad de la información que implica juicios, comparaciones o perspectivas temporales. Muchas personas, incluso entre quienes

tienen educación superior, encuentran difícil manejar cifras grandes, porcentajes, tasas y tendencias, datos comparativos o en perspectiva. La memoria es más engañosa de lo que se piensa, y es frágil aun con datos no demasiado antiguos.

- Que estén dispuestos a proporcionar la información

El supuesto relativo a la disposición de las personas a dar verazmente cierta información a quien la solicita tiene que ver con que se trate de cuestiones públicas o privadas, más o menos íntimas, relativas a conductas lícitas o ilícitas, socialmente aceptables o no, y por ende, que dar la información se perciba como amenazante, en lo que también influye la probabilidad percibida de que se difunda la identidad del informante, o la seguridad de que tal cosa no ocurrirá (anonimato, garantía de confidencialidad). Se pueden encontrar también respuestas que no corresponden a la realidad, pero no se deben a que el sujeto mienta, en sentido estricto, sino que inconscientemente pueden tender a responder de una manera inexacta por la tendencia a hacerlo en una forma socialmente aceptable o deseable, o de manera congruente con otras respuestas (efecto halo), o de manera que sistemáticamente se incline a preferir opiniones de tipo afirmativo en lugar de negativo, o viceversa, o bien posturas intermedias entre las posibilidades que se presentan.

Los dos últimos supuestos no dependen del investigador, sino de los sujetos que cubra el estudio, en lo individual o en conjunto: unas personas son más capaces que otras de informar sobre ciertos fenómenos, y algunas están más dispuestas que otras a hablar de cuestiones personales, pero además en unos contextos culturales ciertos temas pueden ser tabú, mientras que en otros contextos no hay problema para hablar libremente de ellos. El primer supuesto, sobre la comprensión de las preguntas por parte de los informantes, sí depende de los investigadores y afecta a los otros dos: la manera de formular una cuestión y de asegurar el anonimato puede hacer una pregunta más o menos comprensible o amenazante. Los investigadores pueden favorecer o dificultar la apertura de los informantes, si en la presentación del cuestionario establecen la seriedad del trabajo y la credibilidad de la institución que lo promueve, subrayan la importancia de la colaboración del informante y aseguran el anonimato. También es posible suavizar las preguntas más sensibles.

Tipos de preguntas y su calidad

Los cuestionarios se usan sobre todo en encuestas, porque son un medio eficiente de obtener información en estudios de gran escala, aplicándolos a un buen número de sujetos, que solo deben responder preguntas cuyas respuestas se pueden procesar con

herramientas informáticas. Su limitante es que con ellos se pueden explorar solamente cuestiones que reúnan las condiciones mencionadas —que los informantes entiendan las preguntas, conozcan la información sobre la que se les pregunta y estén dispuestos a proporcionar la información— cuestiones sobre las que los informantes pueden expresar lo que piensan mediante respuestas que se pueden plasmar en alternativas claras y unívocas. Aspectos más complejos de la realidad implicarán otros acercamientos, como entrevistas.

Se distinguen, pues, dos tipos básicos de preguntas: abiertas y cerradas. Las primeras suponen que los informantes formulen una respuesta en sus propias palabras, lo que hace necesario procesar después las respuestas, y construir *categorías* para clasificar las respuestas con base en ellas. El rasgo distintivo de las preguntas cerradas es que la respuesta de los informantes se limita a seleccionar una posibilidad entre varias que se le presentan. Las opciones pueden ser sí o no, verdadero o falso, o frases previamente formuladas entre las que se escoge una.

No es indispensable que cada elemento del cuestionario tenga forma interrogativa; puede ser una aseveración que puede ser completada de varias maneras. Por eso en lugar de *preguntas* puede hablarse de *estímulos* o de *ítems*. Indagar sobre el lugar de nacimiento de una persona puede hacerse preguntándole "¿dónde nació usted?", y dándole como opciones de respuesta Aguascalientes, la Ciudad de México, Jalisco, Zacatecas u otro lugar, o bien planteándole la frase "Mi lugar de nacimiento es...", ofreciéndole las mismas opciones para que la complete.

Cuando varias preguntas tienen en común las mismas opciones de respuesta, pueden presentarse agrupándolas, sin repetir cada vez las alternativas de respuesta, en lo que se suele llamar una *batería*, que no hay que confundir con el instrumento que se presenta en el inciso siguiente, la *escala*. En este último caso, como se verá, un grupo de ítems pretende captar información sobre una sola dimensión o variable, mientras que cada una de las preguntas que forman una batería se refiere a una variable distinta.

Unas preguntas de un tipo especial son la que se llaman de filtro, o de contingencia, que se distinguen porque responderlas en una u otra forma dan lugar a un grupo de preguntas subsecuentes, en vez de otras. Si se pregunta a unos alumnos de bachillerato si después de terminar ese nivel educativo piensan seguir estudiando, a quien responda afirmativamente se le puede preguntar a qué tipo de institución o carrera pretende entrar; eso no procedería preguntarlo a quien responda en sentido negativo, a quien se puede preguntar, por ejemplo, si piensa trabajar o casarse.

Se entiende que una pregunta es buena si está formulada de forma que los sujetos a quienes se dirija la entiendan tal como pretendía el investigador, y todos de igual

forma, de suerte que se maximiza la probabilidad de obtener información objetiva y consistente, sin olvidar que habrá que procurar que las preguntas se refieran a temas que los sujetos conozcan, y que estén dispuestos a informar sobre ello.

El que los informantes entiendan lo que se les pregunta varía según a quién se dirija un cuestionario. Tratándose de alumnos de corta edad, se deberá cuidar especialmente la claridad de las preguntas y su extensión, ya que instrumentos largos o complejos producirán información de baja calidad. En el caso de maestros y directores, en cambio, será más importante cuidar la tendencia a responder lo deseable en lugar de lo real, así como los temas delicados o amenazantes.

Un caso delicado es el relativo a la aplicación de cuestionarios a poblaciones de contextos lingüísticos diferentes, como personas de lengua materna indígena, con distinto grado de castellanización, para quienes pueden ser confusos términos de sentido muy claro para aquellos sujetos cuya lengua materna es el español.

No se puede valorar en abstracto si una pregunta es o no inteligible: hay que hacerlo en función de las personas a quienes se aplique. Para juzgar si una pregunta es adecuada hay que verificar que las personas a quienes se dirigirá la entiendan de la misma manera. Por eso muchas veces hay que adaptar las preguntas formuladas en un contexto cuando se quieren aplicar en uno diferente. Un ejemplo es el caso de las adaptaciones terminológicas que hay que hacer al aplicar cuestionarios en contextos en que se hablan variantes dialectales del español. Un indicador del nivel socioeconómico puede referirse al medio de transporte que se usa. Para explorar el tema se puede formular una pregunta con varias opciones de respuesta: a pie, en bicicleta, en autobús, en metro, taxi, o automóvil propio. Unas opciones pueden ser lógicas en un contexto urbano, pero no en uno rural o viceversa; al mismo tipo de transporte en un lugar se le puede llamar autobús, en otro camión, en otro guagua, y en otros más combi, pesera, etcétera. Para asegurar una buena comprensión, sobre todo con alumnos de corta edad, el uso de un término u otro no es indiferente.

Para garantizar que una pregunta se entienda igual no basta evitan los errores más burdos al formularla, y si esto no se asegura las respuestas no serán comparables, porque en realidad los sujetos no responderán la misma pregunta, y sus respuestas no se podrán agregar ni analizar como muestras de unos rasgos de los informantes.

RECUARO 3.1. LAS PRUEBAS DE RENDIMIENTO

En un cuestionario que solo busca información sobre cierto tema no hay respuestas correctas e incorrectas. En cambio, en los instrumentos que exploran conocimiento de un tema (pruebas de rendimiento), y lo hacen con *preguntas de opción múltiple*, una de las opciones es la respuesta correcta a la pregunta formulada, y las otras son incorrectas, o distractores. El tratamiento de la información obtenida con este tipo de instrumentos es distinto al que se da a las respuestas de un cuestionario simple, e involucra complejos desarrollos estadísticos (Teoría Clásica de los Tests, Modelos de Respuesta al Ítem, etc.) de los que se tratará brevemente en el Cap.4.

Organización del conjunto del cuestionario

Pueden distinguirse tres partes que estarán presentes en cualquier cuestionario:

- Una parte inicial que incluye la presentación del instrumento, con información sobre los propósitos del estudio en el que se inscribe; la manera en que se seleccionó los sujetos a quienes se aplica; alguna motivación relativa a la importancia de la información que se obtendrá; la credibilidad de la institución que patrocina el trabajo; cómo se garantizará la identidad de los informantes (confidencialidad), entre otras cosas. En esta parte se sitúan, desde luego, las instrucciones necesarias para responder el instrumento.
- Una parte central, con el grueso de las preguntas, con una secuencia que tenga en cuenta tanto el orden lógico como el psicológico, tratando de que las preguntas más sencillas y menos sensibles vayan antes de las más complejas y potencialmente amenazantes. Si el instrumento tiene partes que sólo deberán responder algunos sujetos, en función de sus respuestas a preguntas filtro, las instrucciones al respecto deberán ser claras.
- Una parte final en la que se agradece la colaboración de los participantes, se reiteran las seguridades de confidencialidad, informando sobre la manera en que se difundirán los resultados, asegurando que los participantes podrán tener acceso a ellos, y se ofrece algún medio de contactar a los responsables. Antes de esto podrán ir preguntas sobre datos demográficos y personales de los participantes, que no conviene sean las primeras del instrumento, porque pueden hacer que surjan dudas sobre el tema del anonimato.

La aplicación

En cuanto a la forma de aplicar un cuestionario, se distingue:

- Autoadministración, cuando el instrumento se hace llegar a los sujetos para que cada uno lo responda cuando pueda y lo regrese. Antes esto se hacía por correo; en la actualidad cada vez más vía Internet.
- Aplicación por otras personas, los encuestadores. En este caso la manera tradicional implica encuentro personal del encuestador con el respondiente, pero puede hacerse también por teléfono o en línea.

Cada una de estas formas tiene ventajas y desventajas, sobre todo en cuanto a la tasa de respuestas que se puede esperar obtener, siempre menor en aplicaciones que no se hacen en persona.

La aplicación en línea, cada vez más frecuente, tiene importantes ventajas, ya que los paquetes de software para ello (como *Lime Survey, Survey Mokey,* o las opciones de Google) reducen casi a cero el esfuerzo que de otra manera representa la captura de datos, facilitan el uso de preguntas filtro y pueden simplificar el esfuerzo que para los informantes implica responder, pero suelen tener tasas de respuesta bajas, que no necesariamente compensa la posibilidad de aumentar el tamaño de la muestra, ya que los sujetos que responden pueden ser muy distintos de los que no lo hacen.

La importancia de la aplicación de un instrumento puede apreciarse si se considera que cada etapa de un estudio influye en la calidad de los resultados: si cualquier etapa se hace muy mal, el resultado final se verá gravemente afectado, aunque todas las demás etapas se hayan hecho perfectamente. Por bien hecho que esté un cuestionario, y por rigurosos que sean los análisis que se hagan con la información obtenida, si la aplicación no se hizo bien la calidad de los resultados se verá severamente hipotecada.

Es importante señalar que la dificultad de una aplicación es mayor a medida en que aumenta el número de casos. Si solo se piensa en que la probabilidad de generalizar los resultados de una encuesta aumenta al crecer el número de sujetos, se puede tender a aumentar el tamaño de la muestra sin pensar en que con ello puede reducirse el error *de muestreo*, pero al mismo tiempo crecerán los *errores no derivados del muestreo*, que se pueden deber a defectos de los instrumentos, pero también a fallas de la aplicación.

Una encuesta con una muestra bien hecha de menor tamaño tiene ventajas respecto a un censo o un estudio basado en una muestra muy grande, no sólo por costo, sino también por la calidad de la información que se obtiene. En educación, una buena muestra, además de menor costo, implica menos trabajo para alumnos y escuelas. Para comprender mejor este punto convendrá ver lo relativo a muestreo, en el Capítulo 4.

Reconociendo que preguntas mal formuladas pueden traer consigo el resultado de proporcionar información de mala calidad, independientemente de la muestra, se ha avanzado mucho en lo que se refiere a la formulación de las preguntas. La obra que se considera el culmen de la primera época de las encuestas en este sentido, *The Art of Asking Questions*, de Stanley Payne, se publicó en 1951. El título de esta obra indica que, en aquella fecha, formular buenas preguntas se consideraba una tarea que no era posible desarrollar con bases plenamente rigurosas, sino que pertenecía más bien al terreno artístico. Tres décadas y media más tarde, hace ya más de tres, en la introducción a una obra de Converse y Presser (1986), el editor de la serie en que fue publicada señalaba que, sin que obstara el título de la obra citada, a mediados del siglo pasado Payne consideraba que la formulación de preguntas *no se debería guiar solamente por la intuición y la experiencia, sino también por la evidencia derivada de experimentos rigurosos,* pero añadía que ese tipo de trabajo experimental no se había desarrollado en las tres décadas transcurridas entre ambas publicaciones (Sullivan, en Converse y Presser, 1986: 5); en sentido similar se puede ver Sudman y Bradburn (1987).

Otros textos de fines del siglo XX muestran los pasos que se han dado en lo que se refiere a la formulación de mejores preguntas para instrumentos de tipo cuestionario (*cfr*. Fowler, 1995; Suskie, 1992). Por otra parte, con la revolución cognitiva comenzaron también a aparecer trabajos que muestran la complejidad de los procesos que tienen lugar en la mente de un informante al que se presentan preguntas en un cuestionario o instrumento similar, primero para entenderlas, luego para localizar la información en distintos niveles de memoria, y finalmente para decidir qué responder. Estas ideas se desarrollarán al tratar de las *entrevistas cognitivas*.

Escalas

Un tipo especial de cuestionario, la *escala*, trata de explorar posturas subjetivas de los respondientes, como sus actitudes, y no su conocimiento de hechos objetivos. Para saber algo respecto a esos aspectos más *ocultos* de la realidad es necesario hacer inferencias a partir de aspectos que se puedan observar, como la expresión verbal de opiniones o actitudes, o la manifestación de conductas que las reflejen, lo que contrasta con lo que se puede hacer para captar aspectos de carácter más *objetivo*, o más susceptibles de captarse fácilmente, como la edad o la estatura. Se puede saber que una persona está triste si lo dice, o bien si la vemos llorar, aunque es posible que nos engañe al responder, o porque es un buen actor. Para designar esos aspectos más difíciles de captar, cuya existencia se infiere a partir de sus manifestaciones perceptibles se suele usar la expresión *constructos latentes.*

La idea que subyace la construcción de escalas tiene que ver con el problema que se mencionó sobre la comprensión de las preguntas de un cuestionario: quien tiene una actitud o sentimiento no siempre lo puede expresar verbalmente. Cuando una persona experimenta una emoción fuerte suele decir que *no tiene palabras* para expresarlo. Por ello buscar información sobre esos aspectos con una sola pregunta es poco fiable, ya que es difícil encontrar una formulación que diga exactamente lo mismo a todos los sujetos. Es más confiable la información que se puede obtener de un conjunto de preguntas que la derivada de una sola, con tal que se cumpla una condición básica: que todas se refieran realmente al mismo aspecto de la realidad; en términos técnicos, que la escala cumpla la condición de *unidimensionalidad*.

Si todos los ítems de una escala pertenecen a la misma dimensión, si se refieren a manifestaciones del mismo *constructo latente*, cada uno capturará un matiz distinto del mismo, y es más probable que el conjunto de las respuestas de los sujetos a los que se aplique la escala represente correctamente lo que pretende, que la respuesta a una sola pregunta. Técnicas estadísticas como el análisis factorial o los modelos de respuesta al ítem permiten verificar si una escala cumple la condición de referirse a una sola dimensión, y si permite *observar* lo que no es posible captar a simple vista, así como *medirlo*, con un nivel no solo clasificatorio (nominal), sino ordinal.

La literatura especializada sobre escalas (*v. gr.* Morgenstern y Keeves, 1997; De Vellis, 1991; Morales Vallejo, Urosa Sanz y Blanco Blanco, 2003) distingue las de tipo Gutmann, Thurstone o Likert. Las de uso más frecuente son las últimas, en las que se explora una actitud pidiendo a los sujetos que expresen su postura respecto a un conjunto de aseveraciones relativas a dicha actitud, seleccionado una opción sobre su grado de acuerdo o desacuerdo respecto a cada una.

Cuestionarios con otros tipos de preguntas

Buscando formas de explorar lo que piensan o sienten los informantes respecto a un tema, se han desarrollado modalidades especiales de preguntas para subsanar limitaciones de las más simples y directas. Estas formas especiales de pregunta implican más extensión y, por ende, más tiempo por parte de los sujetos, y más disposición y/o capacidad para leer y comprender estímulos de mayor complejidad. Esta desventaja debería verse compensada por la ventaja de mayor uniformidad en la comprensión de lo que se indaga, por parte de los informantes.

Cuestionario con referentes explícitos. Una forma de reducir el riesgo de que una pregunta sea comprendida de manera desigual por los informantes es simplemente presentar

de manera explícita aspectos que, de no ser así, se dejan a la interpretación de cada sujeto. Un ejemplo se ha usado en los cuestionarios que se aplican a estudiantes, como parte de la evaluación de los maestros. En este caso, típicamente las preguntas se refieren a distintas dimensiones de las actividades de los docentes que se consideran adecuadas o no, como su asistencia y puntualidad, el que presente el programa al inicio del curso, la imparcialidad al calificar a los alumnos, o el carácter respetuoso de la manera en que los trata.

En general las respuestas a tales preguntas no se deben expresar en términos de sí o no, sino de una gradación mucho-poco, con más-menos frecuencia o similares, y es claro que el sentido que tienen las opciones de respuesta no necesariamente es el mismo para todos los informantes. El que un alumno señale que su maestro falta mucho o poco puede significar cosas muy distintas en escuelas en las que el ausentismo sea excepcional, en comparación con otras en las que sea frecuente. Una forma de reducir la equivocidad de las respuestas es formular las opciones de respuesta explicitando los referentes que considera la instancia que aplica el cuestionario, como muestra el ejemplo siguiente:

Anota el número de clases de la materia que imparte este maestro que debe haber semanalmente: ( ) Teniendo en cuenta lo anterior, durante el mes pasado tu maestro:			
( ) Nunca faltó	**( ) Faltó una vez**	**( ) Faltó dos o tres veces**	( ) Faltó cuatro veces o más
Todos los maestros deben entregar a cada alumno un ejemplar del programa del curso al inicio del mismo y explicarlo de manera que todos los alumnos queden enterados de su contenido.			
En cuanto al programa de la materia, tu maestro:			
( ) Lo entregó el primer día de clase	**( ) Lo entregó en la primera semana de clase**	**( ) Lo entregó después de la primera semana de clase**	**( ) No entregó el examen**
También en cuanto al programa de la materia, tu maestro:			
( ) Lo explicó bien y aclaró todas las dudas	**( ) Lo explicó y aclaró algunas dudas**	**( ) Lo explicó mal y dejó muchas dudas**	**( ) No explicó el examen**

Preguntas con factores combinables (factorial surveys). Para distinguir grados de intensidad de cierta opinión o postura, una forma de pregunta consiste en presentar frases similares en que se hacen variar sistemáticamente uno o más elementos, cuya presencia o ausencia se considera indicativa de mayor o menor intensidad del constructo latente que se quiere medir. Esos elementos cambiantes son *factores*. Una desventaja de las preguntas con factores es que incluso un número no muy grande de estos da lugar

a uno mucho mayor de combinaciones, lo que a su vez hace necesario que el instrumento se aplique a una muestra bastante grande, para que haya suficientes respuestas en cada una de las combinaciones posibles.

Ordenamiento de opciones de respuesta (Rank ordering or ranking scales). En estas preguntas se pide a los sujetos que clasifiquen objetos, eventos o personas, según cierto aspecto que pueda dar lugar a un ordenamiento. En estudios de opinión sobre las cualidades de personalidades de la política se puede presentar los personajes a calificar en orden alfabético, y pedir a los sujetos que los ordenen según el grado en que, a su juicio, tengan las cualidades de interés.

Estimación de magnitudes (magnitude estimation scales). En vez de ordenar a las personalidades sobre las que se explora la opinión de los respondientes, se puede pedir a estos asignar a cada personalidad un valor numérico (*v. gr.* de 1 a 10) que represente la opinión del informante sobre el grado en que cada personaje tiene cierta característica. Las respuestas también se pueden expresar ubicando a los sujetos en dibujos que representan un orden, por ejemplo, un termómetro o una escalera, lo que puede ayudar a personas con una mentalidad más gráfica.

Medición de la intensidad de la opinión o actitud. En escalas cuyas respuestas se expresan con opciones como *totalmente de acuerdo, de acuerdo, en desacuerdo, totalmente en desacuerdo*, o similares, es fácil confundir dos dimensiones que se relacionan, pero no son idénticas: el carácter más o menos extremo de la postura, y su intensidad. Una persona puede estar completamente de acuerdo con cierto punto de vista, pero no creer que es particularmente importante. En sentido contrario, alguien puede considerar muy grave cierto asunto, pero no tener una postura bien definida al respecto. Un ejemplo es el de las opiniones a favor o en contra de un tema polémico, como la despenalización del aborto. Un estudio sobre el tema preguntó tanto a los que declararon estar completamente a favor, como a los que dijeron estar totalmente en contra, qué tanta importancia daban al tema, y el resultado fue que los opositores a la despenalización (*pro vida*) tenían seis veces más probabilidad de manifestar que el punto les importaba mucho, en contraste con los que estaban a favor de dicha medida (*pro choice*).

Elección forzada. Se ha constatado que unas personas se sienten más inclinadas a responder afirmativamente, a cualquier pregunta. Esta tendencia se designa en inglés con la expresión *acquiescence response bias*, que se ha encontrado más entre personas de baja escolaridad, y en algunas subculturas. Una forma de reducir el impacto de esa tendencia (que puede ir en sentido contrario, con personas que tienden a responder negativamente a cualquier pregunta (*naysayer, contreras*) es mediante el uso de las

preguntas *de elección forzada*, que plantean expresamente las dos (o más) opciones de respuesta, y obligan a los informantes a pronunciarse:

> *Considera Usted que, en general, los hombres están mejor preparados para la política emocionalmente, que hombres y mujeres están igualmente preparados, o que las mujeres están mejor preparadas emocionalmente para la política:*
> *( ) En general los hombres están mejor preparados*
> *( ) En general hombres y mujeres están igualmente preparados*
> *( ) En general las mujeres están mejor preparadas*

Otras variantes de instrumentos escritos

Autoreportes, bitácoras y diarios. Son otra variante del cuestionario y, como el término mismo indica, consisten en informes proporcionados por los informantes, por ejemplo, maestros. Pueden ser no estructurados (en los que los sujetos describen con sus propias palabras lo que han hecho durante cierto lapso) o estructurados (con formatos predefinidos para informar sobre la frecuencia con que llevaron a cabo ciertas actividades).

Como en cualquier instrumento de este grupo, la limitación que se suele atribuir a los *auto-reportes* tiene que ver con el riesgo de que los sujetos no informen sobre lo que realmente hicieron, sino sobre lo que consideran que debieron haber hecho, o sea sobre lo considerado deseable.

En el caso de investigaciones sobre las prácticas docentes, un estudio al respecto (Koziol y Burns, 1986) analizó la calidad de la información reportada por nueve profesores de inglés, comparándola con la que dieron sus alumnos y con la que se obtuvo observando sus clases. El instrumento utilizado era altamente estructurado, con 77 ítems que describían otras tantas prácticas, de las cuales 37 se referían a la enseñanza de vocabulario y deletreo, y 40 a gramática y sintaxis.

Los maestros hicieron *auto-reportes* semanales en seis ocasiones, y uno de resumen al final del período. En los primeros debían señalar, para cada una de las 77 prácticas, cuántos días la habían utilizado en la semana en cuestión, por lo que sus respuestas iban de 0 a 5. En el informe resumen, sobre las seis semanas, las opciones de respuesta para cada práctica se expresaban en términos de nunca (1); una o dos veces en el período (2); al menos tres veces en el período pero no semanalmente (3); al menos una vez por semana (4). Un instrumento similar al de resumen de los maestros se aplicó a los alumnos al final del período del estudio, y se hicieron observaciones semanales no estructuradas de una clase impartida por los maestros estudiados, que luego se codificaron.

La información sobre prácticas de los maestros obtenida con *autoreportes* basados en el instrumento descrito mostró razonable consistencia interna y también fue congruente con la información que dieron los alumnos y la que se obtuvo mediante observaciones en aula, lo que hace ver que el riesgo de obtener información no confiable puede ser evitado.

Bitácoras y diarios son textos en que los sujetos describen sus actividades durante cierto tiempo. Son un tipo de autoreportes, pero estos pueden hacerse en forma esporádica, mientras que en bitácoras y diarios es esencial la periodicidad con que se da información (diaria, semanal) lo que aumenta la confiabilidad, ya que es menos probable que se distorsione la realidad si las actividades se reportan varias veces a lo largo del tiempo. Se puede dejar que cada informante reporte en forma libre sus actividades o usar una guía estructurada para que lo haga.

Los pros y contras de estas alternativas son los mismos que en cualquier cuestionario, pero la carga de trabajo que supone llevar un diario abierto durante muchos días es pesada, mientras que llenar un formato estructurado durante el mismo lapso es más sencillo.

Estas herramientas son útiles para explorar la *oportunidad de aprendizaje*, en el sentido del grado en que *el currículo implementado o enseñado* cubre lo que el *currículo planeado o prescrito* establece, lo que incide en el currículo alcanzado o logrado (Rowan, Camburn y Correnti, 2004; Rowan y Correnti, 2009). *El currículo implementado* se suele explorar con cuestionarios aplicado al final del curso en que los maestros informan retrospectivamente sobre los temas cubiertos, o mediante la observación de clases. En ambos casos la información puede ser de baja calidad:

> A lo largo de un año escolar el maestro de primaria típico dará 140 o más días de clase... a 20 o 30 alumnos de su grupo, a veces en forma diferenciada para alumnos individuales o subgrupos. Aún más: en un día cualquiera, las actividades de enseñanza típicamente se desarrollarán según varias dimensiones; un maestro normalmente cubrirá varios objetivos con distintos niveles de demanda cognitiva en un solo día, trabajando con distintos arreglos conductuales y utilizando una variedad de técnicas de enseñanza propias de cada tema... algunos rasgos se repetirán a lo largo del año, pero otros no, por lo que las prácticas son multidimensionales, pero además altamente variables a lo largo del año. (Según Rogosa, Floden y Willet, 1984, citados por Rowan y Correnti, 2009: 121).

La variabilidad de prácticas de los maestros en distintos momentos y la que distingue unos maestros y otros, harían necesario un número elevado de observaciones (15-30) para asegurar una razonable consistencia de la información; el costo de este

tipo de estudios se eleva en consecuencia, lo que hace atractiva la opción de los diarios. (Rowan, Camburn y Correnti, 2004: 14-17)

Al pedir a los maestros que reporten de inmediato lo que hicieron en el aula cierto día, los diarios ganan en confiabilidad respecto a encuestas aplicadas una vez al año, pues los problemas de recordar el pasado y basarse más en interpretaciones de lo deseable se reducen sustancialmente; y al pedir que los reportes se hagan durante muchos días la muestra de prácticas reportadas es más representativa del universo que la información que dan pocas observaciones al año.

Pero si el número de días a reportar aumenta, la carga para los maestros se hace mayor. Para estimular las respuestas se puede ofrecer algún pago a los maestros, como hicieron Rowan y colaboradores, quienes también pusieron un número de teléfono gratuito a disposición de los maestros para resolver sus dudas. Como resultado se obtuvieron tasas de respuesta del orden de 90% y, según reportan los investigadores, datos de calidad ligeramente inferior a la de los derivados de observaciones en el aula. (Rowan y Correnti, 2009: 122)

Se construyeron escalas que combinaran las respuestas a varios ítems, y el análisis mostró que 72% de la variación del tiempo de enseñanza se dio entre unos días y otros; 23% entre maestros de una misma escuela; y sólo 5% entre diferentes escuelas. La desviación estándar de la distribución del tiempo de enseñanza *entre días* fue de 45 minutos, o sea que en 15 de cada 100 días el tiempo dedicado a la lectura fue realmente de *cero minutos,* aunque en el sistema americano suele suponerse se dedican diariamente 90 minutos a la lectura y la lengua, y aunque en promedio los maestros estudiados dedicaban 80 minutos diarios a esa área.

Estos datos implican que, para tener información suficiente sobre las prácticas de un maestro son necesarios reportes de unos 20 días al año (Rowan y Correnti, 2009: 123); un número similar de observaciones, mucho más costosas, sería igualmente necesario para tener una muestra suficiente de lo que pasa en el aula.

Diarios en línea. El uso de diarios en línea para recabar información sobre prácticas docentes puede reducir de manera importante el peso del trabajo que cada maestro, así como el de los investigadores. Si se cuenta con una plataforma adecuada, esta forma de recabar la información permite enviarla rápidamente y hace innecesario que los investigadores deban capturarla posteriormente para su análisis, ya que se almacena en el sistema al mismo tiempo que es registrada por los informantes. Los principios metodológicos para diseñar un instrumento para recabar información en línea sobre prácticas docentes son los mismos que se deben aplicar tratándose de diarios con lápiz y papel.

Cuestionarios con viñetas. El problema derivado de que las preguntas se entiendan de manera distinta puede tener que ver con el uso de términos técnicos, familiares para quienes elaboran un instrumento, pero no necesariamente para quienes deben responderlo, como maestros o alumnos, lo que agrava el riesgo de que la comprensión de los informantes no coincida con la de los investigadores.

Abundan ejemplos en el campo educativo. Si se pregunta a unos maestros, por ejemplo, si utilizan *trabajo colaborativo* o *evaluación formativa*, y responden afirmativamente, es difícil saber qué entendían exactamente por esos términos, y es probable que haya una diversidad de comprensiones al respecto, lo que invalida seriamente las conclusiones que se saquen con base en tales respuestas. Por ello se ha desarrollado el tipo especial de preguntas llamadas *viñetas* que, en lugar de pedir opiniones o información sobre prácticas en términos teóricos y abstractos, lo hacen mediante descripciones precisas de conductas concretas, contextualizadas, pidiendo a los respondientes que indiquen si sus propias formas de trabajar se aproximan más o menos a las descritas en la pregunta o viñeta.

Una pregunta convencional para maestros de primaria podría indagar si están de acuerdo con la idea de utilizar estrategias de evaluación formativa. Una pregunta alternativa, con un estímulo en forma de *viñeta*, podría preguntarles si están de acuerdo con una situación descrita en términos concretos, como la siguiente:

> Además de los exámenes bimestrales que le da el supervisor, la maestra Rosa aplica cada semana pruebas más cortas que ella prepara. Para corregir las respuestas dadas por los alumnos a cada pregunta, Rosa pide a unos niños que digan cómo la respondieron y los hace opinar sobre cuál es la respuesta acertada y por qué.

Es más probable que los maestros entiendan igual la versión con una viñeta, que la que usa la expresión *estrategias de evaluación formativa*.

Este tipo de preguntas se utiliza desde hace tiempo en estudios de otros temas, como sobre actitudes discriminatorias o laborales (*cfr*. Martin *et al.*, 1991; Martin, 2006), pero en educación su uso es más reciente y no hay mucha investigación sobre la calidad de la información que permiten obtener.

Un trabajo de Stecher y colaboradores (2006) estudió las prácticas de enseñanza, explorando el grado en que eran consistentes con las orientaciones innovadoras definidas en los estándares para la enseñanza de matemáticas y ciencias *(reform oriented*), en contraste con la enseñanza *tradicional*. Para validar la información se utilizó una

combinación de cuestionarios simples, cuestionarios basados en viñetas, diarios de los maestros y observaciones de su trabajo en aula. Se hicieron también entrevistas cognitivas a una submuestra de maestros, en relación con las viñetas.

La preparación de viñetas comenzó con la operacionalización de los conceptos de currículo y práctica de enseñanza innovadora, en una taxonomía de 23 elementos; se identificaron elementos susceptibles de medirse con viñetas; se escogieron dos temas de matemáticas y cuatro situaciones en que se pueden manifestar prácticas innovadoras. Los elementos se integraron en escenarios más amplios, que incluían una contextualización y presentaban las situaciones, seguidas por posibles formas de actuar en forma innovadora o tradicional, pidiendo al maestro indicar qué tan probable sería que actuara en la forma descrita en cada opción. Se definieron dos métodos para asignar una calificación global, situando a cada sujeto en un continuo innovador-tradicional. Se hicieron análisis para estimar confiabilidad y validez, contrastando con la información obtenida mediante los cuestionarios simples, los diarios y las observaciones del trabajo en aula.

Los resultados apoyan la idea de que la información obtenida mediante viñetas es de buena calidad en unos aspectos, pero no en todos; entrevistas cognitivas con maestros participantes mostraron que no siempre su interpretación de las descripciones de las viñetas coincidía con la de los investigadores. Una conclusión es que preparar buenas viñetas implica un trabajo mucho mayor al que supone hacer un cuestionario simple. (Stecher *et al.,* 2006: 120)

La entrevista

En general

Este acercamiento forma parte del grupo de los que buscan obtener información a partir de la respuesta de los informantes a preguntas que se les formulan, pero difiere de los vistos hasta ahora en que no lo hace con apoyo en algún tipo de instrumento escrito, sino en forma oral, lo que implica interacción del sujeto con un entrevistador, sea cara a cara o a distancia, por ejemplo, por vía telefónica. Como en el caso del cuestionario, aplica la distinción entre entrevistas estructuradas, semi-estructuradas y completamente abiertas. El primer caso no se distingue de la aplicación de un cuestionario por un encuestador, por lo que hablar de entrevista refiere realmente a variantes menos estructuradas, aquellas en las que *el propósito es comprender a los informantes en sus propios términos*. Este tipo de entrevistas:

> Contrastan con las encuestas y las pruebas, que pueden administrarse en forma oral, pero tienen un alto grado de estructuración... para identificar elementos particulares de información o determinar la frecuencia de diferentes respuestas según categorías preestablecidas... Las entrevistas abiertas, llamadas también cualitativas, permiten que un informante exprese significados en sus propias palabras, y marque la dirección del proceso de entrevista. (Brenner, 2006: 357)

Una entrevista puede basarse en distintos supuestos teóricos y derivarse de disciplinas, como la antropología cultural, las ciencias cognitivas o la psicología del desarrollo, y a partir de ello distingue dos enfoques: uno *inductivo*, en el que las categorías analíticas se construyen a partir de lo que dice la persona entrevistada (como en trabajos basados en la teoría fundamentada o *grounded theory*), y otro *deductivo*, en el que el investigador usa constructos derivados de una teoría, como la *teoría crítica.* (Brenner, 2006: 357-361)

Kvale y Brinkman identifican siete pasos en el desarrollo de una entrevista abierta:

a. Precisión del tema, con base en el propósito y las preguntas de investigación.
b. Diseño, que se refiere a la preparación de varios tipos de preguntas que se pueden formular en distintos momentos de la entrevista:

- Preguntas introductorias (de calentamiento) y de seguimiento, cuando el entrevistador detecta focos rojos sobre los que quiere profundizar.
- Preguntas generales y preguntas específicas, las segundas para buscar información sobre algún aspecto en particular.
- Preguntas directas e indirectas, estas que al indagar sobre un aspecto esperan obtener información sobre otro.
- Preguntas de estructura, de transición para pasar a un tema distinto, y de interpretación, para aclarar algo sobre el tema que se está tratando.

c. Realización, incluyendo el establecimiento de una buena relación inicial (rapport), el respeto al entrevistado, el cuidado de la apariencia y el lenguaje corporal, el contacto visual y la no invasión del espacio del entrevistado.
d. Transcripción, sea de notas escritas o de una grabación, buscando exhaustividad y fidelidad.
e. Análisis, con etapas de codificación, condensación e interpretación; simple o complejo; con categorías emergentes o basadas en una teoría, etc.

f. Verificación, para cuidar la validez, la confiabilidad y la generalizabilidad.
g. Reporte, para comunicar resultados respetando normas científicas y éticas. (Kvale y Brinkmann, 2009)

El número de sujetos que se puede incluir en un estudio basado en entrevistas abiertas debe ser mucho más reducido que en uno que use cuestionarios y escalas, según Kvale y Brinkmann (2009) no más de 10 a 12.

El *Manual* de Gubrium y Holstein (2002) sobre entrevistas distingue:

- Entrevistas en el marco de encuestas, cualitativas, en profundidad, de historias de vida, y con grupos focales.
- Sujetos especiales: niños y adolescentes, hombres y mujeres, ancianos, enfermos, miembros de élites sociales o personas de ciertos grupos étnicos.
- Contextos: intercultural, en marco periodístico, médico, educativo, con propósitos de empleo, o incluso forense.

La obra discute cuestiones técnicas sobre cómo promover respuestas, el trato de personas reticentes, las diferencias de las entrevistas cara a cara, por teléfono o asistidas por computadora y por internet. En su última edición el Manual analiza tendencias recientes, como la autoetnografía y la auto-entrevista feminista, el uso de técnicas de análisis del discurso, metodología Q o programas de software para análisis cualitativo. (Gubrium, Holstein, Marvasti y McKinney, 2012)

Entrevistas focalizadas y grupos de enfoque (focus groups)

Se trata de un tipo especial de entrevista, que se suele identificar con las que se hacen en grupo, y se asocia con una obra de Merton, Fiskie y Kendall, publicada en forma de libro en 1956, cuyos antecedentes se remontan a la década anterior (Merton y Kendall, 1946).

En un artículo de 1987 retomado como *Introducción a la segunda edición* de la obra (Merton, Fiskie y Kendall, 1990), Merton precisa sin embargo que el trabajo se refería de manera general a la entrevista focalizada, que se puede hacer con individuos, con varios individuos que no formen un grupo real, o también con grupos propiamente dichos.

Recordando que la técnica se desarrolló en el campo de los estudios sobre comunicación y propaganda, Merton *et al.* identifican como distintivo de una entrevista focalizada el que previamente se pide a unas personas que expresen sus ideas sobre ciertos contenidos, y que el investigador analiza las respuestas y formula hipótesis sobre lo que

la situación implica para los informantes, con lo que prepara una guía de entrevista para explorar la experiencia subjetiva de esas personas y captar la forma en que dan sentido a las situaciones vividas. La técnica se entiende, pues, como complemento de una investigación experimental o de estudios sobre la manera en que las personas responden a las situaciones que encuentran en la vida real. (Merton, Fiskie y Kendall, 1990: 3, 11)

Haciendo eco a una observación del filósofo Kenneth Burke —*Una manera particular de ver es, al mismo tiempo, una manera de no ver, porque enfocar la atención en el objeto A implica descuidar el objeto B*— Merton, Fiskie y Kendall destacan que la perspectiva cualitativa de las entrevistas focalizadas se debe complementar con una perspectiva cuantitativa, evitando verlas como *medios fáciles y rápidos de llegar a conclusiones*. Como cualquier técnica, tienen pros y contras,

> [...] el detalle cualitativo que ofrecen las entrevistas focalizadas grupales tiene sus costos porque sólo puede llevar a nuevas hipótesis sobre el carácter de las respuestas y sus razones, y se necesitaría trabajo cuantitativo adicional para someter a prueba esas hipótesis. El punto es que una investigación cualitativa limitada en principio no puede decir nada sobre la distribución y alcance de los patrones de respuesta identificados tentativamente. La medicina descubrió hace tiempo que las observaciones clínicas no sustituyen a la investigación epidemiológica. (1990: XXI-XXII)

Merton destaca además que la entrevista focalizada se usó inicialmente sobre todo en estudios de mercado, pero que no debe limitarse a ese campo de aplicación, sino que es:

> [...] un conjunto de procedimientos para recolectar y analizar datos cualitativos, que nos pueden servir para alcanzar una comprensión ampliada, en una perspectiva sociológica y psicológica, en cualquier esfera de la experiencia humana. (1987, en Merton, Fiskie y Kendall, 1990: XXXI)

La obra se refiere al momento *retrospectivo* de la técnica, cuando se analizan las respuestas previas y se prepara la guía; los criterios para escoger materiales de entrevista apropiados; elementos evocadores a usar, considerando su *especificidad y profundidad*, y el grado en que toman en cuenta el contexto de los entrevistados. Se identifican variantes de entrevista focalizada con grupos, y se recomienda un tamaño de 10 a 12 personas, con una composición homogénea. En cuanto al *arreglo espacial* más con-

veniente, deberá ser uno en el que el entrevistador no ocupe un lugar especial que le confiera carácter de autoridad. Se discuten ventajas (menos inhibiciones, un rango mayor de respuestas, disminución de olvidos) y desventajas (excesivo peso de algunos participantes y discusiones que pueden inhibir la participación de otros, o la dificultad para mantener continuidad). Se recomiendan procedimientos para optimizar la experiencia y propiciar interacción, y distinguir silencios embarazosos improductivos de otros que indican que se están incubando ideas, y se sugiere cómo reaccionar ante interrupciones o contrarrestar el excesivo peso de unos miembros del grupo. (Merton, Fiskie y Kendall, 1990)

Hay abundante literatura sobre la técnica de grupos de enfoque, como variante de la entrevista focalizada (Stewart y Shamdasani, 2015; Krueger y Casey, 2015).

Entrevistas cognitivas

Como se adelantó al fin del apartado relativo a cuestionarios, la revolución cognitiva mostró la complejidad de los procesos que implica responder una pregunta, para entenderla, localizar la información en la memoria, y decidir qué responder.

Un principio clave de la perspectiva conductista es que solo pueden ser objeto de estudio las acciones de los sujetos, *conductas*, manifestaciones externas de aspectos internos como sentimientos o intenciones, inaccesibles a la investigación. Lo que dio un carácter revolucionario a la perspectiva cognitivista fue el cuestionamiento de ese supuesto, con el planteamiento de que es posible, y fundamental, que la investigación intente captar aspectos de la actuación de los sujetos que no son inmediatamente aparentes. Esta postura reconoce que para estudiar esos aspectos es necesario hacer inferencias a partir de manifestaciones externas, pero subraya que esto no es excepcional en la ciencia sino inherente a todo conocimiento humano. En todo nivel de conocimiento hay algún tipo de inferencia, y lo que hay que hacer no es intentar eliminarla, sino sustentar su solidez.

Es imposible entrar a la mente de otra persona para saber si está triste o alegre, pero podemos inferirlo viendo si llora o sonríe. Así lo hacemos habitualmente, y tal apreciación puede ser acertada, pero también equivocada, si nuestra observación es superficial, o nuestro interlocutor un actor consumado, de manera que confundimos un llanto o una risa fingidos con unos reales. Las entrevistas llamadas cognitivas son herramientas que buscan precisamente entender lo que pasa en la mente de una persona, sea en relación con las preguntas que se le presentan en un cuestionario, o en otros contextos.

Al inicio del libro *Actos de significado. Más allá de la revolución cognitiva,* Jerome Bruner dice que quiere contar lo que él y sus amigos pensaban a fines de la década de

1950 de esa revolución, de la que él mismo era un destacado actor. A su juicio se trataba de un esfuerzo:

> [...] por instaurar el significado como el concepto fundamental de la psicología; no los estímulos y las respuestas, ni la conducta abiertamente observable, ni los impulsos y su transformación, sino el significado. No era una revolución contra el conductismo [...] era más profunda que todo eso [...] Su meta era descubrir y describir formalmente los significados que los seres humanos creaban a partir de sus encuentros con el mundo, para luego proponer hipótesis acerca de los procesos de construcción de significado en que se basaban. Se centraba en las actividades simbólicas empleadas por los seres humanos para construir y dar sentido no solo al mundo, sino también a ellos mismos. Su meta era instar a la psicología a unir fuerzas con sus disciplinas hermanas de las humanidades y las ciencias sociales, de carácter interpretativo [...] (Bruner, 2006: 22-23)

Snow y Lohman muestran las implicaciones de la revolución cognitiva para la medición educativa. La expresión *ciencia cognitiva* designa la confluencia de psicología, lingüística, neurofisiología y ciencias computacionales, y el elemento común a las diversas corrientes psicológicas que comprende, es

> [...] la visión de que los temas centrales de indagación para la ciencia psicológica son los procesos cognitivos y los contenidos involucrados en la atención, la percepción y la memoria, en el pensamiento, el razonamiento y la solución de problemas, así como en la adquisición, organización y uso del conocimiento. (Snow y Lohman, 1989: 264)

Los campos que estudia la psicología cognitiva pueden combinarse con contenidos relativos a las aptitudes generales y especiales, el desempeño en lectura, matemáticas o ciencias naturales y otros. Los estudios basados en ella pueden referirse a procesos mentales simples o complejos, y pueden emplear herramientas basadas en registros de conductas derivados de reportes verbales, entrevistas retrospectivas, secuencias de las fijaciones oculares de los sujetos al realizar una tarea, e datos neurofisiológicos obtenidos con electroencefalogramas. (Snow y Lohman, 1989: 272)

En otro texto sobre la relación de la psicología cognitiva con la evaluación, Mislevy señala que los estudios de este enfoque comparten dos premisas complementarias: la de que todas las personas son similares en lo que se refiere a mecanismos y procesos cognitivos, *con notables capacidades y sorprendentes limitaciones,* y que el contenido

del aprendizaje y el pensamiento es moldeado por la cultura. El texto se refiere a las implicaciones de la perspectiva cognitiva para la evaluación,

> [...] con énfasis en ámbitos ricos desde el punto de vista semántico. ¿Qué nos dice la investigación sobre cómo piensan, aprenden y actúan las personas en lo que se refiere a la forma de construir y usar una evaluación? (Mislevy, 2006: 257-258)

De la literatura sobre entrevistas cognitivas pueden mencionarse las obras de Leighton y Gierl (2007) y Leighton (2017). La perspectiva cognitiva se aplica desde la década de 1980 al diseño y mejora de cuestionarios, con la metodología llamada *Cognitive Aspects of Survey Methodology* (CASM, según Willis, 2005: 34-35).

Al responder un cuestionario se identifican cuatro procesos: comprensión de la pregunta; recuperación de información relevante en la memoria; procesos de estimación y juicio; y procesos de formulación de la respuesta. Para estudiar estos procesos se manejan técnicas particulares: la entrevista de pensamiento en voz alta (*thinking aloud*) y la de exploración verbal (*verbal probing*) (Willis, 2005: 36, 42). Se distinguen dos tipos de entrevistas de pensamiento en voz alta:

- Análisis de protocolos: el sujeto verbaliza lo que piensa *al momento* en que responde una pregunta, sin que el entrevistador intervenga, salvo para que no se deje de verbalizar; explora procesos que usan memoria de corto plazo.
- Análisis verbal: entrevista hecha *después* de que el sujeto realiza una tarea (de pocas horas hasta unos días); el entrevistador interroga al sujeto sobre sus creencias y actitudes respecto a la tarea, explorando procesos que usan elementos de la memoria de largo plazo. (Leighton, 2009)

Las ventajas de las entrevistas de pensamiento en voz alta, o análisis de protocolos, incluyen que el entrevistador no necesita entrenamiento y, como su intervención es mínima, no hay riesgo de que sesgue los resultados. Como desventajas, los sujetos sí necesitan entrenamiento, sobre todo si son niños, ya que verbalizar lo que se piensa no es usual; la técnica implica esfuerzo considerable para los entrevistados, que aún después de cierto entrenamiento pueden no verbalizar suficientemente, distraerse de la tarea e incluso olvidar responder lo que se les pregunta. Los sujetos pueden ser propensos a dar respuestas socialmente aceptables, congruentes con otras (efecto halo), afirmativas o negativas. Las ventajas de las entrevistas de exploración o análisis verbal son que el entrevistador

controla el proceso y puede mantener el foco, y que el sujeto no necesita entrenamiento; la desventaja es que el entrevistador necesita entrenamiento cuidadoso, para que pueda evitar el riesgo de sesgo que este tipo de entrevista encierra. (Willis, 2005: 53-57)

La tabla siguiente presenta ejemplos de preguntas de exploración verbal.

TABLA 3.1. EJEMPLOS DE PREGUNTAS DE EXPLORACIÓN COGNITIVA

Tipos	Ejemplos
De comprensión o interpretación	¿Qué significa para usted la expresión paciente externo?
De parafraseo	¿Puede repetir con sus palabras la pregunta que le acabo de hacer?
De confianza en el juicio	¿Qué tan seguro está de que su seguro de gastos médicos cubre tratamientos por alcoholismo o drogadicción?
De memoria	¿Cómo recuerda que ha ido al doctor 5 veces en los últimos 12 meses?
Específica	¿Por qué cree usted que el problema de salud más serio es el cáncer?
Generales	¿Cómo llegó a esa respuesta? ¿Qué tan fácil o difícil fue esa respuesta? Noté que dudó al responder. Dígame lo que pensaba.

FUENTE: WILLIS, 2005: 48. TABLA 4.1.

En síntesis

No sólo en el caso de los cuestionarios, sino también en el de los demás acercamientos a la obtención de información a partir de la que dan los sujetos estudiados, se aplican los principios de que sólo se deben preguntar cosas que los informantes conozcan o de las que puedan tener conciencia, además de estar dispuestos a dar la información, y de que se cumpla el prerrequisito de que entiendan en forma inequívoca lo que se les pregunte.

Las preguntas cerradas cortas, sin contexto, son fáciles de contestar, pero hay riesgo de que se entiendan de manera diferente; las preguntas contextualizadas más largas —por ejemplo, con viñetas— propician una comprensión uniforme, pero su elaboración lleva tiempo, y requieren más esfuerzo para responderlas. Si se prefiere usar respuestas abiertas el tiempo de elaboración se reduce, pero aumenta dificultad de categorización posterior, que exige personal entrenado.

Es ineludible la necesidad de definir categorías para analizar la información, sea que se haga a *priori o a posteriori.* La disposición favorable a responder se podrá atender si se asegura la confianza de los sujetos y se garantiza el anonimato.

Con tales cuidados se podrá obtener información de calidad aceptable y a bajo costo con instrumentos basados en información proporcionada por los sujetos, pero no es posible eliminar del todo el riesgo de distorsión en las respuestas. Por ello los instrumentos del primer grupo no son la mejor opción para estudiar fenómenos cuya complejidad y sensibilidad hacen que esos riesgos sean fuertes, y lleva a recurrir a acercamientos de los otros dos grupos que se presentan en los incisos 2 y 3.

Las distintas variantes de entrevista cognitiva pueden ayudar a mejorar preguntas de instrumentos de este grupo, y en principio también pueden usarse para obtener con ellas la información que se requiera para un proyecto de investigación, advirtiendo desde luego que su uso implica mucho tiempo y personal muy bien preparado, como ocurre con todas las entrevistas no estructuradas.

Acercamientos basados en observación

El segundo grupo de acercamientos para recabar información no se basa en lo que digan los informantes al ser interrogados de alguna forma, sino en la observación por terceras personas, dando en este caso al término observación el sentido que lo restringe a la que se hace mediante los sentidos de la vista y el oído.

Un poco de historia

La observación de las prácticas docentes

Desde el siglo XIX en algunos sistemas educativos se pusieron en marcha sistemas de inspección, para uniformar la enseñanza y asegurar mínimos de calidad. Una tarea habitual de inspectores o supervisores era asesorara y evaluar a los maestros, para lo cual los entrevistaban, revisaban planes de trabajo y material preparado por ellos, cuadernos de alumnos, etc. Para observar cómo trabajaban los maestros y verificar el grado de avance de los alumnos, los supervisores visitaban los salones y hacían preguntas o aplicaban pruebas escritas que ellos mismos preparaban.

Algunos inspectores desarrollaron formatos con los que se podía ubicar físicamente a los alumnos, como ayudas para observar lo que pasaba en el aula (*seating charts*). En 1914 Horn propuso usar símbolos para indicar si un alumno respondía una pregunta o realizaba cierta actividad. Las primeras formas para la observación sistemática de lo que pasaba en las aulas se basan en esas prácticas (*cfr*. Medley y Mitzel, 1963).

Una nueva etapa en el uso de técnicas de observación para estudiar prácticas docentes tuvo lugar en la época en que las ideas conductistas predominaban en psicología, en especial en las décadas de 1950 y 1960. A partir de la premisa de que no es posible

estudiar pensamientos o sentimientos íntimos de las personas, sino solo su manifestación externa en conductas visibles, esta etapa vio el desarrollo de protocolos para observar conductas o acciones particulares, previamente definidas y caracterizadas en términos operacionales, de modo que un observador pudiera detectar con facilidad su presencia o ausencia, o la frecuencia de su ocurrencia en cierto período de tiempo.

La definición de las técnicas de observación de la primera edición del *Handbook of Research on Teaching* refleja estas concepciones basadas en el conductismo, al incluir solo a las que consisten en el registro puntual de conductas particulares, con un mínimo de inferencia, y excluir explícitamente a cualquier otro acercamiento:

> Una técnica observacional [...] es una en la que un observador registra aspectos relevantes de una conducta que tiene lugar en el aula a medida que ocurre o muy poco después (or within a negligible time limit after) con un mínimo de cuantificación entre la observación de la conducta y su registro [...] por consiguiente no se incluyen en esta definición esquemas en los que se pide al observador que califique al maestro, a los alumnos o a la clase en una o más dimensiones, aún si las calificaciones se basan en la observación directa de conductas específicas. (Medley y Mitzel, 1963: 253)

Esos autores decían que "la investigación sobre las conductas que tienen lugar en el aula no es pasatiempo para aficionados, sino ocupación de tiempo completo para profesionales técnicamente competentes". (Medley y Mitzel, 1963: 253)

Medley y Mitzel (1963) distinguían sistemas de observación basados en tiempos (*de categorías*) o en eventos (de *signos*), y describen instrumentos para observar prácticas de enseñanza usados por supervisores escolares en las décadas de 1920-1930, e instrumentos de 1950 y 1960 basados en concepciones conductistas.

Rosenshine y Furst (1973) distinguían técnicas de observación por el procedimiento de registro: sistemas de conteo (categorías o signos) y de calificación (*rating*); por la especificidad de los ítems: sobre conductas específicas (de baja inferencia) o más generales (de alta inferencia); y por tipo de codificación, de una o más dimensiones.

En una perspectiva más amplia, la definición de observación incluye procedimientos con escalas de calificación, técnicas de registro y post-codificación, y técnicas cualitativas. Remmers (1963) describe otro tipo de técnicas de observación, que la definición de Medely y Mitzel descarta: escalas y técnicas que implican una calificación (*rating*) de las conductas.

Rosenshine y Furst (1973) informan sobre el desarrollo de acercamientos de observación en otra década. Con la revolución cognitiva los estudios de inspiración

conductista disminuyeron y se multiplicaron los de enfoque cualitativo, lo que se refleja en los trabajos de Erickson (1986) y Everston y Green (1986).

Las observaciones estructuradas basadas en el conductismo no daban cuenta de la complejidad de lo que ocurre en las aulas, y en particular de aspectos importantes. Por ello los estudios intensivos y en profundidad parecían preferibles, con la desventaja del excesivo costo que supondría su aplicación en gran escala, y no solo en pocos casos que típicamente se utilizan en esos trabajos. Los trabajos recientes reflejan el desarrollo de los enfoques cualitativos para estudiar la enseñanza; siguen desarrollándose diversos enfoques, unas veces oponiéndose como excluyentes, otras buscando complementarse. (Floden, 2001)

La grabación de imagen y/o sonido

Desde fines del siglo XX, y en particular como parte de la búsqueda de formas confiables de evaluar el desempeño de los maestros, surgieron esfuerzos por desarrollar protocolos de observación estructurada que informen no solo sobre conductas puntuales fáciles de identificar, sino sobre prácticas complejas e interacciones en el aula con precisión y confiabilidad, y con costos que hagan posibles aplicaciones en una escala que permita generalizar los resultados. Estos esfuerzos se benefician por la posibilidad de apoyarse en grabaciones de imagen y sonido de lo que ocurre en las aulas, grabaciones que son calificadas después por observadores capacitados. Esto tiene desventajas, y requiere que la calidad de la imagen y del sonido sea alta, pero sus ventajas son también claras.

El uso de grabaciones se remonta a los inicios del desarrollo de disciplinas que estudian al hombre y la sociedad. Los sociólogos usaban cuestionarios porque muchos de los sujetos que estudiaban, en contextos como el de Chicago, sabían leer y escribir. Los etnólogos y antropólogos, que estudiaban grupos cuyas lenguas a veces ni siquiera conocían la escritura, recurrían a observación intensiva por largos períodos de tiempo.

Inicialmente la observación no se apoyaba en tecnologías para el registro dinámico de imágenes y sonido, que aún no se habían desarrollado, pero una vez que surgieron, el uso de esas tecnologías como herramientas para el trabajo de los antropólogos comenzó pronto. A fines de la década de 1930, Margaret Mead y Gregory Bateson tomaron más de 25,000 fotografías y cientos de horas de película en su trabajo antropológico en Bali, utilizando por primera vez cámaras de 16mm. (NRC, 2001: 5). Años antes Franz Boas había filmado danzas rituales de los indígenas de la costa del Pacífico, con cámaras más pesadas y sin sonido. Erickson (2011) describe cómo el avance tecnológico permitió

estudiar fenómenos sociales con las primeras grabadoras de voz de cinta magnética, en 1949, y luego con videograbaciones en la década de 1960.

La mayor portabilidad y calidad de la grabación por medios electrónicos, y luego digitales, hizo que en la segunda mitad del siglo XX el uso de las videograbaciones aumentara rápidamente. Rosenstein presenta una clasificación de una gama de usos actuales del video en estudios de antropología, etología, lingüística, medicina, psicología, sociología, urbanismo y educación, en diversos niveles y áreas del currículo, como matemáticas y ciencias, o para formación de docentes (2002: 29-30). Jewitt (2012) distingue uso en programas de intervención de videograbaciones hechas previamente, para estimular el recuerdo o como base para la reflexión (*video elicitation*), y usos para "trabajo de campo basado en videos" (*video-based field work*), que *supone la recolección de datos ocurridos naturalmente utilizando cámaras de video, tal vez el uso más establecido de videos para la recolección de datos en las ciencias sociales.*

En los medios educativos el uso de grabaciones de imagen y sonido tiene décadas de existencia, con propósitos de formación docente. Hacia 1960 se popularizó la *microenseñanza*, basada en la idea de que la dificultad de observarse a sí mismo impide que el maestro sea consciente de su práctica y pueda mejorarla. El principio de la microenseñanza era que la grabación de clases, o partes de clase, sería una herramienta de mejora, al facilitar la toma de conciencia de los maestros grabados en cuanto a puntos débiles y fuertes de su enseñanza (Sherin, 2004).

Al aumentar la calidad de las videograbaciones, con la reducción del tamaño y el costo de los equipos, su uso para formación docente aumentó también, pero su uso para investigación se extendió más lentamente, y principalmente para apoyar trabajos de tipo intensivo o cualitativo.

Algunos sistemas de observación

Este inciso describe brevemente algunos ejemplos de sistemas de observación, de diferente antigüedad y complejidad, que se han usado para registrar lo que ocurre en un salón de clases. En el Apéndice del Capítulo se presentan con más amplitud.

La estructura que sirvió para organizar estos sistemas se toma de Stallings (1977), que también es autora de unos de los que se describen. Según Stallings, más allá de su variedad, los sistemas de observación tienen elementos en común:

- Persona foco de la observación: A quién miras o escuchas: el maestro, un niño, varios... Qué actividades, materiales o factores ambientales registras...

- Contenido focal: Sobre qué quieres saber algo: Desarrollo motor, cognitivo o socioemocional, ambiente físico, actividades...
- Unidad de codificación. Cuánto tiempo observar antes de registrar algo y durante cuánto tiempo observar: tres segundos, cinco minutos, cinco horas.
- Medio para registrar: Cómo registras los datos: con grabación de audio o de video, con lápiz y papel.
- Entorno: Dónde registras: en el aula, en el patio, en una cancha de juegos...
- Propósito: Para qué observas: para estudiar a un alumno, evaluar un programa, entrenar a otras personas, desarrollar un proyecto de investigación. (1977: 6)

Se distinguen descripciones narrativas; instrumentos simples para observaciones estáticas; sistemas para captar la dinámica del aula; y sistemas para registrar interacciones.

Descripciones narrativas

Representan la forma menos estructurada de recabar y registrar información sobre lo que pasa en un aula. La investigadora (en el ejemplo de Stallings 1977 ella misma, en su papel de educadora) se limita a observar lo que ocurre, tanto lo que hacía ella como lo que hacían los niños, a partir de preguntas poco estructuradas, de carácter bastante general, y registrándolo en seguida. No tiene que ser la educadora misma quien haga la observación, sino que puede ser otra persona.

Instrumentos simples para observaciones estáticas

Lista de cotejo de Stallings. Como alternativa más estructurada que la descripción narrativa anterior, para recabar información sobre un tipo particular de acciones Stallings propone una lista de cotejo "Sobre conductas para evitar actividades de aprendizaje" (*Observed Learning Avoidance Behaviors*), con un formato que permite registrar lo que hace un niño en un lapso de cinco minutos y tres días diferentes, en cuanto a 19 conductas. (Stallings, 1977: 11-13)

Mapas de asientos (*Seating charts*). Estas herramientas, las primeras para observar lo que pasa en las aulas, fueron desarrolladas por supervisores escolares, pues sus funciones profesionales incluían observar clases para evaluar al docente. Consisten en un cuadro con un número de casillas igual o mayor al de los alumnos del grupo a observar. Recuérdese que, en las escuelas de principios del siglo XX, los lugares eran fijos y los niños ocupaban regularmente el mismo sitio; los asientos estaban incluso sujetos al piso. Con símbolos predefinidos, el observador podía registrar las acciones de cada alumno

de un grupo, aunque no conociera sus nombres, con base en el lugar que ocupaba en el aula, que los números de las casillas del *Mapa de asientos* permitían identificar. Con ayuda del docente, el observador podía anotar en cada casilla del Mapa el número de lista del alumno que lo ocupaba, para poder identificarlo. Con los símbolos el observador podía registrar diversas conductas. En el Apéndice se presentan dos ejemplos de Mapa de Asientos: uno propuesto en 1928 por R. C. Puckett, en un artículo titulado *Haciendo objetiva la supervisión*, y otro de 1934, debido a J Wrightstone. (*cfr*. Medley y Mitzel, 1963: 254)

Sistemas para captar dinámica del aula con muestreo de tiempo

Los sistemas de observación anteriores tienen en común su carácter estático, ya que sólo captan lo que ocurre en el aula en un momento dado, lo que hace que a veces se utilice para designarlos el término de "instantánea" (*snapshot*), en el sentido que se utiliza en fotografía. A diferencia de una película, una fotografía instantánea congela lo que ocurría en el momento en que se tomó.

La forma de utilizar la *Lista de cotejo sobre conductas para evitar actividades de aprendizaje* de Stallings, con registros en días diferentes, es un intento por superar esa limitación, con la lógica que llevó al desarrollo de las películas animadas: con instantáneas separadas por intervalos pequeños es posible capturar movimientos.

Esto es lo que se hace en los sistemas de observación con muestreo de tiempo (*time sample*), que tienen en común el que se registren observaciones sucesivas sobre una misma situación, buscando captar secuencias en una forma dinámica. Se presentan cinco sistemas:

Mapa del Tiempo (Time Chart) de A. S. Barr (1929). Para identificar características que distinguen prácticas docentes buenas y deficientes, se registra la duración de ciertas conductas de maestro y alumnos, durante 30 minutos de una clase. Cada renglón representa un minuto; las columnas distinguen períodos de 10 segundos. (Medley y Mitzel, 1963: 258-259)

Sistema multidimensional de observación (Cornell, Lindvall y Saupe, 1952). El sistema considera ocho dimensiones, una de las cuales (*Variedad*) comprende 23 tipos de conductas de maestro o alumnos. Los renglones de la forma se refieren a dichas dimensiones, y las columnas al momento en que ocurren las conductas, considerando períodos de cinco minutos en una hora de observación. (Medley y Mitzel, 1963: 275-276)

Sistema para observar interacción en matemáticas (Wright y Proctor, 1961). Este sistema de observación con muestreo de tiempo fue desarrollado para comparar las

interacciones verbales que tienen lugar en clases de matemáticas. La observación se hace durante 45 minutos, lapso en el que deben hacerse 90 observaciones por sesión, cada una de las cuales lleva medio minuto, dividido en dos partes: 15 segundos para observar y otros 15 para registrar lo observado. Para controlar el tiempo se usa un cronómetro. Las interacciones observadas se clasifican según tres dimensiones: contenido, proceso y actitud. (Medley y Mitzel, 1963: 288-290)

Sistema de Elizabeth Prescott (1973). Instrumento para observar a niños de tres y cuatro años de edad en guarderías. Las conductas de un niño se registran cada 15 segundos, indagando autonomía, dependencia, agresión, participación social, persistencia en una tarea, habilidad para resolver problemas y curiosidad. Este sistema de observación aporta información sobre el crecimiento y desarrollo de los niños que están en guarderías, incluyendo conductas que no pueden ser medidas con pruebas estandarizadas. (Stallings, 1977: 13)

Sistema sobre necesidades de educación especial (Croll y Moses, 1985). Sistema desarrollado para estudiar las diferencias que pueda haber entre las actividades que llevan a cabo y las interacciones en que participan en el aula los niños con necesidades educativas especiales o sin ellas. En cada aula se observan de cuatro a diez alumnos, sin que el maestro sepa quiénes son; en total dos horas por niño y 20 horas por aula. (Croll y Moses, 1985)

Sistemas para el registro de interacciones

Un sistema de interacción registra *lo que las personas dicen o hacen entrando en relación unas con otras;* la atención debe centrarse en una persona, de la que el observador registra todo lo que dice o hace en un lapso de tiempo, *que suele ser de cinco a diez minutos* (Stallings, 1977: 13). Estos sistemas suelen usar elementos de los ya descritos, como listas de cotejo, esquemas de asientos y muestreo de tiempo.

Sistema de Flanders (Flanders Interaction Analysis Category System, FIACS). Cada tres segundos el observador anota un número que corresponde a un tipo de conducta de profesor o alumnos. Las anotaciones se hacen siguiendo las columnas del formato, cada una de las cuales tiene 20 renglones, que en total corresponden a un minuto (tres segundos cada renglón). El formato tiene espacio para media hora de observación (30 columnas). Para observar clases más largas se usan formatos adicionales. Los períodos dedicados a una misma actividad se llaman episodios, como el lapso en que el profesor pasa lista, para luego iniciar una exposición, o un período dedicado a dirigir preguntas a los alumnos o a responder dudas. Flanders recomienda observar a un profesor al menos

durante seis clases, y de preferencia durante ocho. Es un sistema que se utilizó mucho en las décadas de 1960 y 1970, y todavía ahora se le usa ocasionalmente; ha inspirado otros, como el Snapshot de Stallings. (Medley y Mitzel, 1963: 271-273)

Sistema OScAR (Observation Schedule and Record) Medley y Mitzel (1958). Este sistema es un esfuerzo por captar el mayor número posible de aspectos de lo que ocurre en un salón de clases. El formato en que se registran las observaciones es una hoja impresa por los dos lados, en los que hay una serie de bloques o secciones para registrar varios elementos: en el anverso actividades, agrupamientos, señales y materiales; en el reverso de la hoja la relación del profesor con los alumnos (verbal y gestual), y las materias enseñadas. La observación se realiza durante unos 40 minutos, distinguiendo períodos pares e impares de cinco minutos, en los que la atención se debe centrar en ciertos aspectos particulares.

La complejidad de este protocolo es evidente. El sistema fue diseñado buscando:

> [...] permitir registrar tantos aspectos significativos como fuera posible de lo que ocurre en el aula, sin buscar relacionarlos con algunas dimensiones o escala. La única preocupación del observador era ver y oír todo lo que pudiera de lo que estaba ocurriendo, y registrar todo lo que pudiera sin hacer supuesto alguno sobre su importancia relativa, o su relevancia respecto a cualquier dimensión conocida. (Medley y Mitzel, 1963: 280-281)

En cuanto a la calidad de la información:

- La confiabilidad: estimada con un análisis de varianza, mostró ser mejor si la observación es hecha por dos personas, aunque esto es costoso.
- La validez de las categorías utilizadas: con un análisis factorial, se agruparon en tres dimensiones: *Clima emocional, Énfasis verbal* y *Estructura social.*

Esto muestra la sofisticación metodológica a la que se llegaba a fines de la década de 1950, pero al mismo tiempo permite ver los límites de los acercamientos desarrollados en el marco del paradigma conductista, Medley y Mitzel reconocen honestamente hasta dónde podían llegar sus resultados:

> Un defecto principal del sistema OScAR es que no consigue captar ningún aspecto de la conducta del aula que esté relacionado con el grado en que los alumnos logran los objetivos cognitivos. Las tres dimensiones que mide representan lo que son probablemente las diferen-

> cias más obvias entre las clases: qué tan ordenadas y relajadas son, en qué formas se agrupan los alumnos, y el contenido general de las lecciones que se enseñan. Medir estos aspectos fue relativamente fácil; medir diferencias más sutiles y cruciales con OScAR será probablemente más difícil. Sin embargo, no hay razón para pensar que sea imposible. (1963: 286)

Sistema de instantáneas (Snapshot) de Jeane Stallings (1977). Este es un ejemplo más de los alcances y límites de los sistemas de este grupo, y reviste interés porque comenzó a usarse hace poco en un proyecto del Banco Mundial para América Latina, y en 2017 era utilizado en el sistema educativo de la Ciudad de México. El *Snapshot* fue desarrollado por Stallings y colaboradores en el Stanford Research Institute, para evaluar programas educativos que participaban en un proyecto del Congreso de Estados Unidos (*Follow Through Planned Variation Project 1967*), buscando identificar programas que pudieran reforzar y extender los avances académicos que habían conseguido hacer niños de medio económico desfavorable en el programa *Head Start* y otros dirigidos a alumnos de preescolar.

Para ver cuáles programas favorecían el desarrollo de los niños, se escogieron 22 que representaban toda la gama de teorías educativas presentes en el proyecto *Follow Through:* modelos de modificación conductual basados en la teoría de Skinner; uno basado en la teoría de Piaget; un modelo de escuela abierta basado en la teoría inglesa de escuela para infantes; y modelos basados en varias combinaciones de las teorías de Piaget, Dewey, Rogers y la escuela inglesa. En la evaluación se pretendía observar si se respetaban los lineamientos de cada programa en cuanto al tipo de materiales a utilizar y a las formas de agrupar a los niños para trabajar con el maestro o con auxiliares, así como en cuanto al tipo de interacciones verbales de maestros y alumnos.

Después de analizar los sistemas de observación disponibles, los autores juzgaron que ninguno de los entonces existentes era suficientemente amplio y flexible para observar una variada gama de programas, y decidieron desarrollar un instrumento nuevo, con apoyo de representantes de ocho de los promotores de los programas en cuestión. El sistema se aplicó por primera vez en 1969 y se aplicó en cuatro años escolares, de 1970 a 1973, ajustándose cada vez. En su versión final, el sistema comprende tres instrumentos: Formato sobre el entorno físico del aula (*Physical Environment Information,* PEI); Lista de cotejo (*Classroom Check List,* CCL); Formato para registrar interacciones cada 5' (*Five-Minute Interaction,* FMI), que se llena cuatro veces por hora, después de la CCL. (Stallings, 1977)

La complejidad del Sistema y su afinidad con los anteriores es evidente, al igual que la dificultad que supone aplicarlo. La versión para América Latina es simplificada,

conservando solo parte de la original, lo que facilita su aplicación, pero reduce su alcance. En Apéndice puede verse la versión usada en la Ciudad de México. La reflexión obligada es que, si la información que da un instrumento como el OScAR es pobre, la de una versión simplificada del Snapshot está lejos de responder a lo que suelen esperar sus patrocinadores y los sistemas educativos que la usen.

Protocolos recientes

Estos sistemas buscan no limitarse a registrar conductas simples, como los que se basan en el conductismo, pero tampoco cubrir un número reducido de casos, como los acercamientos intensivos. Surgidos en el marco de esfuerzos por evaluar a los maestros superando los sistemas escalafonarios o basados en juicio de directores y supervisores, estos protocolos buscan estudiar prácticas docentes complejas con precisión, confiabilidad y costos que hagan posible su aplicación en gran escala.

Se han desarrollado muchos instrumentos, pero no abunda información que permita valorarlos. Los que se presentan tienen en común que fueron usados en el proyecto *Measures of Effective Teaching* (MET), desarrollado entre 2009 y 2012 con apoyo de la Fundación *Bill & Melinda Gates,* para recoger información sobre las prácticas videograbadas en aula de unos 3,000 maestros y 100,000 estudiantes, permitiendo análisis sin precedentes sobre la calidad de la información obtenida. (Kane, Kerr y Pianta, 2014). Se presentan dos sistemas de propósito general (FFT y CLASS), y tres orientados en particular a un área curricular: PLATO, MQI y QST.

Framework for Teaching (FFT). El *Marco para la Enseñanza* se deriva del trabajo de Charlotte Danielson y otros para el instrumento *Praxis* III del ETS. Sus autores describen como una herramienta:

> [...] estructurada según componentes de la instrucción derivados de la investigación, alineada a los estándares del consorcio interestatal (INTASC), y basada en una perspectiva constructivista del aprendizaje y la enseñanza [...]. (Goe, Bell y Little, 2008: 21-22)

El FFT tiene cuatro dominios: Planeación y preparación de la clase; Ambiente del aula; Enseñanza; y Responsabilidades profesionales. Estos dominios se dividen en 22 componentes. Los dominios y sus componentes son los siguientes:

TABLA 3.2. DOMINIOS Y COMPONENTES DE LA PRÁCTICA PROFESIONAL EN EL FFT

1. Planeación y preparación de la clase	2. Enseñanza
• Demostrar conocimiento del contenido y la pedagogía. • Demostrar conocimiento de alumnos. • Selección de metas de instrucción. • Demostrar conocimiento de recursos. • Diseñar una enseñanza coherente. • Evaluar aprendizaje de los alumnos.	• Comunicarse clara y cuidadosamente. • Usar técnicas de interrogación/discusión. • Involucrar a estudiantes en aprendizaje. • Ofrecer retroalimentación a los alumnos. • Demostrar flexibilidad y capacidad de respuesta.
3. Ambiente del aula	**4. Responsabilidades profesionales**
• Crear ambiente de respeto y relación. • Establecer una cultura de aprendizaje. • Manejar procedimientos para la gestión del aula. • Gestionar la conducta de los alumnos. • Organizar el espacio físico.	• Reflexionar sobre la enseñanza. • Mantener registros cuidadosos. • Comunicarse con las familias. • Colaborar con la escuela y el distrito. • Crecer y desarrollarse profesionalmente. • Mostrar profesionalismo.

FUENTE: DANIELSON, 1996: 61.

Los 22 componentes se precisan con 76 indicadores, y el *Marco* incluye rúbricas detalladas para evaluar a un maestro en cuanto a cada uno de los 76 indicadores en cuatro niveles de desempeño, que se definen como *insatisfactorio, básico, avanzado (proficient) y sobresaliente (distinguished)*. Danielson promovió el uso de su sistema de observación, permitiendo que se usara en gran número de escuelas y distritos escolares y aceptando que se le hicieran modificaciones para adecuarlo a las circunstancias locales, por lo que seguramente es el sistema más extendido, pero no había mucha información sobre la calidad de la información obtenida, lo que el Proyecto MET subsanó. (Danielson, 1996)

Classroom Assessment Scoring System (*CLASS*). Este instrumento busca estudiar las actividades del docente y las interacciones que tiene con sus alumnos. Fruto del trabajo de Robert Pianta y sus colegas, una versión previa se denominó *Classroom Observation System* (Pianta y Hamre, 2009).

El sistema se desarrolló inicialmente para observar grupos de preescolar o los grados inferiores de primaria; posteriormente se hicieron versiones para grados superiores de primaria y para secundaria. (Goe, Bell y Little, 2008: 24)

A diferencia del FFT, el uso del CLASS se promueve de manera controlada, requiriéndose a quienes quieran aplicarlo que se capaciten y certifiquen para hacerlo. Por ello es probablemente el sistema sobre el que hay más trabajos de investigación, antes y después del Proyecto MET.

La conceptualización de las actividades e interacciones en que se basa el análisis del constructo *calidad de la dinámica del aula* distingue tres dominios, que sintetiza el cuadro siguiente, incluyendo los aspectos que comprende cada dominio.

TABLA 3.3. DOMINIOS Y ASPECTOS DE LA CALIDAD DE LA DINÁMICA DEL AULA DEL CLASS

	Apoyo emocional	Organización del aula	Apoyo pedagógico
Preescolar y primeros grados de primaria	Clima positivo Clima negativo Sensibilidad del maestro/a Consideración hacia perspectivas de los alumnos/as	Gestión de conducta Productividad Formatos didácticos para el aprendizaje	Desarrollo de conceptos Calidad de comentarios Ejemplificar el lenguaje

FUENTE: PIANTA, LAPARO Y HAMRE (2012).

La versión para los últimos grados de primaria y para secundaria tiene los mismos dominios y aspectos similares, con adecuaciones para tener en cuenta que se trata de estudiantes de mayor edad. En particular, en el dominio Apoyo Pedagógico se incluye la comprensión de contenidos, el análisis y solución de problemas y el diálogo instruccional, además de mantener la calidad de los comentarios.

La observación se lleva a cabo en períodos de media hora; en cada período se dedican 20 minutos a observar y tomar notas, y el tiempo restante a la calificación. Según los autores bastan cuatro ciclos de observación para contar con una muestra representativa de lo que ocurre en un salón de clases.

Protocol for Language Arts Teaching Observation (PLATO). Como su nombre indica, el "Protocolo para la Observación de la Enseñanza de las Artes del Lenguaje" es una herramienta orientada en especial al área curricular de lengua. (Grossman, Loeb, Cohen *et al.*, 2010)

Con base en investigaciones previas, el sistema PLATO se estructura a partir de cuatro factores que se considera subyacen la enseñanza: la demanda que el área plantea a las prácticas del aula y el discurso relativo; el andamiaje para apoyar la enseñanza de los contenidos de lengua; las representaciones de los contenidos y el uso que se hace de ellos; y el ambiente del aula. (MET Project, 2010a)

El sistema identifica 13 elementos que constituyen dimensiones independientes, y permite evaluarlos con una rúbrica, en una escala de uno a cuatro:

Propósito
Desafío intelectual
Representación del contenido
Conexiones con el conocimiento previo
Conexiones con la experiencia personal
Modelamiento
Estrategias explícitas de enseñanza
Práctica guiada
Discurso que se tiene en el aula
Instrucción basada en textos
Lenguaje académico
Gestión de las conductas
Manejo del tiempo

El sistema prevé observaciones independientes de 15 minutos cada una, durante una clase: dos en clases de 45 minutos y tres en las de 90. La investigación sobre la herramienta es más reducida que en los casos anteriores, y se ha desarrollado principalmente en escuelas de Nueva York. (MET Project, 2010a y 2010b)

Mathematical Quality of Instruction (*MQI*). Protocolo para observar prácticas de enseñanza en el campo de las matemáticas (Hill *et al.* 2010; *Learning Mathematics for Teaching Project*, 2011). Se identificaron cinco puntos, como los más relevantes:

- Uso preciso y rico del lenguaje matemático;
- Ausencia de errores e imprecisiones;
- Presencia de explicaciones matemáticas correctas y participación de los alumnos en la construcción de significado y en el razonamiento matemático;
- Conexión del trabajo del aula con ideas importantes en el campo; y
- Esfuerzos para garantizar que todos los alumnos del grupo, y no sólo algunos, tengan acceso al conocimiento matemático. (MET Project, 2010c)

El MQI explora tres tipos de relaciones: de los maestros con los contenidos; de los alumnos con los contenidos; y de los maestros con los alumnos. Al valorar esas dimensiones, el MQI ofrece una visión completa y equilibrada de los elementos que, en conjunto, conforman la enseñanza de calidad en el área. (MET Project, 2010c)

Durante el desarrollo del MQI se encontró que, además de conocimiento general de matemáticas, el maestro debe dominar un conocimiento matemático particular, *el necesario para la enseñanza de esta área*, que tiene que ver con una comprensión precisa de los obstáculos que dificultan a los alumnos el aprendizaje. Por ello en paralelo se ha desarrollado otra herramienta para medir ese tipo de conocimiento: el MKT (*Mathematical Knowledge for Teaching*). (Hill *et al.*, 2008 y 2004)

Quality of Science Teaching (QST). Protocolo diseñado para estudiar la enseñanza de ciencias, teniendo en cuenta los nuevos referentes establecidos para esa área, los llamados *Next Generation Science Standards* (NGSS). En su versión original, el QST comprendía seis dominios y 18 indicadores de calidad de la enseñanza:

1. Evidencia de conocimientos del contenido a enseñar y pedagogía.
2. Acciones para involucrar (engaging) en aprendizaje a los alumnos.
3. Acciones para facilitar razonamiento y discurso científico.
4. Promover indagación basada en trabajo de laboratorio.
5. Ofrecer oportunidades para aplicar la ciencia en la vida.
6. Monitorear aprendizaje de los estudiantes. (Schultz y Pecheone, 2014: 444)

Al no haber variación en los dominios 3 y 5, en que sólo se encontraron prácticas muy pobres, en el proyecto MET se eliminaron los indicadores respectivos, para reducir la dificultad del trabajo de codificación. Con el mismo fin, los indicadores del dominio 6) se integraron con los otros, de manera que la versión del QST usada en el proyecto MET (QST-MET) comprendió solo tres dominios, con cuatro indicadores cada uno:

TABLA 3.4. DIMENSIONES E INDICADORES DEL QST-MET

Dominios (clusters)	Indicadores. El maestro…
Evaluar los conocimientos y la pedagogía del maestro	Define el contexto y focaliza aprendizaje en conceptos científicos clave
	Utiliza representaciones
	Demuestra que tiene conocimiento de los contenidos
	Ofrece retroalimentación para el aprendizaje
Hacer que los estudiantes participen activamente en el aprendizaje de ciencias	Promueve interés y motivación de los alumnos para aprender ciencias
	Asigna tareas para promover el aprendizaje y atiende las demandas que implican las tareas
	Utiliza formas varias para enseñar los conceptos científicos
	Genera evidencias del conocimiento y la comprensión conceptual de los alumnos
Promover la indagación basada en trabajo de laboratorio	Inicia investigaciones
	Ofrece orientación para conducir una investigación y recopilar datos
	Orienta el análisis y la interpretación de los datos
	Genera evidencias del conocimiento y la comprensión conceptual de los alumnos

FUENTE: SCHULTZ Y PECHEONE, 2014: 451. TABLA 14.1.

Incluso con la reducción de dominios, el QST-MET fue el protocolo de observación que arrojó resultados menos confiables, por lo que los primeros análisis del Proyecto MET no lo incluyeron. La menor confiabilidad, y sus implicaciones para la validez de los resultados, se explican en parte por la ausencia de un piloteo del QST-MET antes de su aplicación en las escuelas que participaron en el Proyecto MET. La conclusión es que el instrumento necesita correcciones y pasar por estudios adicionales, para que la calidad de la información que proporciona sea mejor.

Estudios con videograbaciones

En principio, los protocolos descritos en el apartado anterior pueden dar información de buena calidad sobre prácticas complejas, pero para alcanzar niveles aceptables de confiabilidad deben ser aplicados por personal calificado y observar un número mínimo de sesiones. Esto se facilita si se usan videograbaciones, lo que lleva a la necesidad de analizar las implicaciones de su uso, sus ventajas y desventajas.

El potencial del video para el estudio en gran escala de las prácticas de enseñanza se puso en evidencia con el TIMSS *Video Study* (TVS), asociado a la aplicación del *Third International Mathematics and Science Study* de 1995 (Stigler *et al.*, 1999; Stigler y Hiebert, 2009). Con la repetición del TIMSS en 1999 se realizaron nuevos estudios con videos de matemáticas y ciencias. El primer estudio incluyó aulas de Alemania, Estados Unidos y Japón; en el segundo participaron en ciencias Australia, Estados Unidos, Japón, Países Bajos y República Checa, y en matemáticas además Hong Kong y Suiza. (Roth *et al.*, 1999; Roth, 2009; Hiebert *et al.*, 2003)

Como el propósito de ambos estudios no era evaluar maestros individualmente, sino identificar características que distinguieran las formas de enseñar en diferentes culturas, en el primer caso se planeó grabar una sola clase de 100 profesores distintos. En el segundo estudio se planeó grabar una clase de matemáticas y/o una de ciencias en 100 escuelas de cada uno. Con ello fue posible detectar rasgos culturales interesantes (Stigler, Gallimore y Hiebert, 2000).

Varios trabajos sacan lecciones metodológicas del proyecto del TVS, como Siegel (2004) y Jacobs, Hollingsworth y Givvin (2007). Con base principalmente en el TVS, un texto del *National Research Council* incluye también consideraciones metodológicas en cuanto a los alcances y límites del video como herramienta de obtención de información; sobre la importancia de la información contextual; la necesidad de combinar análisis cualitativos y cuantitativos de la información registrada; el tamaño de la muestra y la necesidad de tener en cuenta la variabilidad al interior de un país;

cuestiones de confidencialidad y privacidad; el uso de los videos para el desarrollo profesional; y sobre la relación entre las prácticas de enseñanza y el aprendizaje de los alumnos. (NRC, 2001)

Además de los asociados al TIMSS, otros trabajos han usado videos con protocolos estructurados, aplicados en escala amplia. El Instituto Leibniz de Educación en Ciencias y Matemáticas (IPN por sus siglas en alemán) de la Universidad de Kiel, desarrolló el *IPN Video Study* para estudiar procesos de enseñanza y aprendizaje en física, con un diseño longitudinal que implicó videograbaciones repetidas a lo largo de un año, así como el uso de datos de otras fuentes, incluyendo pruebas y cuestionarios, reportes de los alumnos y juicios de los maestros, con una muestra de tamaño suficiente para aplicar modelos lineales jerárquicos. (Seidel *et al.*, 2004).

En Alemania y Suiza se realizó el *Estudio Pitágoras,* para comparar las prácticas de enseñanza de matemáticas en ambos países. Aprovechando la experiencia del TVS, se retomaron las técnicas de grabación y de codificación de inferencia baja y alta, con ajustes para controlar los contenidos a grabar, seleccionándose tres clases de introducción al Teorema de Pitágoras y tres en los que se resolvían problemas de álgebra presentados por los investigadores. (Klieme, Pauli y Reusser, 2004).

Entre 2004 y 2009, el Centro de Investigaciones Educativas (CPV por sus siglas en checo) de la Universidad Masaryk, en la República Checa, llevó a cabo cuatro estudios con videograbaciones sobre la enseñanza en inglés, educación física, geografía y física (Najvar *et al.*, 2009).

En México se han hecho al menos dos estudios en gran escala sobre las prácticas docentes de maestros de primaria que utilizaron videograbaciones.

- Como parte de la evaluación cualitativa del Programa Escuelas de Calidad (PEC), se videograbaron clases de español y matemáticas en 456 escuelas mexicanas en tres ocasiones —en los ciclos 2001-2002, 2002-2003 y 2003-2004— videos que se usaron tanto para retroalimentación de los docentes, como para un estudio en el que se analizó su práctica pedagógica y los eventuales cambios que sufrió a lo largo del lapso mencionado. (Loera Varela *et al.*, 2006; Loera Varela *et al.*, 2007)
- En el marco de un estudio del BID que incluyó a Paraguay y a la República Dominicana, se videograbaron clases de matemáticas y ciencias en 101 escuelas del estado de Nuevo León que participaron en el estudio SERCE del Laboratorio Latinoamericano de Evaluación de la Calidad Educativa. Con la metodología del TVS se analizó la práctica pedagógica y se comparó la de escuelas públicas

y privadas, la de escuelas cuyos alumnos tuvieron resultados altos, medios y bajos en las pruebas ENLACE, y se compararon sus prácticas con las de países participantes en el TVS. (Loera Varela, 2012; Loera Varela, A., Näslund-Hadley, E. y Alonzo, H., 2013)

El Proyecto MET, ya mencionado, buscó mejorar la evaluación de la eficacia de los docentes combinando medidas del aprendizaje de los alumnos con medidas de la práctica docente obtenidas con los instrumentos de observación descritos, y con encuestas aplicadas a alumnos y docentes. Se aprovecharon las ventajas de la videograbación, generando 7,500 videos de clases, que se analizaron aplicando los protocolos de observación mencionados (Kane, Kerr y Pianta, 2014).

Este estudio ofrece información metodológica y técnica importante para el estudio de las prácticas de enseñanza en gran escala. La obra de Kane, Kerr y Pianta (2014) incluye textos con aportaciones metodológicas derivadas del Proyecto MET:

- Gitomer, Phelps, Weren, Howell y Croft (2014) se refieren a la evidencia de validez de contenido que tienen las evaluaciones de maestros.
- Bell, Qi, Croft, Leusner, McCaffrey, Gitomer y Pianta (2014) discuten retos para la observación debidos al pensamiento de los observadores.
- Joe, McClellan y Holtzman (2014) muestran la influencia de la duración y el foco de las observaciones sobre la confiabilidad de la información recogida.
- Soo Park, Chen y Holtzman (2014) se refieren a los esfuerzos para minimizar el sesgo de los calificadores.
- Schultz y Pecheone (2014), como se ha apuntado, explican los limitados resultados de la aplicación del protocolo QST.

En síntesis

La idea de que la observación es el acercamiento ideal para el estudio de procesos como la práctica docente, y que el único obstáculo para emplearlo es su costo, debe matizarse advirtiendo que, aún con buenas videograbaciones, no es fácil observar confiablemente conductas complejas que suponen inferencias que deben hacer observadores cuyos registros no coinciden del todo, por mucha capacitación que reciban; además, el mejor observador no puede saber lo que los actores piensan cuando actúan de cierta forma, para lo que hay que interrogarlos.

Por otra parte, al estudiar prácticas docentes importa observar las acciones del maestro, pero también las de los alumnos, y las interacciones entre docente y alumnos y de estos entre sí. Por ello las grabaciones deben tener buena calidad de imagen y sonido. Si esto se asegura, la ventaja respecto a observación en vivo es que una misma grabación puede ser vista por distintos observadores, aplicando diferentes protocolos, y tantas veces como se quiera, sin interferencia o con una muy baja. El uso de video, sin embargo, implica fuentes de error propias, al ser imposible captar exhaustivamente lo que pasa en el aula, con cualquier combinación de fuentes de imagen y sonido, incluso con responsables de la grabación altamente calificados. La videograbación no disminuye la complejidad de la tarea de observar las prácticas docentes. Por ello, y tanto para observaciones en vivo como para las que usen video, son necesarios esquemas conceptuales que precisen los aspectos de lo que ocurre en el aula en los que deberá centrarse la atención. *Si usted no sabe en qué fijarse no verá gran cosa.* (Good y Brophy, 2010: 17-18)

Acercamientos basados en análisis de materiales

Los acercamientos al estudio de las prácticas docentes de los dos apartados anteriores se basan en lo que reportan los maestros mismos o en la observación de su trabajo. Otros acercamientos se basan en el análisis de los productos de las prácticas, como planes de clase, cuadernos de alumnos, exámenes aplicados o tareas encomendadas. Es, pues, una categoría distinta tanto de trabajos basados en lo que los docentes informan (en bitácoras, diarios u otro tipo de auto-reportes, o al responder las preguntas que se les dirigen con cuestionarios), como de los estudios que se basan en cualquier forma de observación.

En este tercer grupo de acercamientos se ubican unos muy antiguos y clave para trabajos de tipo histórico, basados en información documental (recogida en archivos o en documentos informales como cartas, diarios o cuadernos de alumnos), pero también otros más recientes y menos convencionales, que analizan materiales no textuales, como diversos tipos de objetos o artefactos; en este caso se distinguen los que se deben a la erosión o desgaste, y los que son producto de la acumulación. Los antecedentes de los acercamientos de este tipo incluyen trabajos relativamente antiguos como los basados en el análisis de restos de basura o los de huellas que dejan los niños en los cristales de los exhibidores de un museo. (*Cfr*. Webb, Campbell, Schwartz y Sechrest, 1966)

En este apartado, y siempre con el ejemplo del estudio de las prácticas docentes, se presentan solamente algunos trabajos que se aproximan a ese objeto de estudio a partir del análisis de materiales.

Los portafolios y sus variantes

Los portafolios son una herramienta ampliamente utilizada para la evaluación de alumnos, pero también de maestros o escuelas, inspirada en las carpetas en que pintores y otros artistas gráficos reúnen muestras de su obra para que los clientes se formen una idea de su producción. En el caso de la evaluación de maestros, los portafolios consisten en conjuntos de evidencias de su trabajo, como planes de clase, cuadernos de alumnos, registros de clases impartidas, entre otros.

Los portafolios contienen materiales seleccionados por el sujeto evaluado, del que se espera una reflexión sobre su situación, una *autoevaluación*, con base en los materiales incluidos en el portafolios, que luego serán valorados por otras personas, autoridades, pares u otros evaluadores externos.

La experiencia sobre el uso de portafolios ha mostrado su valor como herramienta de desarrollo profesional, enfatizando la dimensión formativa de la evaluación que se puede hacer mediante esa herramienta. Como ocurre con cualquier instrumento, el uso de portafolios para propósitos de investigación implica precisar los aspectos que interesan, por ejemplo, las dimensiones y los componentes particulares de la práctica que se quiere observar.

Algunos sistemas de evaluación de maestros piden a estos preparar portafolios, cuyo contenido se analiza como evidencia de ciertas dimensiones de la práctica. El sistema *Praxis* del *Educational Testing Service,* además de pruebas de lectura, escritura, matemáticas y conocimiento pedagógico, incluye un portafolios. El grado en que los maestros cumplen estándares de práctica profesional desarrollados por el Interstate *New Teacher Assessment and Support Consortium* (INTASC) se evalúa con portafolios, al igual que los estándares del *National Board for Professional Teaching Standards* (NBPTS). (Porter, Youngs y Odden, 2001: 263-265)

En Latinoamérica destaca el sistema de evaluación del Ministerio de Educación de Chile. El *Sistema de Evaluación del Desempeño Profesional Docente* (SEDPD) comenzó a desarrollarse en 2003, con base en un conjunto de estándares, el *Marco para la Buena Enseñanza*. Los documentos del SEDPD definen para qué se evalúa; qué se evalúa; qué consecuencias tiene la evaluación para los docentes; quiénes deben evaluarse; y que instrumentos y fuentes de información deben utilizarse. Los instrumentos incluyen una autoevaluación; una entrevista por un evaluador par; un informe de referencia de terceros (el director y el jefe de la Unidad Técnico Profesional de la escuela); y un portafolio. Para cada uno existen pautas que son aplicadas por personal entrenado para ello. (Manzi, González y Sun, 2011).

En seguida se presentan sistemas basados en el análisis de materiales, evidencias o artefactos, que constituyen variantes particulares de un portafolios, con propósitos específicos, siempre en el ámbito de los estudios sobre la práctica docente.

Análisis de tareas y similares

Estudiar las prácticas de un docente a partir del análisis de las tareas que pone a hacer a sus alumnos se basa en la idea de que esas tareas pueden revelar muchas cosas sobre la práctica del maestro, como la forma en que entiende su papel, el dominio que tiene de los contenidos, su concepción de la evaluación, el grado en que ofrece retroalimentación más o menos rica a los alumnos, entre otras. Esta idea se manifiesta en el subtítulo de un ejemplo temprano de este tipo de acercamiento: "Abriendo una ventana sobre las prácticas del aula". (Matsumura y Pascal, 2003)

Los autores, del *National Center for Research on Evaluation, Standards and Student Testing* (CRESST) de la Universidad de California en Los Ángeles, definieron la calidad de las tareas según tres dimensiones: nivel de demanda cognitiva; claridad de los objetivos de aprendizaje a los que una tarea debería contribuir; y claridad con que se especificaban criterios de calificación (Matsumura y Pascal, 2003).

Las puntuaciones asignadas en una escala en cada una de estas dimensiones se combinaron para formar una puntuación de calidad global. De cada maestro participante (181 de 35 escuelas, de los grados 4°, 7° y 10°) se recogieron materiales de tres tareas (dos sobre comprensión lectora y una sobre expresión escrita), incluyendo información sobre la planeación, las instrucciones dadas a los alumnos, las tareas entregadas por cuatro alumnos (dos de alto y dos de medio rendimiento) y se observó a los maestros en clase dos veces. Se analizaron las tareas recolectadas y los resultados se sintetizaron para formar indicadores de la práctica docente. Se encontró un nivel aceptable de consistencia entre los analistas, siendo necesarias tareas de tres o cuatro alumnos de cada maestro para tener estimaciones estables, y sólo si las tareas habían sido diseñadas por el maestro, y no tomadas de otra fuente.

La calidad de las tareas se asoció con la de la enseñanza, según las observaciones de la misma, y también con la calidad del trabajo de los alumnos; aquellos estudiantes cuyo maestro dejaba tareas cognitivamente más exigentes y aplicaba criterios de calificación más claros, mostraron también mayor avance en evaluaciones externas. Sin embargo, en general la calidad de las tareas asignadas por los maestros no fue muy alta. (Matsumura y Pascal, 2003).

Un acercamiento similar es el que consiste en el análisis de libretas o cuadernos de los estudiantes, desarrollado en especial para clases de ciencias, y en la perspectiva de la evaluación formativa, por Araceli Ruiz Primo y colaboradores. (*cfr*. por ejemplo Ruiz-Primo y Li, 2013).

Instructional Quality Assessment (IQA). Es un desarrollo del trabajo de Matsumura y colaboradores, que centra la atención en comprensión lectora y matemáticas, en el nivel medio. El sistema comprende protocolos para evaluar la enseñanza mediante observaciones en aula, así como la calidad de las tareas asignadas por el maestro. Se reportan trabajos para estudiar la confiabilidad y potencial validez de la información obtenida; coincidiendo con otros, encuentran amplia variación de la calidad de la enseñanza en las áreas estudiadas, con un nivel promedio no muy alto. La calidad de la información fue mejor en matemáticas que en lectura. Los resultados de la aplicación del IQA predicen los obtenidos por los alumnos en algunas evaluaciones externas, después de controlar otros factores posiblemente asociados. La solidez de los resultados es limitada porque se derivan de un número reducido de casos. (Matsumura *et al.*, 2006)

Scoop Notebook, EQAS

Scoop Notebook (Borko, Stecher y Kuffner, 2007). Instrumento para estudiar prácticas de enseñanza en matemáticas y ciencias en secundaria (middle school).

Analizar la expresión con la que se designa ayuda a entender el sentido del instrumento. El segundo término, *notebook*, alude simplemente a que los materiales recogidos se ponen en una libreta o carpeta. Interesa más el otro término: *scoop* quiere decir "red" y, en particular, una para atrapar mariposas u otros insectos o sacar peces de una pecera. Así como un biólogo se dedica primero al trabajo de campo, recopilando el mayor número posible de especímenes, que luego analizará en el laboratorio, así los investigadores de la práctica docente pueden primero dedicar tiempo a recolectar todas las evidencias que puedan (materiales, *artifacts*) del trabajo de docentes y alumnos, para luego estudiarlas en detalle.

El propósito del *Scoop Notebook* fue desarrollar un acercamiento alternativo al estudio de la práctica docente, empleando artefactos y materiales, para alcanzar a hacer una representación de dicha práctica de tal manera que una persona, sin observar directamente al maestro en el aula, pudiera hacer juicios válidos sobre algunos aspectos particulares, únicamente con base en esos materiales. Además de incluir instrucciones para los maestros

sobre cómo reunir los materiales, un componente clave es el que constituyen las guías de calificación que los jueces utilizan para analizarlos. El trabajo desarrollado se refiere en particular a las áreas de matemáticas y ciencias. (Borko, Stecher y Kuffner, 2007)

Las dimensiones para estudiar prácticas de ciencias (1. Agrupamientos de alumnos. 2. Estructura de clases. 3. Uso de recursos científicos. 4. Actividades con materiales (Hands-on). 5. Actividades indagatorias. 6. Profundidad cognitiva. 7. Presencia de discurso científico compartido. 8. Manejo de explicaciones y justificaciones. 9. Formas de evaluación utilizadas. 10. Conexiones con la vida o aplicación de conocimientos) se inspiraron en los Estándares Nacionales para Enseñanza de Ciencias (NRC, 1996) y la operacionalización de Le *et al.* (2006), por lo que sus categorías son anteriores a las de modelos más recientes, que llevaron a los Next Generation Science Stardards, pero son congruentes con ellos (Martínez, Borko y Stecher, 2012).

Quality Assessment in Science (eQAS). Se desarrolló para un estudio sobre prácticas de evaluación utilizadas en clases de ciencia. Como el Scoop Notebook, se basa en el análisis de evidencias que forman un portafolio, con la peculiaridad de que las evidencias o artefactos se recolectan por medio de dispositivos computarizados móviles. (Martínez *et al.* 2011 y 2012). Las 10 dimensiones en que se operacionaliza el concepto general de prácticas de evaluación son similares a las que se han comentado del Scoop Notebook. (Martínez *et al.*, 2012, Appendix, Tabla A1: 129)

En síntesis

El tercer grupo de formas de obtener información incluye investigación documental con material de archivos, cartas y otros, y técnicas que aprovechan otros objetos, a partir de la idea de que alguien puede engañar a un entrevistador, o dar respuestas falsas a un cuestionario, pero que suele dejar rastros de sus acciones, a partir de los cuales se pueden conocer de manera fidedigna algunos comportamientos.

Como todo instrumento, los de este grupo, tienen pros y contras, por lo que se deben ver como posibles complementos de los otros, no como alternativas excluyentes.

- Como la observación, el uso de materiales como fuentes de información tiene la ventaja de que no lo afecta el riesgo de captar lo deseable más que lo real.
- Como la observación que se hace con videograbaciones, los materiales pueden ser revisados una y otra vez por distintos analistas, utilizando distintos protocolos.
- El estudio de materiales o evidencias, también como la observación, no capta lo que pensaban los maestros al realizar las actividades de las que se derivaron

los productos, lo que sólo se podrá explorar a través de la versión de los actores, con acercamientos del primer grupo.

Se distinguen estudios de materiales producidos por erosión o por acumulación, que se distinguen porque se trata de *medidas de baja interferencia* (*unobstrusive*), como expresa el título de una obra clásica, que lleva como subtítulo *Nonreactive reserarch in the social sciences*. (Webb, Campbell, Schwartz y Sechrest, 1966)

Las nuevas tecnologías y la obtención de información

Al tratar de cercamientos basados en interrogación se mencionaron diarios en línea; en el inciso sobre observación se habló de videograbaciones; sistemas basados en análisis de materiales usan tabletas para recopilar evidencias de la práctica de los maestros. Pero la creciente importancia de la tecnología hace necesario este inciso.

Hace más de diez años una obra señalaba las posibilidades de recolección de datos en Internet (Best y Krueger, 2004). Vehovar y Lozar Manfreda recuerdan, por su parte, que *la investigación siempre ha estado abierta a los nuevos avances tecnológicos, comenzando con las encuestas telefónicas en la década de 1960, y las asistidas por computadora en la de 1980* (2011: 177). Estos autores distinguen modalidades de lo que genéricamente llaman *Computer-assisted survey information collection* (CASIC), que incluye aplicación de cuestionarios o entrevistas con apoyo de computadora realizadas presencialmente, por teléfono o por videoconferencia, pidiendo que se respondan preguntas de opción múltiple mediante las teclas del teléfono, o respondiendo cuestionarios digitales enviados por correo electrónico, o a los que se accede en una página web. (Vehovar y Lozar Manfreda, 2011: 178)

Paquetes de software para crear y aplicar cuestionarios en línea (*Survey Monkey, Lime Survey, Question Pro, Qualtrics* y otros), tienen versiones limitadas gratuitas, y más completas para instituciones. Estas opciones tienen ventajas, como reducir costos de envío y recuperación de cuestionarios o captura manual de respuestas, pero también desventajas, en cuanto a representatividad de las muestras y tasas de respuesta. El piloteo es indispensable en versión de lápiz y papel o electrónica.

Benfield y Szlemko (2006) discuten *promesas y realidades* de técnicas de obtención de información vía Internet, pero las posibilidades de las tecnologías digitales incluyen mucho más: etnografía virtual (Hine, 2011); grupos focales en línea (Gaiser, 2011); investigación de blogs (Wakeford y Cohen, 2011; Bruns y Burgess, 2012); *experience sampling* (Hektner, Schmidt y Csikszentmihalyi, 2007); rastreo de huellas digitales (Welser, Smith,

Fisher y Gleave, 2011; Janetzko, 2011). Sharpe y Benfield (2012) mencionan la variante de "amigo corresponsal" (*pen pal*) para entrevistas por correo electrónico, o envío de mensajes de texto para obtener respuestas con una frecuencia imposible de alcanzar por otros medios, lo que permite un muestreo fino de la actividad de una persona en el tiempo.

Una obra reciente ha llamado la atención las posibilidades del uso de las enormes bases de datos que generan y acumulan redes digitales como Google y otras. La obra lleva el título de *Todos mienten* (*Everybody Lies*), y el subtítulo *Big Data*, *New Data, and what the Internet can tell us about who we really are*. El autor escribe:

> La gente miente sobre cuantos tragos tomó camino a casa, qué tanto va al gimnasio [...] o si ya leyó tal libro. Se reporta enferma cuando no lo está [...] Dice que te ama cuando no es verdad; que está contenta cuando está en la basura; que le gustan las mujeres cuando en realidad le gustan los hombres. La gente miente a sus amigos; miente a sus jefes; a sus niños; a sus padres; a su médico; a su marido y a su mujer. Y sin la menor duda (damn sure) miente a las encuestas [...] (Stephens-Davidowitz, 2017: 105)

El autor considera que, al navegar en la red, los usuarios dejan rastros, y que al hacerlo es más probable que actúen de acuerdo a su forma real de pensar, que al responder un cuestionario o al ser entrevistados.

Recordando la vieja noción de *respuestas socialmente deseables,* Stephens-Davidowitz opina que mentir en una encuesta puede ser hoy más frecuente, y aduce el dato de que 99% de jóvenes que en una encuesta dijeron que usaban una prótesis por haber perdido un miembro, confesó que había mentido simplemente por bromear. (2017: 108, nota)

Stephens-Davidowitz ejemplifica su punto de vista comentando que el fracaso de las encuestas para predecir el resultado de la elección presidencial que llevó a Donald Trump a la Casa Blanca subestimó el peso del racismo, que un análisis de las búsquedas en las redes digitales muestra que está más extendido de lo que se pensaba, y esto tanto en estados que tradicionalmente votan a favor del Partido Republicano, como en los que lo hacen a favor del Demócrata.

Otro ejemplo se refiere a la proporción de hombres con tendencias homosexuales, que según las encuestas es muy superior en lugares con disposiciones legales tolerantes, en comparación con lugares en que hay disposiciones discriminatorias, mientras que la proporción de personas que hacen búsquedas en Google sobre páginas porno de contenido homosexual, o sobre tests de homosexualidad, es muy similar en todos los estados. (Stephens-Davidowitz, 2017: 1-22 y 112-122)

En el prefacio que escribió para el libro de Stephens-Davidowitz, Steven Pinker señala que los filósofos han especulado sobre un "cerebroscopio", un aparato mítico que mostraría los pensamientos de una persona, y añade:

> Este libro trata de una manera totalmente novedosa de estudiar la mente. Los Big Data que producen las búsquedas en internet y otras respuestas en línea no son un cerebroscopio, pero Stephens-Davidowitz muestra que ofrecen una mirada sin precedentes a la psique de la gente. En la privacidad de su teclado, la gente confiesa las cosas más extrañas, unas veces (como en las páginas de citas románticas o en las búsquedas de asesoría profesional), porque tienen consecuencias en la vida real, otras precisamente porque no tienen consecuencias: la gente puede sentirse liberada del peso de algún deseo o temor cuando no hay alguien que reaccione con consternación o en una forma peor. En ambos casos la gente no está simplemente apretando un botón o girando una perilla, sino escribiendo una secuencia de caracteres entre trillones posibles, para expresar sus pensamientos en toda su explosiva inmensidad combinatoria. Y todavía mejor, dejan esas huellas digitales en una forma que es fácil de agregar y analizar [...] Pueden ser parte de experimentos no invasivos (unobstrusive) que varían los estímulos y tabulan las respuestas en tiempo real [...] (2017: x-xi)

El cuidado de la calidad de la información

El riesgo de que una persona responda errónea o falsamente una pregunta es muy conocido, pero la información derivada de observaciones también dista de ser perfecta. Así lo muestra un trabajo de Chabris y Simons (2009), que describe un experimento reportado por primera vez en la revista *Perception* en 1999, y repetido muchas veces después, siempre con los mismos sorprendentes resultados.

En el experimento se mostraba un video de alrededor de un minuto de duración, en el que aparecían dos equipos de personas, con playeras blancas y negras, que se movían y se pasaban una pelota de basquetbol. Se pedía a los observadores que contaran las veces que las personas con camiseta blanca se habían pasado la pelota, y al terminar se les interrogaba al respecto, pero además se les preguntaba si habían notado algo inusual, lo que alrededor de la mitad de los sujetos negó.

El elemento inusual que muchos observadores no detectan era que, hacia la mitad del video, entraba por un lado del escenario un personaje disfrazado como gorila, caminando lentamente entre los jugadores, se golpeaba el pecho y seguía caminando hasta salir por el lado opuesto. Chabris y Simons escriben:

> Para nuestra sorpresa, ¡alrededor de la mitad de los sujetos [...] no habían notado el gorila! Cuando volvieron a mirar el video [...] lo detectaron fácilmente y quedaron atónitos. Algunos, de manera espontánea, dijeron: "¿No vi eso?" o "¡No puede ser!" (2011: 25)

La sorpresa hizo incluso que algunos sujetos acusaran a los investigadores de tener dos videos, uno en el que el gorila no aparecía y otro en el que sí lo hacía. El video puede verse en Internet (www.theinvisiblegorilla.com) y es fácil comprobar que una considerable proporción de personas cometen el mismo error, desde luego si no se les advierte previamente de qué se trata.

> ¿Cómo puede la gente no ver un gorila que camina delante de ellos, gira para mirarlos, se golpea el pecho y se va?... Este error de percepción proviene de una falta de atención hacia el objeto no esperado, por lo que en términos científicos se lo denomina "ceguera por falta de atención [...]" las personas no ven el gorila, pero no debido a un problema en sus ojos. Cuando dedican su atención a un área o aspecto particular [...] tienden a no advertir objetos no esperados, aun cuando estos sean prominentes, potencialmente importantes y aparezcan justo allí donde ellos están mirando. (2011: 25)

Un creciente número de estudios sobre la calidad de la información que ofrecen los testigos presenciales de un delito sobre lo que vieron, revela los límites tanto de la capacidad de observación como de la fidelidad de la memoria, contra lo que se suele creer de tales testimonios.

Un especialista recuerda que, en 1981, un juez de la Suprema Corte de Estados Unidos escribió: *No hay nada tan convincente como un ser humano que toma el estrado, señala con el dedo al acusado y dice: "ese fue",* y añade:

> Según cientos de estudios de los últimos 30 años, casi no hay nada tan poco fiable como lo que un testigo presencial piensa que vio. La memoria no es una videograbación. Podemos creer que recordamos cosas con precisión, pero la mayor parte de nuestros recuerdos mezcla lo que pensamos que observamos con la información a la que hemos estado expuestos desde entonces. La situación es peor en escenas de crímenes, en las que interfieren con la precisión variables como el estrés y la presencia de un arma. Si se considera la memoria como evidencia, como hace la mayoría de psicólogos, entonces es la evidencia más frágil y fácil de contaminar. (Starr, 2012: 40)

Estudios controlados para valorar la calidad de testimonios muestran que muchas personas se equivocan al tratar de identificar a alguien que vieron en un lugar bien iluminado entre varios sospechosos, o que su respuesta cambia según las palabras exactas que se usen al interrogarlos. Gracias a esos trabajos, los procedimientos para interrogar testigos pueden mejorar, con las implicaciones que tiene p*ara los millares de personas inocentes que han sido condenados... y para todos nosotros, cuando la persona que cometió realmente un delito sigue libre.* (Starr, 2012: 64)

En el terreno que nos ocupa, la conciencia de los límites de todo acercamiento a la obtención de información debe llevar al investigador a hacer lo posible por mejorar la calidad de sus datos, evitando al máximo los errores que la ponen en riesgo. En la terminología del campo la calidad de la información se concreta en términos bien conocidos, de apariencia engañosamente sencilla: confiabilidad y validez.

En este último apartado del capítulo se presentan conceptualmente ambas ideas, intentando dejar claro lo fundamental y dar una idea de la complejidad que implica su tratamiento pleno, a partir de una visión de su desarrollo a lo largo de más de un siglo. Se deja para el capítulo siguiente lo relativo a las técnicas para valorar tanto la confiabilidad como la validez. Esta división del tratamiento del tema en dos partes en capítulos diferentes se entiende si pensamos que la calidad de la información que se obtiene para una investigación debe atenderse desde que se seleccionan o preparan las técnicas para recabarla, pero también debe revisarse una vez que se ha recabado, lo que es ya un tipo de análisis de la información.

Primer acercamiento a la confiabilidad y la validez

Ambas nociones se refieren a la calidad de la información. Por ello una primera aproximación a ellas puede hacerse a partir de dos tipos de amenazas que pueden afectar dicha calidad o, si se prefiere, dos tipos de errores que se pueden cometer al recoger información: errores aleatorios y errores sistemáticos. Para introducir las dos ideas pueden servir ejemplos del ámbito de la medición de propiedades físicas, como la estatura y el peso de las personas.

Si pedimos a los alumnos de un grupo de primaria que se midan y pesen, con una cinta que marca centímetros y decímetros, y con una báscula de baño de resorte, y hacemos que las dos mediciones se repitan varias veces en distintos días de la semana y diferentes horas del día, los resultados que se obtendrán no mostrarán una coincidencia perfecta, sino ciertas variaciones, de dos tipos: unas variaciones que fluctúan de manera que parece aleatoria, en tanto que en otros casos parecerá haber un patrón definido, con mayor o menor altura o peso en unas u otras.

En el primer caso la posible explicación no se refiere a una causa precisa, y es probable que se deba más bien a errores no sistemáticos atribuibles al descuido de quienes midieron o, tal vez, a una báscula defectuosa.

En el segundo caso hay una razón que explica el patrón detectado. Es posible que en una ocasión se haya encuentre una estatura sistemáticamente menor, porque las medidas se hicieron cuando los alumnos se habían quitado los zapatos para una actividad gimnástica. Una estatura regularmente mayor de un grupo de niñas pudo deberse, a su vez, a que todas se pusieron zapatos de tacón par un bailable. Y podría ocurrir que el mayor peso encontrado en ciertas ocasiones se deba a que se tomó justo después de que los estudiantes habían tomado alimentos.

En el caso de variaciones debidas a errores aleatorios, se aprecia inconsistencia en las mediciones, que no son confiables; en el caso de variaciones sistemáticas hay consistencia, pero el resultado no corresponde del todo a lo que se pretende medir, que no es el tipo de calzado de los estudiantes, o la abundancia de su comida.

En la investigación educativa y social hay ejemplos más realistas de ambos tipos de errores, y concretamente en relación con las técnicas de obtención de información de los tres grupos considerados en apartados anteriores de este capítulo.

Si se analizan las respuestas de un grupo de sujetos a preguntas de un cuestionario, es posible encontrar distribuciones atribuibles a los dos tipos de error. Si se pregunta la edad de los sujetos, y se comparan las respuestas con la edad que consta en el documento de identidad, es posible que se encuentren diferencias no sistemáticas, atribuibles probablemente a errores involuntarios de los informantes. Pero si las respuestas muestran una tendencia regular a dar una edad menor o mayor a la real, entonces la explicación no es un error aleatorio, sino un error sistemático, o sesgo, que lleva a unas personas a declarar una edad menor o mayor a la real, tal vez por coquetería, o para tener acceso a una actividad reservada a los adultos.

Los numerosos casos que muestran que no se puede creer sin más en lo que dicen las respuestas a un cuestionario son ejemplos de estos tipos de error.

Si se pregunta a unas personas cuántas veces han ido al médico en el último año, y se comparan las respuestas con registros fidedignos, las cifras seguramente no coincidirán perfectamente, pero es posible que no aparezca un patrón claro en las diferencias, lo que reflejaría errores aleatorios por la imprecisión de la memoria; si aparece, en cambio, un patrón claro, la explicación será un sesgo que busca tal vez dar una impresión de mayor cuidado por la salud.

Las distorsiones por lo sensible de ciertas preguntas o por el sesgo de deseabilidad social aparecen una y otra vez en distintas áreas. Las fallas de las encuestas de intención de voto se explican en parte porque muchas personas no quieren decir realmente por quién se inclinan. Quienes estudian el fenómeno religioso saben que la capacidad de todas las iglesias, y todos los servicios dominicales de una ciudad, es mucho menor al número de quienes en las encuestas al respecto informan que van a misa cada domingo. En otro campo, Stephens-Davidowitz llama la atención sobre el hecho de que el número de preservativos que se vendió en los Estados Unidos en un año no corresponde al que puede calcularse con base en el número de relaciones sexuales, y la proporción de relaciones que se llevan a cabo con y sin protección, según las encuestas al respecto, en las que por otro lado tampoco hay coincidencia entre lo que dicen las mujeres y los hombres. (2017: 5)

En todos estos casos no estamos ante errores aleatorios debidos a la imprecisión de la memoria o a la ambigüedad de la pregunta, sino ante errores sistemáticos o sesgos: los informantes quieren que quienes les preguntan piensen que van a votar por A y no por B, que son más religiosos, o que tienen una vida sexual más activa y más responsable de lo que es en realidad. En estos casos las respuestas pueden ser muy consistentes, pero no están ofreciendo información verídica sobre lo que se pregunta. En realidad, no se está midiendo lo que se quiere medir, sino otra cosa, lo que es la definición convencional de validez, como se verá más adelante.

Fallas como estas no son exclusivas de las encuestas por cuestionario. Si se hacen preguntas similares en entrevistas no hay razón para esperar respuestas mejores; el riesgo de deseabilidad social, o la incomodidad de preguntas sensibles, es sin duda mayor. Las ventajas de que un entrevistador puede obtener mejor información al tener la posibilidad de insistir, preguntar algo de diferentes formas, o granjearse la simpatía del entrevistado, tiene las desventajas correlativas de que un informante se puede sentir intimidado por un entrevistador de distinto sexo, raza o edad, o bien que puede sentir la tentación de engañarlo, aunque sea simplemente por bromear.

En una primera aproximación, pues, la *confiabilidad* se refiere a la consistencia de la información obtenida que se consigue cuanto no hay errores aleatorios, o no sistemáticos. La *validez*, por su parte, es la cualidad de la información que se debe a la ausencia de errores sistemáticos, o sesgos.

Veamos en seguida las dos nociones de una manera más técnica, con una idea de su complejidad, a partir de su desarrollo histórico.

Confiabilidad: consistencia y precisión de las mediciones

Hoy es claro que confiabilidad y validez son ideas distintas, aunque relacionadas, pero cuando se comenzaron a usar a principios del siglo XX no era claro. En la siguiente cita, un pionero de la psicometría, Truman Kelley, define la confiabilidad en términos que hoy se aplican más bien a la validez, como se verá en seguida:

> Hay que entender por confiabilidad el grado en que una prueba mide lo que realmente mide, que no es necesariamente lo que quiere medir. By reliability is to be understood the extent to which the test measures that which it in reality does measure-- not necessarily that which it is claimed to measure. (1921: 370)

Las dos nociones surgieron en relación con los primeros esfuerzos por extender la metodología de las pruebas de inteligencia a la evaluación de los aprendizajes y a la medición de actitudes. Para mediados del siglo XX habían alcanzado su lugar como conceptos centrales en la *Teoría Clásica de la Medición* y, de manera más general, en las ciencias sociales y de la conducta. Luego las dos nociones han seguido precisándose en formas cada vez más complejas.

El concepto de confiabilidad fue introducido en 1904 por Charles Spearman, que seis años más tarde lo definió como *el coeficiente de correlación entre una mitad y la otra de varias mediciones de la misma cosa*. Edward L. Thorndike retomó la noción en un libro publicado también en 1904, con el que *lanzó la medición educativa y psicológica en los Estados Unidos*... (Stanley, 1971: 370 y 371)

En la primera edición de la obra *Educational Measurement*, Robert L. Thorndike comenzaba definiendo la confiabilidad a partir de su opuesto, diciendo:

> Cada vez que medimos algo [...] esa medición tiene cierta cantidad de error aleatorio, grande o pequeño, pero omnipresente [...] las discrepancias pueden expresarse en millas o en millonésimas de milímetro, pero aparecerán siempre, si las unidades son suficientemente finas en relación con la precisión de las medidas. El que conjuntos repetidos de medidas nunca se dupliquen exactamente es lo que se quiere decir con la expresión "no confiabilidad". Al mismo tiempo, medidas repetidas de una serie de objetos o individuos mostrarán, por lo general, cierta consistencia [...] lo opuesto a la variación a la que nos acabamos de referir, y que designaremos como "confiabilidad". (Thorndike, 1951: 560)

Thorndike presenta seis grupos de posibles fuentes de variación de los puntajes de una prueba y procedimientos para estimar su confiabilidad, mediante coeficientes de correlación, a partir de la aplicación de formas equivalentes, de la aplicación reiterada de una forma o de subconjuntos de una misma prueba, y a partir del análisis de la varianza entre ítems. (1951: 568, Tabla 8)

Thorndike permite distinguir dos cualidades relacionadas, pero no idénticas, de una medición, su precisión y su consistencia, al señalar que se puede calcular el tamaño de los errores de medición mediante la desviación estándar de la distribución de los resultados —*el error estándar de la medición*— o estimar la consistencia entre dos conjuntos de puntuaciones, según el coeficiente de correlación correspondiente, que pasa a ser un *coeficiente de confiabilidad*. Thorndike explica la relación entre coeficiente de confiabilidad y Error Estándar de Medición (EEM), que ayuda a evitar interpretaciones simplistas del primero. (Thorndike, 1951: 610).

La relación entre el valor de un coeficiente de correlación y el tamaño *relativo* del EEM correspondiente ayuda a interpretar correctamente los coeficientes de confiabilidad, que no debe hacerse como si se tratara de porcentajes. Lo anterior implica que, tratándose de mediciones de un gran número de casos y diferencias pequeñas de los puntajes de esos casos, aunque el coeficiente de confiabilidad de la medición de que se trate sea alto, el tamaño del EEM podrá ser suficiente para que un ordenamiento de los casos basado en esos resultados sea muy impreciso, ya que habrá muchos traslapes entre los intervalos de confianza.

El capítulo sobre confiabilidad de la segunda edición de *Educational Measurement* comienza retomando el primer párrafo del texto de Thorndike citado antes (Stanley, 1971: 356), lo que indica que el concepto no sufrió cambios en las dos décadas que separan ambas ediciones. Esto se confirma al ver que Stanley mantiene la distinción entre los dos enfoques para estimar la confiabilidad, a partir del error estándar de la medición o del coeficiente de confiabilidad, y que retoma literalmente la tabla en la que Thorndike presenta las fuentes de variación de los puntajes de una prueba. El resto del capítulo presenta procedimientos estadísticos para calcular confiabilidad, mostrando diversas fórmulas, sus ventajas, la dificultad para calcularlas en una época en que no había computadoras, y la equivalencia de algunas, incluyendo el coeficiente alfa, introducido por Lee Cronbach en 1951. (Stanley, 1971)

En la tercera edición de *Educational Measurement*, el capítulo sobre confiabilidad aporta dos elementos: desarrollos del enfoque correlacional para casos particulares, presentando doce coeficientes de consistencia interna: cuatro para casos de subdivisión

de la prueba en dos partes; uno para subdivisión en tres partes y siete para un número indefinido de partes (Feldt y Brennan, 1989: 115); y la presentación de la *Teoría de la Generalizabilidad* (TG), basada en los trabajos de Cronbach, Rajaratnan y Gleser (1963) y Gleser, Cronbach y Rajaratnam (1965), y desarrollada ampliamente por Cronbach *et al.* (1972). Según Feldt y Brennan, la TG:

> [...] puede ser vista como una extensión y liberalización de la teoría clásica, que se logra básicamente gracias a la aplicación del análisis de varianza a los datos de la medición. En la teoría clásica el error de medición se ve como una entidad unitaria, global, aunque se reconoce que se deriva de una combinación de fuentes. En contraste, los modelos y métodos de la teoría de la generalizabilidad se interesan por los errores derivados de esas múltiples fuentes como entidades separadas [...] (1989: 127-128)

El capítulo sobre confiabilidad de la 4a edición de *Educational Measurement* señala que los principios de las teorías clásica y de la generalizabilidad siguen siendo válidos, y que la mayoría de los materiales nuevos son desarrollos de los anteriores, como los relativos al error de medición en el caso de medias grupales, y con la confiabilidad de clasificaciones y de observaciones. (Haertel, 2006: 99-103)

En otro trabajo reciente, Brennan señala que, al igual que la validez, la confiabilidad tampoco es una propiedad que se pueda predicar de una prueba u otro instrumento de obtención de información. La consistencia con que se define la noción se refiere a los datos que se obtienen, no al instrumento con el que se obtuvieron. El autor considera las implicaciones que tiene para la confiabilidad la noción de *réplica*, en el sentido de un proceso que duplique *lo más exactamente que sea posible* las condiciones de una aplicación previa, partiendo de que es imposible conseguir una réplica perfecta, ya que una nueva aplicación implica inevitablemente cambio en al menos algunos aspectos del proceso. Esta idea es similar a la idea central de la TG, de que no hay un solo tipo de error en el resultado de cualquier medición sino varios, que se derivan de múltiples fuentes: el instrumento, desde luego, pero también las ocasiones en que se hace una aplicación, incluyendo la original y sus réplicas, los aplicadores o calificadores, entre otras. Por ello el autor sostiene que la noción de réplica es fundamental para definir la confiabilidad, que él expresa como sigue: "la confiabilidad es una medida del grado de consistencia de los puntajes de los sustentantes en las réplicas del procedimiento de medición". (Brennan, 2001: 296)

Es todo el procedimiento de medición, y no sólo el instrumento, lo que afecta tanto la precisión como la consistencia de los resultados. Brennan concluye:

> En mi opinión no puede haber respuestas significativas a las preguntas sobre la confiabilidad sin una consideración expresa de la naturaleza de las réplicas (planeadas y efectivas) de un procedimiento de medición. Un marco coherente para conceptualizar, calcular e interpretar la confiabilidad requiere, por lo tanto, que se responda la pregunta de qué constituye una réplica de un procedimiento de medición. (Brennan: 2001: 313)

La validez, o medir lo que se quiere

La definición más simple de validez es la que la concibe como la cualidad que consiste en que un instrumento de obtención de información mida efectivamente lo que pretende en principio medir.

Esta conceptualización distingue con claridad la validez de la confiabilidad, y tiene en cuenta la idea clave de que el conocimiento humano no se reduce simplemente a lo que capta nuestra percepción, sino que todo concepto supone algún tipo de salto o inferencia, a partir de lo percibido, pero que no se reduce a ello.

Obviamente el desarrollo de la noción de validez ha sido más complejo, incluyendo variantes y dimensiones cada vez más numerosas. Para mostrar esa evolución, se retoman los capítulos respectivos de las ediciones de *Educational Measurement*, de 1951 a 2006, así como referencias más recientes, lo que permite explorar cómo se ha modificado y enriquecido el concepto a lo largo de siete décadas.

El capítulo que interesa de la primera edición de la obra citada comienza diciendo que la pregunta clave sobre la validez de una prueba es *qué tan bien ejecuta la función para la que fue empleada* (Cureton, 1951: 621). Una prueba puede usarse con distintos propósitos y su validez puede ser alta para uno, moderada para otro y baja para un tercero, por lo que no se puede etiquetar en general como de alta, moderada o baja validez. Por otra parte, la validez no acaba de distinguirse de la confiabilidad, y se conceptualiza con base en su correlación con un criterio:

> La validez tiene dos aspectos, que se pueden denominar relevancia y confiabilidad. Relevancia se refiere a la cercanía o coincidencia entre lo que la prueba mide y la función que se pretende medir con ella. Confiabilidad se refiere a la precisión y consistencia con que mide lo que sea, en el grupo en el que se usa [...] La validez se define, pues, en términos de la correlación entre los puntajes que arroja la prueba y los puntajes criterio verdaderos (true criterion scores). Un puntaje verdadero es la parte del puntaje obtenido que no es error de medición [...] (Cureton, 1951: 622-623)

En 1954 la *American Psychological Association (*APA) publicó recomendaciones entre las que incluyó 19 estándares para cuidar la validez. La *American Educational Research Association* (AERA) hizo lo propio en el área de la educación en 1955, distinguiendo cuatro tipos de validez: de contenido, predictiva, concurrente y de constructo. En 1966 APA y AERA, publicaron una versión conjunta y revisada de los criterios, titulada *Standards for Educational and Psychological Tests*, en la que los tipos de validez se integraron en tres: de contenido, de criterio y de constructo.

En la 2ª edición de *Educational Measurement*, validación "es el proceso de examinar la exactitud de una predicción específica o una inferencia hecha a partir del puntaje de una prueba o de los resultados de instrumentos de medición, como cuestionarios, observaciones, calificaciones de desempeño, etc.". (Cronbach, 1971: 443)

La validez se refiere a la robustez de las interpretaciones descriptivas o explicativas hechas con base en los resultados de un instrumento de medición. Para explicar las puntuaciones de una prueba hay que utilizar una teoría que ayude a explicar cierto desempeño y sus implicaciones relacionadas con el fenómeno que se mide. Validar las interpretaciones de los resultados de una prueba es similar a la evaluación de cualquier teoría científica: se privilegia la validación o refutación de la teoría, y no el procedimiento de medición empleado. Cronbach advertía que, en la visión estrecha que prevaleció hasta 1950, validar era comparar el puntaje en una prueba con otra observación que sirviera como criterio. Se trataba de predecir ese criterio y el mérito de una prueba se juzgaba simplemente por la precisión de la predicción. (1971: 443)

Cronbach consideraba cinco tipos de validación, según el uso de la prueba. Si el foco se pone en la solidez de las interpretaciones descriptivas, los tipos de validación son: validez de contenido, importancia educativa y validez de constructo. Si la prueba se usa para tomar decisiones sobre personas —a partir de un criterio—, los tipos de validación son validez para la selección y validez para la colocación. Los estudios de validez, según el tipo de validación en la que se enfoquen, tendrán preguntas claves a responder e implicarán hacer juicios adecuados a partir de las respuestas. (1971: 446. Tabla 14.1)

El autor señala que cuando alguien emplea la frase *validación de una prueba* refleja un mal entendimiento del concepto. El investigador no valida una prueba, sino la interpretación de datos derivados de un procedimiento específico. Un instrumento puede ser usado de diferentes maneras. Una prueba de lectura puede ser utilizada para seleccionar aspirantes a una carrera profesional, para planear instrucción remedial en lectura, medir la efectividad de un programa instruccional, entre otros propósitos. Dado que cada uso se basa en una interpretación diferente, la evidencia que justifica

una utilización puede tener poca relevancia para otra. Y como cada interpretación tiene su propio grado de validez, nunca se puede llegar a la simple conclusión de que una determinada prueba «es válida». (Cronbach, 1971: 447-448)

Algunos diseñadores de pruebas, erróneamente, hacen un análisis de correlación para reducir la cantidad de ítems que se correlacionan de manera positiva. Sin embargo, nada en la lógica de la validez de contenido requiere que el universo de contenidos a evaluar sea homogéneo. Si el universo de contenidos a evaluar es heterogéneo, de hecho, una alta correlación entre los ítems indica un muestreo de contenidos inadecuado. La correlación de los reactivos con un criterio es irrelevante para la validez de contenido. (Cronbach, 1971: 457-458)

Cronbach agrupa los procedimientos para examinar interpretaciones en tres clases: correlacionales, experimentales y lógicos. Un procedimiento correlacional determina cómo difieren las personas con altos o bajos puntajes en una prueba. Una persona con un alto puntaje en un test que mide un constructo latente debería puntuar alto en otros indicadores del mismo constructo, o en otras pruebas que midan lo mismo. Por ello el análisis factorial es importante en los estudios correlacionales, ya que los ítems que correspondan a una misma dimensión deberán tener cargas sustanciales en un solo factor (1971: 469). Los procedimientos experimentales buscan modificar el rendimiento de una persona en una prueba con una intervención controlada. Los procedimientos lógicos del contenido de una prueba o de las reglas de calificación, pueden revelar influencias preocupantes en la puntuación. Un caso es el de ciertas medidas de logro, que son inválidas porque tienen un techo bajo; por ejemplo, los alumnos que en una prueba previa obtienen altos puntajes, sólo pueden ganar unos pocos puntos en la prueba posterior. (Cronbach, 1971: 465-475)

Así pues, al menos desde los años cincuenta y hasta inicios de los noventa del siglo pasado, en la noción de validez se solían distinguir tres tipos, uno de ellos con dos subtipos, a saber: validez de contenido, validez de criterio —predictiva o concurrente— y validez de constructo.

Samuel Messick comienza con esta definición el capítulo *Validez* de la 3ª edición de *Educational Measurement*:

> La validez es un juicio evaluativo integral del grado en que la evidencia empírica y los argumentos teóricos apoyan lo adecuado y apropiado de inferencias y acciones basadas en puntajes de pruebas y otras formas de evaluación [...] la expresión puntajes de pruebas se usa aquí genéricamente, para significar en el sentido más amplio cualquier consistencia

observada no solo de pruebas en el sentido usual, sino de cualquier medio de observar o documentar conductas o atributos consistentes. (1989: 13)

El autor añade que la validez es cuestión de grado, no de todo o nada. Además, con el tiempo, la evidencia de validez se fortalece o debilita por nuevos hallazgos, y las proyecciones de las posibles consecuencias sociales de las evaluaciones se transforman a partir de la evidencia sobre las consecuencias reales en la actualidad y las condiciones sociales cambiantes. Entonces, inevitablemente, la validez es una propiedad en evolución y la validación es un proceso continuo. (1989: 13)

Luego retoma las formulaciones de APA-AERA sobre los tipos de validez:

- La *validez de contenido* valora qué tan bien cubre el instrumento situaciones o temas sobre las que se harán conclusiones. Se basa en el juicio profesional sobre la relevancia del contenido del test para medir un comportamiento en particular o un campo de interés y si las tareas que comprende representan bien dicho campo. No tiene que ver con el proceso para responder, la estructura interna y externa del test, las diferencias en el desempeño, la respuesta a cierta intervención o a las consecuencias sociales.
- La *validez de criterio* se evalúa al comparar los puntajes de la prueba con una o más variables externas —criterios— que se considera proveen una medición directa de las conductas o características en cuestión; comprende:
  - *Validez predictiva*, que indica el grado en que cierto nivel desempeño en la variable criterio es predicho a partir de una prueba aplicada antes.
  - *Validez concurrente*, sobre el grado en que una prueba estima el real desempeño de un individuo en la variable criterio, con base en la relación empírica —asociación o correlación— entre los puntajes de la prueba y las puntuaciones consideradas criterio. No se trata del grado en que la variable evaluada con la prueba se asocia con otras, sino con la misma variable (criterio) que se considera medida directamente.
- La *validez de constructo* indaga el grado en que un instrumento mide un constructo o variable compleja. Se basa en la integración de toda evidencia que apoye la interpretación o significado de las puntuaciones, que no son consideradas equivalentes al constructo que se mide, sino que son posibles indicadores de una variable latente. (Messick, 1989: 16-17)

La *validez de constructo* subsume la de contenido —relevancia y representación de un dominio— y la de criterio, porque la información que se obtiene mediante ellas contribuye a la interpretación de los puntajes. Por tanto, la validez de constructo incluye la mayoría de las evidencias de validez. Messick deja claro, sin embargo, que su propósito es analizar los límites de esa clasificación tradicional de los tipos de validez, y proponer una visión alternativa, a partir de un concepto unificado de esta fundamental cualidad de una buena medición. (1989: 16)

Que las evidencias de validez de contenido y criterio estén incluidas en la validez de constructo deja una sola gran categoría, pero en la práctica, la evidencia general que soporta la validez de constructo necesita apoyarse por evidencia específica de la relevancia de la prueba para ser aplicada con cierto propósito y sobre su utilidad en cierto escenario. El que las evidencias de validez de contenido y de criterio aporten a la validez de constructo y que por ello se considere que toda evidencia de validez es de constructo, puede ocasionar que no se distingan los matices o aportaciones específicas de los primeros dos tipos de evidencia, como adjuntas a la validez de constructo, para justificar los usos de una prueba (Messick, 1998).

Por otro lado, al ver cómo evolucionó la noción de validez hasta que la dimensión de constructo fue omnipresente, se notará que la única fuente de evidencia que no es explícitamente incorporada es la validez que evalúa las consecuencias sociales. Según Messick es irónico que los estudios de validez hayan puesto poca atención a los usos y consecuencias, porque la validez en sus inicios fue concebida en términos funcionales: qué tan bien la prueba hace la tarea para la que fue diseñada.

Partiendo de que la validez es un concepto unitario, Messick propone un marco que distingue dos facetas relacionadas. Una es la justificación de la prueba, que se basa en la evaluación de evidencias o consecuencias. La otra es la función o resultado de la prueba, que es la interpretación de sus puntajes o los usos que se hacen de ellos. Al cruzar las facetas se obtienen cuatro clasificaciones.

TABLA 3.5. FACETAS DE LA VALIDEZ

Resultados / Justificación	Interpretación de la prueba	Uso de la prueba
Basada en evidencias	Validez de constructo	Validez de constructo + Relevancia/Utilidad
Basada en consecuencias	Implicaciones de valor	Consecuencias sociales

FUENTE: MESSICK, 1989: 20. TABLA 2.1.

Messick aclara expresamente lo que entiende por evidencia:

> Por evidencia se entiende tanto los datos o hechos, como los argumentos que conjuntan esos datos en una justificación de las inferencias de unos puntajes. Otra forma de decirlo es que los datos no son información; esta es el resultado de la interpretación de ciertos datos [...] O como dice Kaplan, lo que sirve como evidencia es el resultado de un proceso de interpretación; los hechos no hablan por sí mismos; sin embargo, hay que escuchar los datos o se pierde el carácter científico del proceso de interpretación. (1989: 15-16)

Messick concibe la validación como un proceso de investigación, y como una tarea científica, y discute ampliamente las cuestiones filosóficas implicadas en ello (1989: 23-34), y a partir de ello en el resto del capítulo desarrolla el contenido de las cuatro celdas de la tabla anterior, así como a los rubros del encabezado de las dos columnas y de sus dos renglones. (1989: 34-92)

Por último, en el capítulo Validación de la 4ª edición de *Educational Measurement* Michael Kane aclara que validación y validez suelen tener dos usos distintos pero relacionados. En el primero, validación se usa para generar evidencia que sustente las interpretaciones y usos que se quiere hacer de los resultados de un instrumento; en esta óptica validar sirve para *defender* interpretaciones y usos. El segundo uso implica una valoración más o menos objetiva de la evidencia. (2006: 17)

Para Kane "validar una interpretación o uso de los puntajes de una prueba es evaluar la plausibilidad de las afirmaciones que se harán a partir de esos puntajes. Por lo tanto, la validación requiere una clara declaración de los propósitos para los que se emplearán las interpretaciones y usos de los resultados". (2013: 1)

El *enfoque de validación basado en argumentos* (*Argument-based approach to validity, ABAV*) que planteó Cronbach (1988) y desarrollaron Messick (1989) y Kane (2001 y 2006), provee un marco para evaluar los usos de las puntuaciones de una prueba. La idea central es establecer explícitamente, mediante argumentos, las interpretaciones que se pretende dar a los resultados y los usos que se piensa hacer de ellos, y luego evaluar la plausibilidad de esos propósitos.

El proceso que propone el ABAV es simple: primero se establecen las afirmaciones que se harán a partir de las interpretaciones y usos —*argumento interpretativo*— y después se evalúan dichas afirmaciones a partir de las evidencias que justifican o no su uso para cierto propósito: *argumento de validez*.

Los argumentos de interpretación/uso (*Interpretation/Use Arguments,* IUA) de este enfoque de validación tienen el propósito de explicitar y especificar la manera en que se interpretarán y utilizarán los resultados de una prueba con una población.

Según Kane (2013) el ABAV se basa en ocho ideas: 1) Lo que se valida no es la prueba misma o sus puntajes, sino la interpretación y el uso que se haga de ellos. 2) La validez de una interpretación o uso de resultados depende de lo bien que la evidencia apoya las afirmaciones hechas. 3) Afirmaciones más ambiciosas requieren de mayor evidencia que las soporten que las afirmaciones menos ambiciosas. 4) Afirmaciones más ambiciosas —*v. gr.* interpretación de constructos— suelen ser más útiles que las menos ambiciosas, pero son más difíciles de validar. 5) Las interpretaciones y usos pueden cambiar con el tiempo, en respuesta a necesidades y comprensiones nuevas, lo que conduce a cambios en las evidencias necesarias para la validación. 6) La evaluación del uso de los puntajes requiere una evaluación de las consecuencias de los usos planeados, y consecuencias negativas se pueden traducir en un uso inaceptable de las puntuaciones. 7) El rechazo en el uso de un puntaje no necesariamente invalida una interpretación de la puntuación subyacente. 8) La validación de la interpretación de un puntaje no valida el uso que se haga de los resultados.

Las conceptualizaciones recientes sobre la validez incluyen las consecuencias sociales e individuales —deseadas y no previstas— que trae consigo el uso de una prueba (Kane, 2013; Moss, 2008; Sireci, 2013). Esta dimensión atendió una preocupación razonable, pero supuso una complejidad que para algunos hizo más confuso el panorama. Se denominó *validez de consecuencias*, y apareció en los estándares AERA-APA-NCME de 1999, dando lugar a fuertes discusiones.

A partir de la noción de validez de consecuencias, otras posturas advierten el cambio derivado de pasar del ámbito psicométrico al terreno de las políticas. Un resultado de ello es que la valoración global de un programa de evaluación educativa no puede limitarse a lo técnico, sino que debe incluir lo relativo a las consecuencias de las evaluaciones.

Una dimensión más de la noción es la que denota la expresión *validez cultural*, definida como el grado en que el diseño, el proceso de desarrollo y el contenido de una prueba toman en cuenta la forma en que factores de naturaleza cultural, lingüística y socioeconómica, distintos de los constructos de interés, influyen en la manera en que se interpreta el contenido de los ítems y la forma en que se responden. (*cfr*. Basterra, Trumbull y Solano-Flores, 2011)

A principios del siglo XXI tienen lugar fuertes discusiones entre los estudiosos del tema que mantienen posturas diferentes, e incluso encontradas. El volumen editado por

R. Lissitz (2009) presenta un abanico de tales posturas. Algunas variantes del concepto se desarrollan en relación con usos particulares del mismo. Paul Newton (2013) llega a enumerar 149 acepciones del término validez, que incluyen desde los grandes tipos clásicos hasta variantes mínimas de ellos, junto con muchos casos de sentido muy particular. En relación con lo anterior, cuestionamientos radicales proponen abandonar el concepto por considerar que la amplitud que ha alcanzado es incompatible con un mínimo de coherencia y rigor. (Newton, 2013; Michell, 2000)

Relación entre confiabilidad y validez

Se acepta generalmente que puede haber confiabilidad sin validez, pero no al contrario: la ausencia de confiabilidad impide que haya validez. Para comprender esta idea conviene remitirse a la definición más sencilla de validez, que dice que esta consiste en medir realmente lo que se quiere.

Puede parecer ilógico que alguien pueda medir algo que no quiere, pero si se reflexiona sobre la complejidad de muchas variables que se estudian en ciencias humanas, así como en su carácter no evidente sino latente (*constructo*), se podrá estar de acuerdo en que las definiciones operacionales de esas variables, y los indicadores en que se concretan, no siempre reflejan adecuadamente la realidad subyacente, por lo que la información que se podrá obtener con un instrumento desarrollado a partir de tales operacionalizaciones medirá en realidad algo distinto de lo que el investigador pretendía medir. Esa medición podrá ser consistente, o sea que podrá tener confiabilidad, pero carecerá de validez.

Ahora bien: la falta de confiabilidad de una medición indica que la proporción de error o de ruido en la información obtenida es demasiado grande. La ausencia de confiabilidad indica que no se está midiendo en realidad ninguna variable, ni la que se pretendía ni otra, ya que los resultados se deben al azar tanto o más que a cualquier factor determinado. Por ello se considera que la falta de confiabilidad implica también ausencia de validez. Una buena validación no podrá considerarse suficiente si no incluye un sólido análisis de la confiabilidad.

En sentido opuesto Moss (1994) señala que mediciones complejas, como las que usan, por ejemplo, evaluaciones de desempeño o basadas en portafolios, pueden ser menos confiables que, por ejemplo, un instrumento estandarizado, pero pueden captar mejor diversos aspectos de constructos complejos, o sea que pueden tener mayor validez. En estos casos, sin embargo, se trata de dos mediciones diferentes, y parece razonable aceptar una que pueda ofrecer mayor validez, con menor confiabilidad que

otra que dé información más confiable con menor validez. Tratándose de una misma medición, sin embargo, la idea de que la ausencia de confiabilidad implica también la de validez no parece cuestionable.

La calidad de la información, resultado de todo el proceso

Cuidar la confiabilidad y la validez no es simplemente verificar que el valor de tal coeficiente se sitúe dentro del rango considerado aceptable; implica mucho más.

En cuanto a confiabilidad no se puede reducir al cálculo de un coeficiente como el alfa de Cronbach, sin tener claro lo que significa y sin complementarlo con el análisis del error estándar de medición. Es preferible un acercamiento que considere las diferentes fuentes de error con la Teoría de la Generalizabiliadd, y la confiabilidad como propiedad no del instrumento, sino de todo el proceso de medición, lo que implica que el cuidado pleno de esta cualidad básica requiere estudios complejos.

En cuanto a validez, AERA-APA-NCME (1999: 9-24; 2014: 13-22) proponen obtener evidencia de cinco fuentes, que se concretan en 24 estándares que todo instrumento de medición debería satisfacer. Las cinco fuentes de evidencia son: de contenido, de procesos de respuesta, de estructura interna, de relaciones de los puntajes del instrumento con otras variables y de consecuencias. Esas cinco fuentes no son distintos tipos de validez, sino que esta es un concepto unitario, pero su fundamentación requiere obtener evidencias de múltiples fuentes, lo cual implica que un estudio de validación sólido, aunque no exhaustivo, supone el trabajo de equipos de investigación durante lapsos de tiempo prolongados.

Asegurar la calidad de la información que se obtiene con ciertos instrumentos, incluyendo su consistencia y precisión (confiabilidad) y el que soporte sólidamente las interpretaciones que se quieran hacer para conseguir ciertos propósitos en un contexto determinado (validez) implican, pues, investigaciones que consideren el proceso de medición en todas sus etapas, como Cronbach señalaba:

> La sencilla expresión "validación de una prueba" parece implicar que el puntaje que uno interpreta es producido por un instrumento desnudo. En realidad, el instrumento es solo un elemento de un procedimiento, y un estudio de validación debe examinar el procedimiento como un todo. Cada aspecto del contexto en que se aplica una prueba, y cada detalle del procedimiento para hacerlo puede tener influencia en el desempeño, y por lo tanto en lo que se mide. (1971: 449)

En el mismo sentido Crooks, Kane y Cohen proponen un enfoque "*paso por paso*":

> La validez es la cualidad más importante de una evaluación, pero es frecuente que se descuide su valoración. El enfoque paso-por-paso que se sugiere ofrece una guía estructurada a quienes deban validar evaluaciones. El proceso de evaluación es como una cadena de ocho etapas eslabonadas entre sí: administración de la prueba, calificación, agregación de resultados, generalización, extrapolación, juicios de valor, decisiones e impacto. Valorar la validez del conjunto implica considerar con cuidado las amenazas a la validez asociadas a cada eslabón [...] El modelo de la cadena sugiere que la validez del conjunto se ve limitada por el eslabón más débil, y que los esfuerzos por hacer particularmente fuertes sólo algunos eslabones pueden ser estériles e incluso dañinos [...] (1996)

En la tabla siguiente las etapas de un proceso de medición se sistematizan en cinco, cada una de las cuales comprende tres pasos, incluyendo en las etapas finales los que mencionan Crooks, Kane y Cohen, añadiendo otros y distinguiendo dos variantes: una que se aplica en particular a una prueba de aprendizaje, y la otra a cualquier instrumento de obtención de información. La manera como se formulan los pasos hace referencia a cuestionarios escritos de aplicación colectiva, pero es fácil hacer la transferencia a cualquier otro instrumento.

TABLA 3.6. ETAPAS Y PASOS DEL PROCESO DE MEDICIÓN

Etapas	Pasos particulares	
	Prueba de aprendizaje	**Instrumento de medición**
Planeación de la medición	Precisión del propósito(s)	Precisión del propósito(s)
	Definición de población objetivo y, en su caso, muestra	Definición de población objetivo y, en su caso, muestra
	Decisiones técnicas: tipo de prueba, modelo psicométrico...	Decisiones técnicas: tipo de instrumento...
Diseño del instrumento	Definición de los dominios a evaluar con la prueba	Identificación de las variables a medir con el instrumento
	Especificación de dominios	Operacionalización de variables
	Diseño de las pruebas: especificación de ítems, ítems, escalas, niveles de desempeño, juicios por comités expertos, aplicación piloto...	Diseño de instrumento: ítems, conductas a observar, rasgos a analizar en ciertos materiales... Jueceo(s) y piloteo(s)

Recolección de la información	Reproducción de las pruebas	Reproducción del instrumento
	Preparación de aplicación: capacitación, logística...	Preparación de aplicación: capacitación, logística...
	Aplicación	Aplicación
Procesamiento de la información	Calificación de respuestas	Captura de respuestas
	Agregación de resultados	Agregación de resultados
	Generalización, extrapolación	Generalización, extrapolación
Usos de los resultados	Juicios de valor	Juicios de valor
	Decisiones	Decisiones
	Impacto	Impacto

FUENTE: ELABORACIÓN PROPIA.

La noción del eslabón más débil que apuntan Crooks, Kane y Cohen es importante y se puede ilustrar con la tabla anterior: aunque todos los pasos de un proceso sean perfectos, basta con que uno sea muy deficiente para que el resultado se vea comprometido. Por bien diseñados que estén una prueba o un cuestionario, errores graves en su impresión pueden afectar seriamente la calidad de la información que se obtenga. Lo mismo puede decirse si los aplicadores no están bien capacitados, si el aparato que se use para leer respuestas está mal calibrado, si se usa una clave de respuestas equivocada, o si se cometen errores importantes al analizar los datos.

En otro nivel, se puede llegar a juicios de valor injustos con información correcta, y se pueden tomar decisiones desafortunadas y producir daños considerables. Por ello el cuidado de la confiabilidad y la validez debe estar presente en los diversos momentos de la construcción y la prueba del instrumento, pero no sólo en ellos. Debe incluir aplicación, procesamiento de la información y uso de los resultados.

En estudios de grandes dimensiones, como los que incluyen aplicar pruebas de aprendizaje en gran escala, la validación debe incluir estudios de gran profundidad que implican tiempos y recursos considerables, dada la importancia del impacto potencial de los resultados. En proyectos de investigación modestos una validación razonable es igualmente importante, pero los recursos disponibles obligan a adoptar un enfoque sólido pero accesible, de manera pragmática.

La forma convencional de conseguirlo es cuidando la calidad de los instrumentos mediante procesos de verificación por conocedores (jueceos) y de aplicaciones de prueba con sujetos de características similares a las de la población objetivo (piloteos). Para que los procesos de jueceo y piloteo se puedan organizar en forma adecuada, es necesario precisar antes cómo servirán para asegurar que se obtenga información de

la mejor calidad con los instrumentos en desarrollo, o sea para que la información que se obtenga sea confiable y sustente inferencias válidas.

El diseño de un instrumento comienza con la definición del concepto a observar, que por lo general será complejo, multidimensional, y no podrá ser captado en forma directa e inmediata (se tratará de un *constructo latente*). Por ello el siguiente paso consistirá en operacionalizar o especificar el constructo, identificando las dimensiones que lo componen, en su caso subdimensiones, e indicadores de cada dimensión y subdimensión. Esa operacionalización será el punto de partida para diseñar una primera versión del instrumento, formulando preguntas (ítems) o precisando conductas o rasgos observables, con lo que se podrá elaborar un cuestionario, escala o guía de entrevista, un protocolo de observación o una guía para el análisis de materiales, entre otras posibilidades.

El siguiente paso será el *jueceo*, presentando la primera versión a la consideración de conocedores, para que señalen deficiencias y hagan sugerencias de mejora. Luego, con igual propósito, vendrá el *piloteo*, aplicando la versión mejorada a personas similares a las que serán estudiadas. Tanto la revisión por jueces como la aplicación piloto deben cuidar varios puntos:

Los jueces (10-12, combinando expertos e *insiders*) deben tener claro qué pretende medir el instrumento que se les presenta, para lo que no basta que reciban un ejemplar del mismo, sino que necesitan un documento que presente los constructos objeto de estudio, sus dimensiones y la definición operacional que se traduce en ciertos indicadores, así como los ítems del instrumento que corresponden a cada dimensión e indicador. Así los jueces podrán opinar sobre la probable validez de la información que se obtendrá con el instrumento en cuestión, al valorar si los ítems corresponden a indicadores y dimensiones resultantes de la operacionalización. Las opiniones de los jueces se referirán también a la confiabilidad, al valorar la claridad de la redacción de cada ítem, su posible carácter amenazante, el riesgo de generar respuestas estereotipadas o socialmente deseables, etc., ya que ese tipo de deficiencias producirán error aleatorio o ruido, al tener efectos diferenciados, e imposibles de predecir, en los sujetos. Los jueces expresarán sus opiniones con anotaciones al margen del instrumento o con cuestionarios de segundo nivel, pero además deberán discutir sus opiniones (*debriefing*).

En cuanto a la aplicación piloto, los sujetos con los que se haga no deberán ser una muestra estadísticamente representativa de cierta población, porque no se pretende tener resultados generalizables; el número y tipo de sujetos será el necesario para los análisis que se harán, empleando coeficientes de confiabilidad apropiados y, de

ser posible, utilizando la teoría de la generalizabilidad para explorar la importancia de distintas fuentes de error. Convendrá incluir análisis factoriales para verificar si los datos empíricos coinciden con las dimensiones definidas teóricamente. El piloto deberá hacerse en condiciones similares a las que tendrá la aplicación definitiva, y se deberá revisar también la calidad de las preguntas, la adecuación de las opciones de respuesta, el tiempo que lleva responder y la claridad de las instrucciones. Se podrán hacer entrevistas o aplicar cuestionarios *ad hoc* sobre la prueba.

Un buen piloteo es fundamental para que la aplicación se haga correctamente, lo cual es tan importante para la calidad de la información que se obtendrá, como el instrumento con el que se recabe.

Para entender lo que implica la aplicación para la confiabilidad, y recordando lo que dice Brennan sobre la consistencia de resultados en réplicas, se puede pensar lo que pasa si una misma prueba se aplica a dos grupos de alumnos en condiciones distintas, de tiempo, ruido y temperatura: en un grupo los alumnos tienen dos horas para responder, en condiciones de poco ruido y temperatura ambiente moderada; en el otro grupo se da sólo una hora y media para responder, hay mucho ruido y hace mucho calor. Obviamente los resultados variarán de forma tal que no se puede afirmar que un alumno que respondió bien más preguntas en el primer caso sabe realmente más que otro con menos aciertos en el otro grupo.

La comparabilidad de resultados de una prueba con versiones equivalentes que se aplican anualmente es un caso de réplicas cuya confiabilidad debe analizarse. Si las condiciones de aplicación cambian a lo largo del tiempo los resultados no serán comparables. Puede haber diversos cambios: el más obvio será, desde luego, el que los instrumentos no tengan el mismo grado de dificultad, como resultado de fallas de equiparación. Otros problemas tienen que ver, por ejemplo, con el momento del ciclo escolar en que se haga la aplicación, el grado de familiaridad de los alumnos o su motivación, o el que haya preparación para la prueba, todo lo cual puede afectar seriamente la comparabilidad de los resultados.

En cuanto a validez, si esta consiste en que se mida efectivamente lo que se quiere medir, pensemos en una prueba que pretende medir los conocimientos de unos alumnos sobre ciertos temas; si un alumno recibe ayuda de otra persona (otro alumno, el maestro, o quien sea), entonces los resultados no medirán lo que ese alumno sabe, sino lo que sabe la persona que le ayudó.

Para que las condiciones de la aplicación de un instrumento no afecten la calidad de los resultados es necesario *estandarizarlas*. La noción de *prueba estandarizada* se

entiende muchas veces incorrectamente, como si se refiriera solo a que las pruebas tengan las mismas preguntas. El concepto correcto es distinto: dos pruebas podrán considerarse estandarizadas si son equivalentes, aunque no contengan las mismas preguntas, pero cuyas condiciones de aplicación sean *uniformes*.

Estandarizar una aplicación puede ser difícil por la diversidad de circunstancias y la cantidad de personas involucradas cuando se trata de aplicaciones grandes, con experiencia desigual, que reciben capacitación limitada y no siempre dominan todos los detalles del proceso. Por ello es importante preparar manuales detallados de procedimientos, que definan en forma clara las tareas a realizar y las funciones de cada actor; tener procesos de capacitación, que consigan habilitar a los aplicadores para que consigan la buscada uniformidad; definir estándares de calidad precisos para cada paso de la aplicación; y tener instancias de control interno y externo que verifiquen el cumplimiento de los estándares.

Los investigadores educativos y de las ciencias sociales suelen estar familiarizados con las nociones de confiabilidad y validez. En muchos casos, sin embargo, la comprensión que tienen al respecto es limitada, reduciéndose a versiones de hace más de medio siglo, con desconocimiento de los desarrollos recientes, así como de herramientas estadísticas desarrolladas hace décadas. Los paquetes estadísticos computarizados ponen hoy al alcance de toda persona herramientas que hace no mucho estaban reservadas a los centros de investigación más avanzados, pero eso no sustituye una cabal comprensión del sentido de esas herramientas.

Como dice Brennan en sus comentarios conclusivos del artículo sobre *la historia y el futuro de la confiabilidad desde la perspectiva de las réplicas*, antes citado:

> Las matemáticas y la estadística ofrecen poderosas herramientas para examinar los aspectos sintácticos de las cuestiones de medición, en particular de la confiabilidad, pero por sí mismas no pueden responder las preguntas semánticas al respecto [...] Tengo la impresión de que frecuentemente pasamos por alto la importancia de estas cuestiones en nuestro discurso, nuestras publicaciones y nuestros manuales técnicos. A medida que en el futuro se desarrollen nuevos tipos de procedimientos de medición —y estoy seguro de que así ocurrirá— no dudo de que los investigadores encontrarán formas estadísticas creativas de estimar la confiabilidad. Estoy más preocupado por nuestra tendencia a pasar por alto nociones conceptuales fundamentales, sobre la confiabilidad como réplicas. Me preocupa tener una estimación de algo sin una idea clara de lo que es ese algo. (2001: 313)

Conclusión

Los términos *cuantitativo y cualitativo* para clasificar un procedimiento de obtención de información son inadecuados, pero apuntan a diferencias reales entre dos tipos de técnicas, que se trata de explicitar y sistematizar mediante la tabla siguiente.

TABLA 3.7. CARACTERÍSTICAS DE LAS TÉCNICAS CUANTITATIVAS Y CUALITATIVAS

Características	Tipos de técnicas	
	Llamadas cuantitativas	Llamadas cualitativas
Del objeto y su acotamiento		
Amplio/reducido	Acercamiento extensivo enfoque macro: muchos casos	Acercamiento intensivo, enfoque micro, pocos casos
Simple/complejo	Pocas variables, analítico	Muchas variables, holístico
Del procedimiento		
Estructuración	Inicial: operacionalización fina *a priori*, lógica deductiva	Final: categorías *a posteriori,* lógica inductiva
Inferencia	Baja	Alta

FUENTE: ELABORACIÓN PROPIA.

Las dimensiones principales, en los renglones del cuadro, y sus subdimensiones, se relacionan de manera que unas opciones implican o descartan otras:

- Un acercamiento extensivo, macro, que abarque muchos casos, difícilmente podrá simultáneamente considerar muchos aspectos de cada uno, muchas variables, como un acercamiento intensivo, micro.
- Mayor estructuración inicial se asocia lógicamente a baja inferencia, e implica una operacionalización de tipo deductivo, mientras los enfoques con estructuración final, que incluyen la construcción a *posteriori* de categorías, con lógica inductiva, se asocia a procedimientos de alta inferencia.

Los tipos ideales que se construyen con estas dimensiones y subdimensiones son prototipos de las técnicas llamadas *cuantitativas y cualitativas.* Se puede apreciar que los procedimientos de medición a niveles superiores al nominal, que suponen cuidadosos procesos previos de operacionalización, se asocian con la familia de técnicas *cuantitativas*, lo que explica el uso de esa etiqueta para designarlas.

El esquema muestra que puede haber combinaciones de características menos convencionales que las de los tipos ideales; muestra también que no se puede hablar de la superioridad o inferioridad de una combinación u otra: lo que hay es mayor o menor adecuación de ciertas combinaciones de características a las condiciones de ciertos objetos de estudio. Como no puede decirse sin más si un teleobjetivo es mejor que un gran angular o viceversa, sino que depende de lo que quiera destacar el fotógrafo, así también no existe técnica alguna que tenga sólo ventajas; si se gana en detalle se pierde en perspectiva y viceversa.

Se puede llegar también a la conclusión de que no hay una diferencia fundamental entre los acercamientos etiquetados como cuantitativos y los llamados cualitativos a partir de una consideración sobre los criterios que utilizan quienes trabajan en una u otra de estas tradiciones para juzgar la calidad de sus hallazgos.

¿Con qué criterios juzgar la calidad de un estudio? Un acercamiento simplista *cuanti* tenderá a aplicar criterios técnicos: una investigación será mejor si usa herramientas estadísticas sofisticadas: una con técnicas multivariadas será mejor que otra limitada a correlaciones, y esta superior a una que se quede en procedimientos descriptivos univariados; una que ni siquiera maneje medidas de tendencia central y dispersión será inaceptable. En trabajos simplistas *cuali* la calidad podrá depender de la riqueza de la descripción etnográfica, o de la profundidad del compromiso.

Como hemos visto, la tradición cuantitativa, en especial experimental, presta mucha atención al tema, con criterios de calidad como o*bjetividad, confiabilidad, o validez interna y externa.* En la tradición cualitativa, Lincoln y Guba han propuesto una extensión de esos criterios para valorar la calidad de un trabajo, su *trustworthiness*; estos autores sugieren términos paralelos a los mencionados; Miles y Huberman proponen otros similares en la forma que resume esta tabla:

TABLA 8. TÉRMINOS CONVENCIONALES Y ALTERNATIVOS DE CRITERIOS DE CALIDAD

Aspecto	Término convencional	Guba-Lincoln	Miles-Huberman
Valor de verdad	Validez interna	Credibilidad	Autenticidad
Aplicabilidad	Validez externa Generalización	Transferibilidad	*Fittigness*
Consistencia	Confiabilidad	Dependencia	*Auditability*
Neutralidad	Objetividad	Confirmabilidad	Confirmabilidad

FUENTE: GUBA-LINCOLN, 1985; MILES-HUBERMAN, 1984.

Cada tradición ha desarrollado criterios particulares, de naturaleza procedimental:

- En trabajos cuantitativos la confiabilidad se cuida controlando la consistencia de los resultados de mediciones sucesivas, los de partes del instrumento contra los de otras, o los juicios de un evaluador y los demás. Se atienden las dimensiones o fuentes de validez mediante jueceos o correlaciones con medidas criterio y se busca eliminar la influencia de factores que la amenacen, como los efectos de la historia, la maduración de los sujetos, la mortalidad de la muestra, los de la regresión estadística, los defectos de implementación o los efectos de la situación de laboratorio sobre investigadores y sujetos experimentales. Se procura detectar el carácter espurio de una correlación, por la intervención de otras variables, antes de proponer una interpretación en términos causales.
- En la tradición cualitativa no se da por bueno cualquier trabajo: hay criterios distintos pero exigentes, como el uso de triangulaciones, el tiempo de inmersión en el terreno y de la observación, la densidad de la descripción, el análisis de casos negativos, contrastantes o extremos, la discusión de resultados con otros investigadores y demás participantes, la verificación del sustento que tiene en los registros lo que aparece en el reporte, y la auditoría de investigación.

No olvidar que cualquier criterio será aplicado por personas concretas, por lo que la calidad dependerá, en última instancia, de los controles cruzados de que habla Polanyi, en comunidades de investigadores más o menos consolidadas, alrededor de un paradigma (Kuhn) o programa de investigación, a la manera de Lakatos.

Apéndice. Ejemplos de protocolos de observación

Descripciones narrativas (Stallings, 1977)

Con frecuencia yo deseaba tener un recuento confiable de lo que pasaba con un niño(a) con problemas para relacionarse y aprender, porque era difícil entender lo que experimentaba un niño mientras había otros 35 que necesitaban atención. ¿A quién se acercaba y hablaba el niño en cuestión? ¿Quién se acercaba a ese niño? ¿Cómo era su interacción? ¿Cómo le respondía yo, como maestra? ¿En qué actividades participaba más? ¿Cuáles rechazaba? ¿Cuánto tiempo duraba en una actividad? ¿Qué tanto cambiaba de lugar? ¿Qué tan frecuentemente estaba contento o enojado, absorto o desganado? Las respuestas a este tipo de preguntas son valiosas para planificar el programa educativo de cada niño.

Esas respuestas pueden derivarse de descripciones escritas sistemáticas de los niños, hechas por el maestro o por un observador externo. Yo desarrollé descripciones escritas de este tipo de un alumno llamado Frank, a lo largo de siete meses, durante períodos de 20 minutos y al menos dos veces por semana. Mientras yo observaba se encargaba de la clase un maestro auxiliar. Yo registraba también eventos más cortos en los que Frank estaba involucrado, en cualquier momento que ocurriera, para lo cual siempre traía en mi bolsa una pequeña libreta para anotar la fecha y hora del evento o interacción que describía. Cada anotación tenía, por lo general, sólo unas pocas frases, pero esa información era importante para formar una imagen más completa de la vida de Frank en la escuela, y para preparar un programa educativo para él.

En una ocasión la autora contrató a un estudiante universitario para observara a un alumno difícil de segundo grado, y con las descripciones que obtuvo:

[...] pedí una reunión con los padres del alumno, su médico, un especialista en lectura y el psicólogo de la escuela. La narración escrita de su conducta me permitió presentar información factual con un mínimo de inferencias. Como resultado de la reunión se diseñó un programa educativo que ayudó al alumno en su aprendizaje, que incluyó que un especialista en lectura trabajara con él tres veces por semana, fuera del aula, al tiempo que dentro del aula yo seguía el programa de lectura que indicó el especialista. El pediatra prescribió una dieta sin alimentos con aditivos artificiales para reducir la hiperactividad. El psicólogo escolar recomendó a los padres que fueran pacientes y ejercieran en casa la mayor influencia tranquilizadora que pudieran, y que tuvieran expectativas realistas en cuanto al progreso del niño hacia una conducta más controlada [...] se decidió filtrar las influencias distractoras dentro de la escuela. Como el aula era grande, decidimos hacer tres cabinas de aprendizaje [...] con el espacio justo para un pequeño escritorio, y con la entrada cerrada con una cortina de tela no inflamable. Dijimos al grupo que las cabinas eran lugares especiales para poder aprender mejor, para que el niño pudiera usarlas sin dar la impresión de que se le marginaba o castigaba. De hecho, las cabinas se volvieron tan populares con los niños que para ocuparlas hubo que organizar una lista de espera.

Lista de cotejo (Observed Learning Avoidance Behaviors, Stallings, 1977)

Mi segundo intento para observar fue más sistemático. Me interesaba ver cómo trataban de evitar las actividades de aprendizaje los niños de primer grado durante los períodos de instrucción formal de lectura en pequeños grupos. Hice una lista de las conductas que

yo había advertido que ocurrían con más frecuencia durante los tiempos dedicados a la lectura, en comparación con otros períodos de la jornada. Con esa lista desarrollé una forma que podía emplear para observar a cada niño. Con ella podría observar la conducta de un niño por períodos de cinco minutos y registrar las observaciones junto con la fecha y hora en que habían ocurrido.

Cada vez que un niño se chupaba un dedo, por ejemplo, hacía una marca en la columna correspondiente a la observación y en el renglón de "Chuparse el dedo". La forma permitía registrar observaciones durante varios días, y las conductas para evitar actividades de aprendizaje de un niño se podían comparar a lo largo del tiempo, comparación importante porque una presencia alta de cierta conducta en un día puede no ser representativa, pues el niño podía estar especialmente nervioso por ser su cumpleaños o porque debía ir al dentista.

Los datos se analizaban simplemente contando cuantas veces había ocurrido un comportamiento en los cinco minutos observados. Una tasa elevada de conductas para evitar actividades de aprendizaje en tres días diferentes podía sugerir que el niño se sentía muy inseguro en cuanto a su capacidad para leer, o que el material era inapropiado para su estilo de aprendizaje o sus intereses, o que tenía un problema físico [...]

Yo podía identificar fácilmente a los niños que realmente evitaban actividades de aprendizaje comparando las observaciones de varios alumnos, pero eso era sólo el principio. Había que entender la causa de tales comportamientos para poder emprender acciones correctivas. Primero verificaba los registros de salud de los alumnos para ver si había reportes de problemas de agudeza visual o auditiva [...] En segundo lugar, trataba de detectar problemas de percepción, para lo que aplicaba a cada alumno de primer grado de mi grupo una sencilla prueba diagnóstica sobre secuencias [...]

Los niños con problemas de secuencias visuales parecían avanzar más si la enseñanza de la lectura comenzaba con un enfoque global [whole word]. Las habilidades fonéticas se deben introducir y usar para mejorar las habilidades de secuencias de un alumno, pero un niño que tiene problemas para secuenciar porciones pequeñas de información puede tener un bajo rendimiento si se le obliga a comenzar a leer con un enfoque fonético.

Hay que advertir que es importante, desde luego, observar y evaluar a los niños al inicio del año escolar, de manera que sea posible seleccionar bien los métodos de instrucción apropiados. Los niños no deberían experimentar varios meses de fracaso antes de recibir la atención que necesiten. (Stallings, 1977: 11-13)

El formato usado por Stallings para registrar lo que hace un niño en un lapso de cinco minutos y tres días diferentes, en cuanto a 19 conductas, es el siguiente.

LISTA DE COTEJO SOBRE CONDUCTAS PARA EVITAR ACTIVIDADES DE APRENDIZAJE

Alumno: _____	**Fecha _____ Hora _____**	**Fecha _____ Hora _____**	**Fecha _____ Hora _____**
Torcer o apretar las manos			
Torcer pies o piernas entre patas de la silla			
Golpetear con los pies o las manos			
Golpetear con un lápiz u otro objeto			
Balancear el cuerpo o la cabeza			
Inclinar la silla o el escritorio			
Caerse de la silla			
Levantarse de la silla para algo irrelevante como sacar punta a un lápiz que no lo necesita Distraer a los vecinos			
Hacer ruidos sin sentido, como zumbidos, canturreos, hablar solo			
Distraerse de la tarea, soñar despierto			
Hacer gestos, entrecerrar los ojos o parpadear			
Bostezar, toser o suspirar			
Escarbarse la nariz			
Chuparse un dedo			
Rascarse			
Jugar con el cabello, enredarlo			
Taparse la cara con las manos o el cabello			
Quejarse de dolor de cabeza, estómago, etc.			

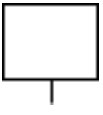
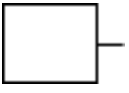
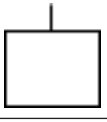
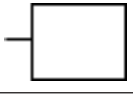
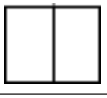
FUENTE: STALLINGS, 1977: 12.

Mapa de asientos (Seating chart) de Puckett (Medley y Mitzel, 1963: 254)

Se reconstruye el formato como sigue:

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48

Clave para la codificación de las conductas:

•	El alumno levanta la mano		
	Idem y el profesor le pregunta		El profesor pregunta a un alumno que no levantó la mano
	Idem y el alumno responde con monosílabo		Idem y el alumno responde con monosílabo
	Idem y el alumno responde en forma aceptable		Idem y el alumno responde en forma aceptable
	Idem y el alumno responde en forma buena		Idem y el alumno responde en forma buena
	Idem y el alumno responde en forma muy buena		Idem y el alumno responde en forma muy buena
>	El alumno pregunta		Idem y el alumno no responde
|	El alumno habla sin que el profesor se dirija a él		

FUENTE: CON BASE EN MEDLEY Y MITZEL, 1963: 254.

Mapa de asientos (Seating chart) de Wrightstone (Medley y Mitzel, 1963: 254)
El cuadro de registro es similar al de Puckett en lo que se refiere a las casillas para registrar conductas individuales de alumnos, añadiendo en la parte inferior un espacio amplio para registrar acciones del maestro dirigidas a todo el grupo.

Claves para codificar conductas del maestro:

A	Permite que un alumno intervenga voluntariamente	**F**	Desalienta o prohíbe la intervención de un alumno
B	Alienta a un alumno a que intervenga	**G**	Llama la atención a un alumno verbalmente o gestualmente
C	Plantea una pregunta o tema a un alumno o al grupo	**H**	Asigna una tarea específica a un alumno o al grupo
D	Indica a uno o varios alumnos que consulten algún material	**I**	Hace preguntas o responde las de uno o más alumnos sobre tareas o sobre el contenido de los libros de texto u otros materiales
E	Sugiere o explica soluciones, actividades o formas de trabajo		

FUENTE: CON BASE EN MEDLEY Y MITZEL, 1963: 255.

Mapa del Tiempo (Time Chart) de Barr (1929)
Se registra la duración de conductas de maestro y alumnos, durante 30 minutos de una clase. Cada renglón representa un minuto; las columnas distinguen períodos de 10 segundos. Las conductas a observar y las claves para registrarlas son:

T: Actividades realizadas por el profesor (Teacher's activities)
Tc: Comentarios del profesor (Teacher's comments)
Tq: Preguntas del profesor (Teacher's questions)
P: Actividades realizadas por uno o varios alumnos (Pupils' activities)
Pc: Comentarios de los alumnos (Pupils' comments)
Pq: Preguntas de los alumnos (Pupils' questions)

Al presentarse por primera vez una conducta se anota su clave en el lugar del *Mapa* que corresponda, y a partir de ese punto se marca una línea horizontal, según el tiempo que dure la misma conducta; cuando se interrumpa se detiene la marcación de la línea y se anota la clave de la conducta que se presente en seguida, continuando con la línea horizontal correspondiente, por toda su duración. Se usará un reloj con segundero.

Minutos	Segundos					
	10	20	30	40	50	60
1						
2						
3						
4						
5						
…						
26						
27						
28						
29						
30						

FUENTE: CON BASE EN MEDLEY Y MITZEL, 1963: 258-259.

Sistema multidimensional de observación (Cornell, Lindvall y Saupe, 1952)

El sistema considera ocho dimensiones; la relativa a Variedad (E) comprende 23 tipos de conductas de maestro o alumnos. Los renglones de la forma se refieren a dichas dimensiones, y las columnas al momento en que ocurren las conductas, considerando períodos de cinco minutos en una hora de observación

Dimensiones y/ conductas	Períodos de cinco minutos de una hora de observación											
	5	10	15	20	25	30	35	40	45	50	55	60
Diferenciación (A)												
Organización social (B)												
Iniciativa alumnos(C)												
Contenido (D)												
Variedad (E)												
1												
2												
3												
4												
5												
6												
7												

8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
Competencia maestro (F)												
Clima (maestro)(G)												
Clima (alumnus)(H)												

Se observa una clase durante una hora, dividida en 12 períodos de cinco minutos. Cada columna contendrá las observaciones correspondientes a esos períodos. Los renglones de la forma corresponden a dimensiones y conductas a observar. En cada cruce de renglón y columna se anota (cada 5') la clave de la tabla que mejor describa lo que pasó en el aula en ese período en lo que se refiere a las dimensiones A, B, C y D. Sobre la dimensión E, se anota una rayita en el renglón que corresponde a cada una de las 23 conductas a observar si esa conducta se presentó en el período de 5' de que se trate. Las tres últimas dimensiones (F, G y H) se observan también, pero marcando sólo una vez las conductas que se presenten a lo largo de la hora completa de observación, independientemente de si se presentan con mucha o poca frecuencia. Las conductas positivas se marcan con su número seguido por el signo + y las negativas con el número que corresponda seguido por el signo -.

Clave para registro de conductas

a. Diferenciación

- Trabajo idéntico, sin ayuda del maestro.

- Trabajo idéntico, con ayuda del maestro.
- Trabajo diferenciado con base en habilidad, pocos grupos, sin ayuda del maestro.
- Trabajo diferenciado con base en habilidad, pocos grupos, con ayuda del maestro.
- Trabajo diferenciado con base en habilidad, individual, sin ayuda del maestro.
- Trabajo diferenciado con base en habilidad, individual, con ayuda del maestro.
- Trabajo diferenciado con base en habilidad e interés, pocos grupos, sin ayuda del maestro.
- Trabajo diferenciado con base en habilidad e interés, pocos grupos, con ayuda del maestro.
- Trabajo diferenciado con base en habilidad e interés, individual, sin ayuda del maestro.
- Trabajo diferenciado con base en habilidad e interés, individual, con ayuda del maestro.

b. Organización social

- Un solo grupo, dirige el maestro, no interacción.
- Varios grupos, dirige el maestro, no interacción.
- Un solo grupo, dirige un alumno, no interacción.
- Varios grupos, dirige un alumno, no interacción.
- Un solo grupo, dirige el maestro, interacción.
- Varios grupos, dirige el maestro, interacción.
- Un solo grupo, dirige un alumno, interacción.
- Varios grupos, dirige un alumno, interacción.
- Iniciativa de los alumnos
- Dominio del maestro, no participación de los alumnos.
- Dominio del maestro, participación menor de los alumnos.
- Control del maestro, participación mayor de los alumnos.
- Control de los alumnos, participación del maestro.
- Control de los alumnos, no participación del maestro.

c. Contenido

- Un libro de texto o cuaderno de trabajo.
- Varios textos o materiales similares.
- Libros de consulta diferentes de los de texto.
- Problemas, unidades o tareas indicadas por el maestro.
- Problemas, unidades o tareas por interés de los alumnos.

d. Variedad

1.	El maestro expone o lee	2.	El maestro hace demostración
3.	El m. pasa video o transparencias	4.	Los a. leen el texto en su lugar
5.	Los a. leen otros libros en su lugar	6.	Los a. trabajan con cuadernos en su lugar
7.	A. resuelven problemas no del texto en su l	8.	A. estudian otros materiales en su lugar
9.	A. dibujan o pintan en su lugar	10.	El m. pregunta, los a. contestan
11.	El grupo discute un tema	12.	Un a. expone o presenta un reporte
13.	Los a. trabajan en el pizarrón	14.	Los a. leen un libro en voz alta
15.	Los a. estudian dibujos, mapas, tablas	16.	Los a. hacen experimentos
17.	Los a. construyen objetos	18.	Los a. decoran el aula
19.	Los a. hacen juegos de roles u otros juegos	20.	Los a. salen de excursión
21.	Los a. trabajan en un aula diferente	22.	Los a. trabajan en pequeños grupos
23.	Los a. responden una prueba		

e. Competencia del maestro

	Positivo. El maestro:	Negativo. El maestro:
1	Sugirió ayudas para el aprendizaje, dio tips	Estuvo inactivo
2	Explicó detalladamente	No consiguió mantener la atención del grupo
3	Dio repasos	Eludió la responsabilidad
4	Aclaró, repitió ideas, mostró relaciones…	Dio explicaciones que confundieron más
5	Puso ejemplos y experiencias	Dio respuestas incompletas o inexactas
6	Dio respuestas completas y satisfactorias	Dio evidencias de escaso domino de temas
7	Dio evidencias de haber planeado y preparado	Evidenció escasa preparación de temas
8	Se mostró confiado, capaz de enfrentar cosas	Pareció inseguro
9	Dio evidencias de conocer la materia	Permitió que la discusión se saliera del tema

f. Clima que propicia el maestro

	Positivo. El maestro:	Negativo. El maestro:
1	Hizo observaciones con cortesía	Simplemente aplicó las normas
2	Mostró respeto por las opiniones	Mostró intolerancia ante sugerencias de a
3	Dio evidencias de paciencia	No dejó hablar a los a
4	Ayudó a los a. en problemas no académicos	Corrigió o criticó excesivamente
5	Expresó simpatía	No mostró simpatía ante fracasos

6	Trató de entender el punto de vista del alumno	Usó amenazas
7	Felicitó a los alumnos	Perdió el control
8	Aceptó bien las críticas	Permitió que los a. se rieran de errores
9	Bromeó con los alumnos	Hizo observaciones sarcásticas, ridiculizó
10	Usó principalmente el "nosotros"	En ocasiones mostró molestia, enojo
11	Prestó atención a todo el grupo	Usó principalmente el "yo"

g. Clima manifestado por alumnos

	Positivo. Los alumnos:	Negativo. Los alumnos:
1	Respondieron gustosamente	Se mostraban inquietos, distraídos
2	Trabajaron intensamente sin distracciones	Eran lentos para responder a demandas del m.
3	Se mostraron dispuestos a participar	Reacios a participar, no se ofrecían para ello
4	Prestaban atención al m. u otros a.	Cuchicheaban u otras señales de no atención
5	Hacían observaciones cortésmente	Hacían observaciones no cortes
6	Aceptaban bien las críticas	Se mostraban pendencieros e irritables

FUENTE: CON BASE EN MEDLEY Y MITZEL, 1963: 275-276.

Sistema para observar interacción en matemáticas (Wright y Proctor, 1961)

Este sistema de observación con muestreo de tiempo busca comparar interacciones verbales en clases de matemáticas. La observación se hace durante 45 minutos, lapso en el que deben hacerse 90 observaciones por sesión, cada una de las cuales lleva medio minuto, dividido en dos partes: 15 segundos para observar y otros 15 para registrar lo observado. Para controlar el tiempo se usa un cronómetro.

Las interacciones observadas se clasifican según tres dimensiones: contenido, proceso y actitud, con las categorías que se sintetizan según la tabla siguiente:

Dimensiones	Categorías
A. Marco para la dimensión de contenidos	Aspectos fundamentales: 1. Estructura 2. Técnicas
	Relaciones: 3. Deductivas 4. Inductivas 5. Declaraciones

	Aplicaciones: 6. Problemas matemáticos 7. Otros aspectos
B. Marco para la dimensión de procesos	Procesos silogísticos: 1. Analíticos 2. Sintéticos
	Procesos clasificatorios: 3. Especializados 4. Generalistas
	Aspectos relevantes sin secuencia lógica clara 5. Información relevante
C. Marco para la dimensión de actitudes	Actitudes que maestro muestra o promueve o que los alumnos muestran: 1. Curiosidad 2. Independencia 3. Receptividad

FUENTE: CON BASE EN MEDLEY Y MITZEL, 1963: 288-290.

En cada período de 15 segundos las interacciones verbales se deben clasificar según alguna categoría de las tres dimensiones. Los resultados se analizan considerando el número absoluto de registros de cada categoría y su proporción en el total. Se pueden hacer análisis de las combinaciones de las categorías de las tres dimensiones que se presentan, o bien de ciertas secuencias, *v. gr.* estructura + deducción + inducción + declaración.

Se construyó también un "Índice de iniciativa", basado en un compuesto ponderado de los puntajes de curiosidad, independencia y receptividad. El sistema implica que los observadores tengan un nivel alto tanto en conocimientos matemáticos como en experiencia de observación.

Sistema de Prescott (1973)

Elizabeth Prescott desarrolló este instrumento para observar a niños de tres y cuatro años de edad en guarderías. En este sistema las conductas de un niño se registran cada 15 segundos, cuando se escucha un sonido en un audífono, sin pretender captar interacciones. Los datos que se recogen con este sistema examinan muchas conductas de los niños, como autonomía, dependencia, agresión, participación social, persistencia en una tarea, habilidad para resolver problemas y curiosidad. Estas conductas se definen operacionalmente y los observadores reciben entrenamiento para registrar códigos particulares para cada conducta. La tabla de la página siguiente muestra los códigos y las conductas que se registran con el Sistema de Observación de Prescott.

> Los códigos del instrumento se dividen en cuatro categorías principales. La primera (Rechazo) incluye conductas que el niño o niña manifiesta cuando se defiende. La segunda categoría (Empuje) incluye conductas proactivas o iniciadas por el niño o niña, identificando a los que son introvertidos o extrovertidos. La tercera categoría (Respuesta) incluye la reacción de los niños a invitaciones, peticiones, preguntas o indicaciones y reglas; en esta categoría también se registran invitaciones, peticiones, etc., que haga el niño o niña. La cuarta categoría (Integración) se utiliza para registrar las ocasiones en que un niño o niña integra varias conductas. Este sistema de observación aporta información importante sobre el crecimiento y desarrollo de los niños que están en guarderías. Muchas de las conductas de niños de tres y cuatro años que pueden ser registradas con este sistema de observación no pueden ser medidas con pruebas estandarizadas. (Stallings, 1977: 13)

Sistema sobre necesidades de educación especial (Croll y Moses, 1985)

El objetivo del estudio para el que se desarrolló este sistema fue estudiar las diferencias entre las actividades que llevan a cabo y las interacciones en que participan en el aula los niños con necesidades educativas especiales o sin ellas. Se observaron 34 aulas, cuyos docentes fueron entrevistados para identificar niños con necesidades especiales. Se aplicaron pruebas de lectura y razonamiento no verbal para identificar otros alumnos con necesidades especiales. Se escogieron al azar hasta seis niños con posibles necesidades especiales, y otros seis (cuatro niños y dos niñas) como control. En total en cada aula se debía observar de cuatro a diez alumnos, sin que el maestro supiera quiénes eran. Se observó sucesivamente a cada niño, en orden aleatorio; en total dos horas por niño y 20 horas por aula. Las variables y las categorías utilizadas para la observación fueron las siguientes:

Variable	Categorías
Tres variables relativas al contexto	
1. Organización de la enseñanza	El maestro con toda la clase; con un grupo en que está el niño observado; un grupo en que está el niño, sin m.; el niño observado solo; no hay actividad organizada.
2. Lectura/no lectura	La actividad que hace el niño observado requiere que lea; actividad no requiere que lea.
3. Contenido curricular	Lengua-lectura; l.-escritura; l. otro; matemáticas; otro ni lengua ni matemáticas; actividad no curricular.
Tres variables relativas a las actividades	

4. Actividad del niño observado	Curricular directa; relacionada con currículo; distraído en algo no relacionado con currículo; agresión física o verbal a persona; agresión a objetos; disciplina; administrativa; otra (pero no "distraído").
5. Interacciones	Nula (el niño observado solo); el observado con otro; con un grupo de niños; con el maestro solo; con el m. en un grupo con papel principal; como parte de toda la clase con el m.; como parte de la clase con el m. pero con papel principal; con adulto que no es el m.
6. Mobilidad-inquietud	El niño observado está moviéndose de lugar; el niño está en su lugar, pero inquieto; ninguna de las dos.

FUENTE: CROLL Y MOSES, 1985.

La observación se hizo en intervalos de 10 segundos, debiendo registrarse lo que hacía en ese lapso el niño observado. Para el registro se utilizó un formato con 25 renglones (A—Y). El primer renglón A (se llena antes) tiene espacios para marcar una clave de dos dígitos para el maestro; una de un dígito para el alumno al que se refiere el formato de observación; y una de dos dígitos, según lista especial, para identificar la sesión a que se refiere el formato. Los primeros espacios del renglón (B) se llenan al inicio de un período de observación con el dígito que corresponda de las variables de contexto, que permanecen constantes. Los últimos espacios del renglón B y los siguientes renglones se llenan con el dígito que corresponda de las variables de actividades. En la parte derecha están categorías de las seis variables a codificar con sus respectivos dígitos. Un formato permite hacer 24 observaciones de 10" de un niño, para un total de 240" = 4'. Luego se llena otro formato para el segundo niño, y así sucesivamente. En una clase de 40' se puede observar a 10 niños. Observando 20 horas por aula, 2 h. por niño = 7,200" = 720 registros de 10"

Sistema de Flanders (Flanders Interaction Analysis Category System, FLACS)
Stallings dice respecto a este sistema, que inspiró los desarrollados por ella:

> [...] es útil en especial para valorar discusiones en grupo dirigidas por el docente, ya que permite apreciar el nivel de participación de los miembros del grupo, así como las estrategias de interrogación y retroalimentación que utiliza el maestro. Las intervenciones del maestro y los alumnos se registran en una matriz y se analizan posteriormente [...] Este tipo de información puede ayudar al docente a ver qué tan frecuentemente acepta, elogia o critica a los alumnos, y se puede utilizar para analizar grabaciones en audio o video de las clases. (1977: 15)

Cada tres segundos el observador anota un número que corresponde a un tipo de conducta de profesor o alumnos. Las anotaciones se hacen siguiendo las columnas del formato, cada una de las cuales tiene 20 renglones, que en total corresponden a un minuto (tres segundos cada renglón). El formato tiene espacio para media hora de observación (30 columnas).

Para observar clases más largas se usan formatos adicionales. Los períodos dedicados a una misma actividad se llaman episodios, como el lapso en que el profesor pasa lista, para luego iniciar una exposición, o un período dedicado a dirigir preguntas a los alumnos o a responder dudas. Cuando el observador identifica el fin de un episodio debe anotar el signo = y explicar qué significa. Flanders recomienda observar a un profesor al menos durante seis clases, y de preferencia durante ocho. Las categorías definidas por Flanders son 10:

Habla el profesor:

En respuesta a conducta de los alumnos:

- Acepta sentimientos
- Elogios o aliento
- Acepta o utiliza ideas de los alumnos

En interacción:

- Formula preguntas
- Iniciando una actividad propia:
  - Expone la lección
  - Da instrucciones a los alumnos
  - Critica a los alumnos o justifica su autoridad

Hablan los alumnos:

En respuesta a acción del profesor:

- Respuesta verbal no espontánea
- Iniciando una actividad propia:
- Intervención verbal espontánea

Silencio:

- Silencio o confusión
- Reglas para la observación

1. Si el observador no está seguro de cuál categoría debe utilizar para registrar lo que ocurre en uno de los lapsos de tres segundos, entre las categorías 1 a 9 (o sea descartando la categoría 10), deberá marcar la categoría más alejada de la 5. Por ejemplo, si duda entre las categorías 2 y 3 deberá marcar 2.
2. El observador deberá abstenerse de involucrar su punto de vista personal.
3. Si una categoría se mantiene más de 3" se debe marcar su código tantas veces como proceda, siempre con intervalos de 3". Si en algún caso ocurren simultáneamente dos conductas (o incluso más) se deben registrar todas.
4. Si un silencio dura más de tres segundos se deberá marcar el código 10 cuantas veces sea necesario.
5. Si el maestro llama a un niño por su nombre se deberá registrar la categoría 4.
6. Si el maestro repite una respuesta de un alumno dando a entender que es correcta se deberá marcar la categoría 2, como un tipo de elogio o aliento.
7. Cuando el maestro escucha lo que dice un alumno y acepta sus ideas para un intercambio se deberá marcar la categoría 3.
8. Expresiones como sí, muy bien, OK, ajá y similares corresponden a categoría 2.
9. Si el maestro cuenta un chiste o hace una broma sin referirse a un alumno se marcará la categoría 2, pero si se refiere a un alumno se usará la categoría 7.
10. Si todos los alumnos responden en grupo preguntas cortas se usa categoría 8.

Lineamientos para construir la matriz de interacciones

Identificados *episodios* (dedicados a una misma actividad) separados por signos de igual (=) en el formato de registro, para construir la matriz de interacciones se prepara una *secuencia* con números registrados cada 3" y se forman *pares traslapados*; se registran en una *matriz de interacciones* que tiene 10 columnas y 10 renglones, que corresponden a las diez categorías del sistema.

Según los registros de las 10 categorías se pueden hacer análisis básicos, como el porcentaje de tiempo que habla el profesor (registros de las categorías 1 a 7 respecto al total); % de tiempo que hablan los alumnos (categorías 8 y 9 respecto al total); % de tiempo de silencio o confusión (categoría 0 respecto al total).

En estudios de Flanders, en Estados Unidos típicamente el profesor hablaba 68% del tiempo; los alumnos 20%; 12% restante corresponde a *Silencio o confusión*. El porcentaje de tiempo que hablaba el profesor con niños de 9 años era 53%; con jóvenes de 12 años, 61%; y con jóvenes de 13 años en matemáticas 70%.

Análisis más finos pueden incluir la proporción de enfoque de enseñanza directo (categorías 5, 6 y 7 respecto al total) proporción de enfoque de enseñanza indirecto /categorías 1, 2 y 3 respecto al total; la relación entre enfoques indirecto y directo (categorías1, 2 y 3 respecto a 5, 6 y 7); tasa de respuesta del profesor (categorías 1, 2 y 3 respecto a 1, 2, 3, 6 y 7); tasa de preguntas del profesor (categoría 4 respecto a 4 y 5); y tasa de iniciativas de los alumnos (categoría 9 respecto a 8 y 9).

El formato de registro del sistema de Flanders consiste en una matriz en la que las columnas corresponden a los minutos (de 1 a 30) y los renglones a períodos de tres segundos (20 por minuto).

Períodos de 3"	Minutos																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																				
3																				

18																				

FUENTE: CON BASE EN MEDLEY Y MITZEL, 1963: 271-273.

Las conductas observadas se registran siguiendo *verticalmente* las columnas, con números del 1 al 10, según la clave. Se usa un reloj con segundero. El tiempo se ajusta al fin de cada bloque de cinco minutos. Al final de un episodio se anota = y después se incluyen comentarios.

Sistema OScAR, Observation Schedule and Record. (Medley y Mitzel, 1958)

Este sistema es un esfuerzo por captar el mayor número posible de aspectos de lo que ocurre en un salón de clases. La forma en que se registran las observaciones es una hoja impresa por los dos lados, en los que hay una serie de bloques o secciones en los que se registran distintos elementos, como sigue:

ANVERSO DE LA HOJA

Sección superior izquierda para registrar ACTIVIDADES, seis partes (A — F):

A. Interacción Profesor-Alumnos y Alumnos-Profesor (*Teacher-Pupils & Pupils-Teacher*, TP-PT).

B. Comunicación unidireccional Profesor-Alumnos (*Teacher-Pupils*, TP).
C. Trabajo del Profesor no orientado a los Alumnos (*Teacher*, T).
D. Trabajo de los Alumnos no orientado al Profesor (*Pupils*, P).
E. Actividades de los alumnos orientadas a otros alumnos (*Pupils-Pupils*, PP).
F. Actividades de los Alumnos que implican desorden (*Pupils Mess*, PM).

Categorías para registrar ACTIVIDADES:

A TP-PT

A1 Maestro trabaja con alumno individual
A2 Maestro trabajo con pequeño grupo
A3 Maestro pregunta, Alumno responde
A4 Maestro responde pregunta de Alumno
A5 Maestro ignora pregunta de alumno
A6 Maestro dirige canto, ejercicio, gimnasia

B (TP)

B1 Maestro expone
B2 Maestro lee, cuenta historia
B3 Maestro habla al grupo
B4 Maestro ilustra algo en pizarrón
B5 Maestro ilustra con mapa o gráfica
B6 Maestro hace demostración
B7 Maestro pone película, filmina, disco
B8 Maestro pasa hojas, libros

C (T)

C1 Maestro trabaja solo en escritorio
C2 Maestro limpia o decora el aula
C3 Maestro escribe o dibuja en pizarrón
C4 Maestro habla con visitante
C5 Maestro sale del aula o regresa a ella

D (P)

D1 Alumno lee o estudia en su lugar
D2 Alumno escribe o trabaja en su lugar
D3 Alumno pinta, cuenta, dibuja…
D4 Alumno trabaja en pizarrón
D5 Alumno decora el aula o pizarrón
D6 Alumno limpia el aula o pizarrón
D7 Alumno descansa, toma refrigerio
D8 Alumno sale del aula o regresa a ella
D9 Alumno pone manos en cabeza…

E (PP)

E1 Alumno habla al grupo
E2 Alumno recita
E3 Alumno repite, da charla preparada
E4 Alumno lee en voz alta
E5 Alumno hace demostración, ilustra
E6 Alumno hace representación o juego
E7 Alumno canta, toca instrumento
E8 Alumno hace ejercicio gimnástico
E9 Alumno escenifica rol, personaje
E10 Alumno dirige la clase

F (PM)

F1 Alumno ignora pregunta del maestro
F2 Alumno pelea o disputa
F3 Alumno cuchichea
F4 Alumno ríe
F5 Alumno juega con papeles, libros…
F6 Alumno habla con visitante

Sección inferior izquierda, para registrar AGRUPAMIENTOS (*Groupings*, G):
Se distinguen dos tipos de agrupamiento, que se registran por separado: a la derecha, grupos administrativos, organizados por el Profesor; a la izquierda grupos sociales, no organizados por el Profesor, pero también implican interacción de los alumnos entre sí y/o con el Profesor.
Categorías para registrar AGRUPAMIENTOS:

G1 Al menos ½ clase trabaja en grupo
G2 Al menos ½ clase habla en grupo
G3 De 4 alumnos a ½ clase trabaja grupo
G4 De 4 alumnos a ½ clase hablan en grupo
G5 2-3 alumnos trabajan en grupo
G6 2-3 alumnos hablan en grupo
G7 Alumnos actúan individualmente

Sección superior derecha, para registrar SEÑALES (S).

- Categorías para registrar SEÑALES:

S2 Maestro hace movimientos amigables
S3 Alumno hace movimientos amigables
S5 Maestro llama la atención a alumno
S6 Maestro muestra afecto a alumno
S7 Alumno muestra afecto a Maestro
S8 Alumno muestra hostilidad a alumno
S9 Alumno muestra hostilidad a alumno
S10 Maestro usa sarcasmo
S11 Maestro grita

Sección inferior derecha, para registrar MATERIALES (M).
Distingue los que usa el Profesor y los que usan los Alumnos; los primeros se registran a la derecha, y los segundos a la izquierda. Incluye las categorías designadas con las letras L, M, N, O, P, Q y R.

Categorías para registrar MATERIALES:

L1 Pizarrón
L2 Mapa, gráfica, imagen
L3 Diapositiva, película
M Material de audio
N5 Objeto
N6 Dispositivo especial
O Nada
P1 Texto, ejercicios
P2 Lectura complementaria
Q Material de escritura
R Objeto artístico, artesanal

REVERSO DE LA HOJA

En la cara posterior de la hoja hay secciones destinadas a registrar otros aspectos.
Sección superior: Relación verbal y gestual del Profesor con los Alumnos (K).
Categorías:

K1 Miradas-gestos que expresen aprobación o afecto hacia alumno
K2 Frases que expresen aliento a alumnos
K3 Frases para estructurar problemas
K4 Actividades varias
K5 Órdenes
K6 Llamadas de atención
K7 Miradas-gestos que expresen reprobación u hostilidad hacia un alumno
K8 Manifestaciones para afirmar la autoridad
Sección inferior: para registrar las Materias enseñadas (T).
Categorías:

T1 Lectura
T2 Matemáticas
T3 Lengua
T4 C. Sociales
T5 Ciencias
T6 Recreo
T7 E. Artísitica
T8 Música
T9 Socialización
T10 Exámenes

Modo de utilización

La observación se realiza durante aproximadamente 40 minutos, de los cuales 30 se dividen en seis períodos de 5', que se identifican con números romanos (I a VI). Durante los tres períodos impares (I, III y V, que corresponden a los minutos 1 a 5, 11 a 15 y 21 a 25), la atención del observador se concentrará en los aspectos que deben registrarse en el anverso de la hoja; durante los tres períodos pares (II, IV y VI, de los minutos 6 a 10, 16 a 20 y 26 a 30), la atención se pondrá en los aspectos que se registrarán en el reverso. En cada período se procederá como sigue:

▪ Períodos impares, I III y V

Después de anotar datos de identificación del grupo a observar, profesor, materia, día, hora y lugar, inicia el primer período de observación, poniendo en marcha el cronómetro. Durante 4' se observan *actividades* que tienen lugar en el aula, según los grupos A, B, C, D, E y F. En la columna I de la forma, que corresponde al primer período, se anota una marca en el renglón que corresponda a cada conducta que se observe según las categorías A1 a F6. Si una conducta se repite no se vuelve a anotar, por lo que no habrá varias marcas en cada categoría, sin importar el número de veces que se haya presentado cada conducta. Al mismo tiempo, el observador prestará atención a la sección de *Señales*, según las categorías S2 a S11, anotando las que observe, pero anotando una marca cada vez que detecte una.

Durante el quinto minuto de cada período el observador centrará su atención en las secciones *Agrupamientos y Materiales*, con las categorías respectivas, anotando en las columnas I y los renglones respectivos los *Agrupamientos* observados y los *Materiales utilizados*. En las secciones a que se refiere este párrafo hay dos bloques de columnas numeradas I, III y V a derecha e izquierda de los renglones centrales en que se plasman las categorías respectivas, distinguiendo grupos administrativos o sociales, y materiales utilizados por el profesor o los alumnos. Por ello deberán hacerse anotaciones en uno y otro lado de cada bloque. Si en el quinto minuto el observador detecta un elemento correspondiente a las secciones de *Actividades* (A—F) o *Sentimientos* (S) completará el registro donde corresponda. Se procederá de la misma forma durante los períodos III (minutos 11 a 15) y V (21 a 25).

▪ Períodos pares, II, IV y VI

Al terminar cada período impar, el observador dará vuelta a la hoja de registro y anotará, en la sección de *Materias enseñadas*, la o las que hayan sido tratadas durante

el período impar anterior. Inmediatamente después el observador pondrá en marcha su cronómetro para iniciar a contar el tiempo del primer período par de observación, y concentrará su atención en la *Relación verbal y gestual* del profesor con los alumnos, anotando una marca por cada frase o expresión gestual que caiga dentro de las categorías K1 a K8.

Al final del período de cinco minutos el observador anotará en la columna correspondiente al segundo período (II) de la sección *Materias* la o las tratadas en el período que acaba de terminar, tras lo cual dará vuelta a la hoja y pondrá en marcha su reloj para el siguiente período de cinco minutos (III), procediendo como se indicó, y así seguirá hasta el final.

La complejidad de este protocolo es evidente. OScAR fue desarrollado para un estudio de seguimiento de docentes recién egresados del programa de formación inicial en el que se prepararon. Según sus autores, el sistema fue diseñado con la intención de:

> [...] permitir registrar tantos aspectos significativos como fuera posible de lo que ocurre en el aula, sin buscar relacionarlos con algunas dimensiones o escala. La única preocupación del observador era ver y oír todo lo que pudiera de lo que estaba ocurriendo, y registrar todo lo que pudiera sin hacer supuesto alguno sobre su importancia relativa, o su relevancia respecto a cualquier dimensión conocida. (Medley y Mitzel, 1963: 280-281)

Para el diseño de este sistema sus autores aprovecharon el de Cornell, Lindvall y Saupe, presentado antes. Con base en la experiencia de la aplicación de uno y otro, por lo que se refiere a la calidad de la información obtenida, Medley y Mitzel dicen que la confiabilidad es sin duda mejor si la observación es hecha por dos personas, pero que esto es costoso, por lo que proponen que el sistema OScAR sea aplicado por un solo observador.

A esta conclusión se llegó después de la primera aplicación, en la que se observó durante diez semanas a 49 maestros (46 mujeres y 3 varones) de primarias de Nueva York. Los observadores fueron seis, organizados en parejas que trabajaron en la misma escuela, pero observando individualmente a un docente distinto. Para analizar los datos obtenidos se combinaron ítems en 20 *claves*.

La confiabilidad de cada clave se estimó con un análisis de varianza en que se consideró a los *maestros* y las *visitas* a su salón como efectos aleatorios. Se utilizó la media de los puntajes asignados por seis observadores en 12 visitas. Después de descartar seis claves, los coeficientes de confiabilidad fueron de .605 a .916.

Con un análisis factorial, las 14 claves se agruparon en tres dimensiones: *Clima emocional* (Confiabilidad .903), *Énfasis verbal* (.770) y *Estructura social* (.826). (Medley y Mitzel, 1963: 280-283)

Lo anterior muestra la sofisticación metodológica a la que se llegaba a fines de la década de 1950, pero en una reflexión que muestra al mismo tiempo los límites de los acercamientos desarrollados en el marco del paradigma conductista, Medley y Mitzel reconocen honestamente hasta dónde podían llegar sus resultados, en relación con las tres dimensiones mencionadas:

> Un defecto principal del sistema OScAR es que no consigue captar ningún aspecto de la conducta del aula que esté relacionado con el grado en que los alumnos logran los objetivos cognitivos. Las tres dimensiones que mide representan lo que son probablemente las diferencias más obvias entre las clases: qué tan ordenadas y relajadas son, en qué formas se agrupan los alumnos, y el contenido general de las lecciones que se enseñan. Medir estos aspectos fue relativamente fácil; medir diferencias más sutiles y cruciales con OScAR será probablemente más difícil. Sin embargo, no hay razón para pensar que sea imposible. (1963: 286)

Sistema de instantáneas, Snapshot. (Stallings, 1977)

- **Formato de entorno físico del aula** (*Physical Environment Information*, PEI). Se llena una vez al día y da información sobre los patrones de organización del mobiliario escolar y la presencia y uso de equipo y materiales.
- **Lista de cotejo** (*Classroom Check List*, CCL). Se llena cuatro veces por hora, cinco horas al día durante tres días; total 60 formas/aula. Para llenar cada una el observador recorre visualmente el aula a partir de la puerta de entrada, en el sentido de las manecillas del reloj, registrando en la forma a todas las personas presentes, atendiendo cuatro aspectos: actividades que tienen lugar; materiales que se usan; personas que participan; y forma de agruparse de niños y docentes:
  - **Actividades**: 21 tipos en otros tantos renglones, y uno más que se marca cuando hay dos observadores para verificar la confiabilidad.
  - **Materiales**: seis tipos, y espacio para registrar actividades de las áreas curriculares de matemáticas, lengua ciencias sociales y ciencias naturales.
  - **Participantes**: tres tipos de adulto identificados con mayúscula TAV (*Teacher, Aide, Volunteer*), y los niños, con minúscula (i = *independent child*). Esta dimensión se recoge en las cuatro líneas en que se subdivide cada uno de los renglones principales, líneas que se identifican con las letras TAVI.

- **Agrupamientos**: *One Child, Two Children, Small Groups*, Large Groups. Las líneas TAVi tienen números: One Child 1, 2, 3; Two Children 1,2,3; Small Groups 1, 2, 3, 4; Large Groups 1, 2. Así es posible registrar lo que hacen adultos y niños, tanto si trabajan todos en una misma actividad o si hay subgrupos en actividades diferentes. Se registra a todas las personas (menos el observador) una sola vez. Cada registro puede hacerse en un minuto lo que teóricamente dejaría espacios de descanso.

- **Forma para registrar interacciones cada 5'** (*Five-Minute Interaction*, FMI). Se llena cuatro veces por hora, después de la CCL. Se considera Quién habla o interactúa, A quién se dirige, Qué dice y Cómo lo hace (*Who-To Whom-What-How*). Las formas tienen 76 cuadros; en c/u se debe hacer un registro cada 5". En cinco minutos se podrán hacer 60; en una hora 4 x 60 = 240; en cinco horas 1,200; en tres días 3,600 registros. Los códigos para registrar interacciones son:

Who-To Whom	What	How
T = Teacher	1 = Command or Request	H = Happy
A = Aide	2 = Open-ended Question	U = Unhappy
V = Volunteer	3 = Response	N = Negative
C = Child	4 = Instruction, Explanation	T = Touch
D = Different Child	5 = Comment, Greetings, General Action	Q = Question
2 = Two Children	6 = Task-related Statement	G = Guide/Reason
S = Small Group (3-8)	7 = Acknowledge	P = Punish
L = Large Group (9 o más)	8 = Praise	O = Object
An = Animal	9 = Corrective Feedback	W = Worth
M = Machine	10 = No Response	DP = Dramatic Play/Pretend
	11 = Waiting	A = Academic
	12 = Observing, Listening	B = Behavior
	NV = Nonverbal	
	X = Movement	
R = Repeat the frameS = Simultaneous actionC = Cancel the frame		

La versión simplificada usada en la Ciudad de México es la siguiente:

Categorías de observación y formato de registro, México

Fecha _____ Hora de inicio _____ Hora de término _____
Secundaria _____ CCT _____
Grado/grupo_____ Asignatura _____ Alumnos _____
Profesor(a) _____Observador _____

FICHA DE OBSERVACIÓN EN CLASE							
PERSONAS		Actividades				Materiales	
			Con material		Sin material		
D	Docente	1	Lectura en voz alta	8	Interacción social	I	Sin material
A	Alumno	2	Exposición y demostración	9	Alumnos no involucrados Anote número	II	Portadores de texto, elementos de lectura
Tamaño del Grupo		3	Pregunta/respuesta, debate, discusión	10	Disciplina	III	Cuaderno/elementos de escritura
T	Todo incluso D	4	Práctica/memorización	11	Administración de clase	IV	Pizarrón
G	Grande (6 o más)	5	Monitoreo/tarea/trabajo individual/ejercicios	12	Clase por sí solo	V	Material didáctico
P	Pequeño (2 a 5)	6	Copiar	13	Interacción social del D, D no involucrado	VI	TIC
1	1 alumno	7	Instrucción verbal	14	Docente fuera	VII	Cooperativo

FORMATOS PARA REGISTRAR 10 OBSERVACIONES INSTANTÁNEAS

Instantánea	Descripción		Act.	Materiales	T. Gpo.
1		D			
		A			
Hora		A			
		A			
		A			
Instantánea	Descripción		Act.	Materiales	T. Gpo.

2		D			
		A			
Hora		A			
		A			
		A			
Instantánea	**Descripción**		**Act.**	**Materiales**	**T. Gpo.**
10		D			
		A			
Hora		A			
		A			
		A			

Referencias

Introducción

Cicourel, A. V. (1964). Method & *Measurement in Sociology*, New York: Free Press.

Ferguson, A. ***et al.*** (1940). Quantitative Estimation of Sensory Events. *Advancement of science.* No. 2, pp. 331-349.

Fishburn, P. (2001). Measurement Theory: Conjoint. En Smelser, Neil J. y Baltes, Paul B. (2001). I*nternational Encyclopedia of the Social and Behavioral Sciences* (pp. 9448-9451). Oxford: Elsevier-Pergamon.

Hand, D. J. (2016). *Measurement. A Very Short Introduction*. Oxford-New York: Oxford University Press.

Keeves, J. P. (1988). The Improvement of Measurement for Educational Research. En Keeves (Ed.). Educational Research, *Methodology and* Measurement. *An International Handbook* (pp. 241-246). Oxford-New York: Pergamon Press.

Keats, J. A. (1988). Measurement in Educational Research. En Keeves, J. P. (Ed.). *Educational Research*, *Methodology and Measurement. An International Handbook* (pp. 253-260). Oxford-New York: Pergamon Press.

Michell, J. (2001). Measurement Theory: History and Philosophy. En Smelser, Neil J. y Baltes, Paul B. (2001). International Encyclopedia of the Social and Behavioral Sciences (pp. 9451-9454). Oxford: Elsevier-Pergamon.

Robinson, A. (2007). *The Story of Measurement*. London: Thames & Hudson.

Sleeper, R. W. (1989). Reseña, *The Closing of the American Mind y Cultural Literacy* de A. Bloom y E. D. Hirsch. *International Journal of Qualitative Studies in Education*, Vol. 2, No. 1 pp. 81-86.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. Science, New Series, Vol. 103, N° 2684 (June 7), pp. 677-680.

Stevens, J. C. (1976). Psicofísica. En Sills, D. L. (Ed.). *Enciclopedia Internacional de las Ciencias Sociales*. Madrid: Aguilar, Vol. 8: 655-661.

Suck, R. (2001). Measurement, Representational Theory of. En Smelser, N. J. y Baltes, P. B. (2001). *International Encyclopedia of the Social and Behavioral Sciences* (pp. 9442-9448). Oxford: Elsevier-Pergamon.

Acercamientos basados en interrogación

Cuestionarios

Converse, J. M. y S. Presse. (1986). Survey questions. *Handcrafting the standardized questionnaire.* Series Quantitative Applications in the Social Sciences. N° 63. Beverly Hills: Sage.

Fowler, F. J. (1995). *Improving survey questions*. Design and evaluation. Applied Social Research Methods Series. N° 38. Newbury Park: Sage.

Koziol, S. M. y Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *Journal of Educational Research*, 79(4): 205-209.

Martin, E. (2006). Vignettes and Respondents Debriefing. for Questionnaire Design and Evaluation. Washington. U. S. Bureau of Census. Research Report Series, Survey Methodology. N° 2006/8.

Martin, E. *et al.* (1991). An application of Rasch analysis to questionnaire design: using vignettes to study the meaning of work in Current Population survey. *Journal of the Royal Statistical Society*. D-40 (3): 265-276.

Morales Vallejo, P., Urosa Sanz, B. y Blanco Blanco, Á. (2003). C*onstrucción de escalas de actitudes tipo Likert. Una guía práctica.* Madrid: La Muralla.

Morgenstern, C. y J. P. Keeves (1997). Descriptive scales. Keeves, J. P. *Educational research, methodology & measurement.* Oxford. Elsevier: 900-908.

Rowan, B. y Correnti, R. (2009). Studying Reading Instruction With Teacher Logs: Lessons From the Study of Instructional Improvement. *Educational Researcher*, Vol. 38 (2): 120-131.

Rowan, B., Camburn, E. y Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: a study of literacy teaching in third-grade classrooms. *The Elementary School Journal*, 105, 75-102.

Stecher, B. *et al.*, (2006). Using Structured Classroom Vignettes to Measure Instructional Practices in Mathematics. *Educational Evaluation and Policy Analysis*, Vol. 28 (2): 101-130.

Sudman, S. y N. M. Bradburn (1987). *Asking questions. A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Suskie, L. A. (1992). *Questionnaire survey research. What works*. Resources for Institutional Research, N° 6. Tallahassee. Association for Institutional Research.

De Vellis, R. F. (1991). *Scale development. Theory and applications*. Applied Social Research Methods Series, Vol. 26. Newbury Park: Sage.

Entrevistas

Brenner, M. E. (2006). Interviewing in Educational Research. En Green, J. L., G. Camilli y P. B. Elmore (Eds.). *Handbook of Complementary Methods in Education Research* (21, pp. 357-370). New York: Routledge.

Bruner, J. (2006). Actos de significado. *Más allá de la revolución cognitiva. Madrid:* Alianza Editorial. (Edición original en inglés, 1990).

Kvale, S. (1996). InterViews: *An Introduction to Qualitative Research Interviewing*. Thousand Oaks: Sage.

Kvale, S. y Brinkmann, S. (2009). InterViews: *Learning the Craft of Qualitative Research* Interviewing. Thousand Oaks: Sage.

Gubrium, J. F. y Holstein, J. A. (Eds.). (2002). *Handbook of Interview Research: Context and Method*. Thousand Oaks: Sage.

Gubrium, J. F., Holstein, J. A., Marvasti, A. B. y McKinney, K. D. (Eds.). (2012). *The SAGE Handbook of Interview Research*: The Complexity of the Craft, 2nd Ed. Thousand Oaks: Sage.

Krueger, R. A. y Casey, M. A. (2015). Focus Groups: *A Practical Guide for Applied Research, 5th Ed.* Los Angeles: Sage.

Leighton, J. P. (2009). *Two types of Think-Aloud Interviews for Educational Measurement: Protocol and Verbal Analysis.* NCME, Abril 14-16, San Diego.

Leighton, J. P. (2017). *Using Think-Aloud Interviews and Cognitive Labs in Educational Research.* Understanding Qualitative Research. New York: Oxford University Press.

Leighton, J. y Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge Univ. Press.

Merton, R. (1987). The focused interview and focus groups: Continuities and discontinuities. *Public Opinion Quarterly*, 51: 550-566.

Merton, R. K., Fiske, M. y Kendall, P. L. (1990). The Focused Interview. *A Manual of Problems and Procedures*, 2nd Ed. New York: The Free Press. (Edición original, 1956).

Merton, R. K. y Kendall, P. L. (1946). The Focused Interview. *American Journal of Sociology*, 51: 541-557.

Mislevy, R. J. (2006). Cognitive Psychology and Educational Assessment. En Brennan, R. L. (Ed.). *Educational Measurement*, 4th Ed. Washington: American Council on Education-Praeger Publishers.

Snow, R. E. y Lohman, D. F. (1989). Implications of Cognitive Psychology for Educational Measurement. En Linn, R. L. (Ed.). *Educational Measurement,* 3rd Ed. Washington: ACE-Macmillan Publishing Co.

Stewart, D. W. y Shamdasani, P. N. (2015). Focus Groups: *Theory and Practice*, 3rd Ed. Applied Social Research Methods Series, N° 20. Thousand Oaks: Sage.

Willis, G. B. (2005). *Cognitive Interviewing. A Tool for Improving Qustionnaire Design*. Thousand Oaks: Sage.

Acercamientos basados en observación

Generalidades, sistemas, simples, dinámicos, para registrar interacciones

Croll, P. y D. Moses. (1985). One in Five: *The Assessment and Incidence of Special Educational Needs*. London: Routledge and Keagan Paul.

Erickson, F. (1989). *Qualitative Methods in Research on Teaching*. En Wittrock, M. C. (Ed.). *Handbook of Research on Teaching*. 3rd Ed. (Pp. 119-161). New York: Macmillan Publ.

Everston, C.M. y Green, J. L. (1986). Observation as Inquiry and Method. En Wittrock, M. C. (Ed.). *Handbook of Research on Teaching. Third Ed.* (pp. 162-213). New York: Macmillan Publ. Co.

Floden, R. E. (2001). Research on Effects of Teaching: A Continuing Model for Reseach on Teaching. En Richardson, V. (Ed.). *Handbook of Research on Teaching, Fourth Edition.* (pp. 3-16). Washington: AERA.

Goe, L., C. Bell y O. Little. (2008). *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. Washington: NCCTQ.

Good, Th. L. y Brophy, J. E. (2010). *Looking in Classrooms*, 10thed.Boston: Pearson. (1st Ed. 2000).

Medley, D. M. y Mitzel, H. E. (1963). Measuring Classroom Behavior by Systematic Observation. En Gage, N. L. (Ed.). *Handbook of research on teaching*. (pp. 247-328). Chicago: Rand McNally.

Rosenshine, B. y N. Furst (1973). The use of direct observation to study teaching. En Travers, R. M. W. (Ed.). *Second Handbook of Research on Teaching*. Chicago: Rand McNally College Publ. Co., pp. 122-183.

Stallings, J. A. (1977). Learning to Look. A Handbook on Classroom Observation and Teaching Models. Belmont, CA: Wadsworth Publiching Co., Inc.

Protocolos recientes

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D., Gitomer, D. H. y Pianta R. C. (2014). Improving Observational Score Quality: Challenges in Observer Thinking. En Kane, Kerr y Pianta (Eds.). *Designing Teacher Evaluation Systems. New Guidance from the Measures of Effective Teaching Project* (pp. 50-97). San Francisco: Jossey Bass.

Danielson, Ch. (1996). *Enhancing Professional Practice. A Framework for Teaching*. Alexandria: Association for Supervision and Curriculum Development.

Gitomer, D., Bell, C., Qi, Yi, McCaffrey, D., Hamre, B. K. y Pianta, R. C. (2014). The Instructional Challenge in Improving Teaching Quality: Lessons from a Classroom Observation Protocol. *Teachers College Record,* 116 (6): 1-32.

Gitomer, D. H., Phelps, G., Weren, B. H., Howell, H. y Croft, A. J. (2014). Evidence on the Validity of Content Knowledge for Teaching Assessments. En Kane, Kerr y Pianta (Eds.). *Designing Teacher Evaluation Systems. New Guidance from the Measures of Effective Teaching Project* (pp. 493-528). San Francisco: Jossey Bass.

Grossman, P., Loeb, S., Cohen, J. ***et al.*** (2010). Measure for measure: The relationships between measures of instructional practice in middle school English Language Arts and teachers' value-added scores. *NBER Working Paper*. N° 16015.

Hill, H. *et al.* (2010). *Mathematical Quality of Instruction (MQI). Coding Tool.* University of Michigan, Learning Mathematics for Teaching.

Hill, H. *et al.* (2008). Mathematical Knowledge for Teaching & Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, Vol. 26 (4): 430-511.

Hill, H., S. G. Schilling y D. L. Ball. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, Vol. 105: 11-30.

Joe, J. N., McClellan, C. A. y Holtzman, S. L. (2014). Scoring Design Decisions. Reliability and the Length and Focus of Classroom Observations. En Kane, Kerr y Pianta. (Eds.). *Designing Teacher Evaluation Systems. New Guidance from the Measures of Effective Teaching Project* [pp. 415-443]. San Francisco: Jossey Bass.

Kane, Th. J., Kerr, K. A., y Pianta, R. C. (Eds.). (2014). *Designing Teacher Evaluation Systems. New Guidance from the Measures of Effective Teaching Project.* San Francisco: Jossey Bass.

Learning Mathematics for Teaching Project (2011). Measuring mathematical quality of instruction. *Journal of Mathematics Teacher Education*. 14 (1): 25-47.

MET Project (2010a). *Overview: Teacher Observation Rubrics.* Measures of Effective Teaching. Teachscape.

MET Project (2010b). *The PLATO Protocol for Classroom Observations.* Bill & Melinda Gates Foundation.

MET Project (2010c). *The MQI Protocol for Classroom Observations.* Bill & Melinda Gates Foundation.

Pianta, R. C. y Hamre, B. K. (2009). Conceptualization, Measurement and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher*, Vol. 38 (2): 109—119

Pianta, R. C., Laparo, K. M. y Hamre, B. K. (2012). *Classroom Assessment Scoring System. Sistema para evaluar la dinámica de las aulas. Manua*l. Pre-K Spanish. Baltimore: Paul H. Brookes Publishing Co.

Schultz, S. E. y Pecheone, R. L. (2014). Assessing Quality Teaching in Science. En Kane, Th. J., Kerr, K. A. y Pianta, R. C. (Eds.). *Designing Teacher Evaluation Systems. New Guidance from the Measures of Effective Teaching Project* (pp. 444-492). San Francisco: Jossey Bass.

Soo Park, Y., Chen, J. y Holtzman, S. L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. En Kane, Kerr y Pianta (Eds.). Designing Teacher Evaluation Systems. *New Guidance from the Measures of Effective Teaching Project* (pp. 383-414). San Francisco: Jossey Bass.

Estudios con videograbaciones

Erickson, F. (2011). Uses of video in social research: A brief history. *International Journal of Social Research Methodology*, Vol. 14 (3): 179-189.

Hiebert, J. *et al.* (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 Video Study.* Washington: National Center for Educational Statistics.

Jacobs, J. K., Hollingsworth, H. y Givvin, K. B. (2007). Video-Based Research Made Easy. Methodological Lessons Learned from the TIMSS Video Studies. *Field Methods*, 19 (3): 284-294.

Jewitt, C. (2012). *An introduction to using video for research.* National Centre for Research Methods Working Paper. Economic & Social Research Council.

Klieme, E., Pauli, Ch. y Reusser, K. (2004). The Pythagoras Study: Investigating Effects of Teaching and Learning in Swiss and German Mathematics Classrooms. En Janik, T. y Seidel, T. T*he Power of Video Studies in Investigating Teaching and Learning in the Classroom* (pp. 137-160). Munster-New York: Waxmann.

Loera Varela A. *et al.* (2006). C*ambios en la práctica pedagógica en escuelas del PEC: videograbaciones de lecciones de matemáticas y español. Análisis de las fases dos y tres de la Evaluación cualitativa del Programa Escuelas de Calidad.* México: Universidad Pedagógica Nacional y Heurística Educativa.

Loera Varela, A. *et al.* (2007). *Cambios en la práctica pedagógica en escuelas del programa escuelas de calidad.* México: UPN.

Loera Varela, A. (2012). *Identificando la brecha en la enseñanza de las ciencias y las matemáticas. Comparación en los desempeños de docentes latino-americanos con participantes de los estudios TIMS* videos. México: BID. https://issuu.com/heconversa/docs/identificando_la_brecha_en_la_ense__c871b5b357e894 Consultado el 20/04/2016.

Loera Varela, A., Näslund-Hadley, E. y Alonzo, H. (2013). El desempeño pedagógico de docentes en Nuevo León: hallazgos de un estudio basado en videos de lecciones de matemáticas y ciencias. *Revista Latinoamericana de Estudios Educativos*. Vol. XLIII (2): 11-41.

National Research Council. (2001). *The power of video technology in international comparative research in education*. Washington: National Academy Press.

Najvar, P. ***et al.*** (2009). CPV Video Study: Comparative Perspectives on Teaching in Different School Subjects. En Janik, T. y Seidel, T. *The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (6, pp. 103-119). Munster-New York: Waxmann.

Rosenstein B. (2002). Video Use in Social Science Research and Program Evaluation. *International Journal of Qualitative Methodology*. 1(3): 1-38.

Roth, K. J. (2009). Using Video Studies to Compare and Understand Science Teaching: Results from the TIMSS Video Study of 8th Grade Science Teaching. En Janik, T. y Seidel, T. T*he Power of Video Studies in Investigating Teaching and Learning in the Classroom* (2, pp. 23-37). Munster-New York: Waxmann.

Roth, K. J. *et al.* (1999). *Teaching science in five countries: Results from the TIMSS 1999 Video Study*. Washington: National Center for Educational Statistics.

Seidel, T., Prenzel, M., Schwindt, K., Rimele, R., Kobarg, M. y Dalehefte, I. (2004). The link between teaching and learning. Investigating effects of physics teaching on student learning in the context of the IPN Video Study. En Janik, T. y Seidel, *T. The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (pp. 161-180). Munster-New York: Waxmann.

Sherin, M. G. (2004). New perspectives on the role of video in teacher education. En Brophy, J. (Ed.). *Using video in teacher education*. (pp. 1-27). Oxford-New York: Elsevier.

Siegel, A. R. (2004). Telling Lessons from the TIMSS Videotape. En Evers, W. M. y Walberg, H. J. (Eds.). *Testing Student Learning, Evaluating Teaching Effectiveness* (pp. 161-193). Stanford: Hoover Institution Press.

Stiegler, J. W., y Hiebert, J. (2009). T*he Teaching Gap. Best Ideas from the World's Teachers for Improving Education in the Classroom* [1st ed. 1999]. New Yory: Free Press.

Stigler, J., Gallimore, R., Hiebert, J. (2000). Using Video Surveys to Compare Classrooms and Teaching Across Cultures: Examples and Lessons from the TIMSS Video Studies. *Educational Psychologist*, Vol. 35 (2): 87-100.

Stigler, J. W. ***et al.*** (1999). *The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan and the United States. Washington: National Center* for Educational Statistics.

Acercamientos basados en análisis de materiales

Borko, H., Stecher, B. y Kuffner, K. (2007). *Using Artifacts to Characterize Reform Oriented Instruction: The Scoop Notebook and Rating Guide.* (CSE Technical Report 707). Los Angeles, UCLA.

Borko, H. *et al.* (2005). Artifact Packages for Characterizing Classroom Practice: A pilot Study. *Educational Assessment*, Vol. 10 (2): 73-104.

Manzi, J., González, R. y Sun, Y. (Eds.). (2011). *La evaluación docente en Chile. Santiago, Chile:* Facultad de Ciencias Sociales, Escuela de Psicología, PUC.

Martínez, J. F., Borko, H., Stecher, B., Luskin, R. y Kloser, M. (2011). Measuring Classroom Assessment Practice Using Instructional Artifacts: A validation study of the QAS Notebook. Educational Assessment. 17, 2-3: 107-131.

Martínez, J. F., Borko, H. y Stecher, B. M. (2012). Measuring Instructional Practice in Science using Classroom Artifacts: Lessons Learned from Two Validation Studies. *Journal of Research in Science Teaching*, 49 (1): 38-67.

Matsumura, Lindsay *et al.* (2006). *Measuring Reading Comprehension and Mathematics Instruction in Urban Middle Schools: A Pilot Study of the Instructional Quality Assessment* (CSE 681). Los Angeles, UCLA.

Matsumura, Lindsay y J. Pascal (2003). *Teacher's Assignments & Student Work: Opening a Window on Classroom Practice.* (CSE 602). Los Angeles, UCLA.

Porter, A. C., Youngs, P. y Odden, A. (2001). *Advances in teacher assessments and their uses. En Richardson*, V. (Ed.). Handbook of Research on Teaching. Washington: AERA, pp. 259-297.

Ruiz-Primo M. A. y Li, M. (2013). Analyzing Teachers' Feedback Practices in Response to Students' Work in Science Classrooms. *Applied Measurement in Education* 26(3). DOI: 10.1080/08957347.2013.793188.

Webb, E. J., Campbell, D. T., Schwartz R. D., y Sechrest, L. (1966). *Unobstrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally College Publishing Co.

Las nuevas tecnologías y la obtención de información

Benfield, J. A. y Szlemko, W. J. (2006). Internet-Based Data Collection: Promises and Realities. *Journal of Research Practice*, Vol. 2 (2) Article D1.

Best, S. J. y Krueger, B. S. (2004). *Internet Data Collection*. Quantitative Applications in the Social Sciences, N° 141. Thousand Oaks: Sage.

Bruns, A. y Burgess, J. (2012). Doing blog research. En A., James, Waring, M., Coe, R. y Hedges, L. V. (Eds.). *Research Methods and Methodologies in Education* (202-208). Los Angeles: Sage.

Gaiser, T. J. (2011). Online Focus Groups. En Fielding, N., Lee, R. M. y Blank, G. (Eds.). *The SAGE Handbook of Online Research Methods (*290-306). Los Angeles: Sage. (1st. ed. 2008).

Hektner, J. M., Schmidt, J. A., y Csikszentmihalyi, M. (2007). Experience Sampling Method. Measuring the Quality of Everyday Life. Thousand Oaks: Sage.

Hine, Ch. (2011). Virtual Ethnography: Modes, Varieties, Affordance. En Fielding, N., Lee, R. M. y Blank, G. (Eds.). *The SAGE Handbook of Online Research Methods* (pp. 257-270). Los Angeles: Sage. (1st. ed. 2008).

Janetzko, D. (2011). Nonreactive Data Collection on the Internet. En Fielding, N., Lee, R. M. y Blank, G. (Eds.). *The SAGE Handbook of Online Research Methods* (pp. 161-173). Los Angeles: Sage. (1st. ed. 2008).

Sharpe, R. y Benfield, G. (2012). Internet-based methods. En A., James, Waring, M., Coe, R. y Hedges, L. V. (Eds.). *Research Methods and Methodologies in Education* (pp. 193-201). Los Angeles: Sage.

Stephens-Davidowitz, S. (2017). *Everybody Lies. Big Data, New Data, and What the Internet can Tell Us about Who We Really Are*. New York: Harper Collins-Dey St.

Wakeford, N. y Cohen, K. (2011). Fieldnotes in Public: Using Blogs for Research. En Fielding, N., Lee, R. M. y Blank, G. (Eds.). *The SAGE Handbook of Online Research Methods* (pp. 307-326). Los Angeles: Sage. (1st. ed. 2008).

Vehovar, V. y Lozar Manfreda, K. (2011). Overview: Online Surveys. En Fielding, N., Lee, R. M. y Blank, G. (Eds.). *The SAGE Handbook of Online Research Methods* (pp. 177-194). Los Angeles: Sage. (1st. ed. 2008).

Welser, H. T., Smith, M., Fisher, D. y Gleave, E. (2011). Distilling Digital Traces: Computational Social Sciences Approaches to Studying the Internet. En Fielding, N., Lee, R. M. y Blank, G. (Eds.). *The SAGE Handbook of Online Research Methods* (pp. 116-140). Los Angeles: Sage. (1st. ed. 2008).

El cuidado de la calidad de la medición

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, Authors.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, Authors.

Basterra, M. R., E. Trumbull y G. Solano (Eds.). (2011). *Cultural validity in assessment: Addressing linguistic & cultural diversity*. New York: Routledge.

Brennan, R. L. (2001). An Essay on the History and Future of Reliability from the Perspective of Replication. *Journal of Educational Measurement*. Vol. 38 (4): 295-317.

Chabris, Ch. F. y Simons, D. J. (2009). *The Invisible Gorilla. How Our Intuitions Deceives Us. New York:* Crown Publ. Traducción al español, *El gorilla invisible y otras maneras en las que nuestra intuición nos engaña*. Buenos Aires: Siglo **xx**I Editores, 2011.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.). *Educational Measurement* (2nd ed., pp. 443-507). Washington: American Council on Ed.

Cronbach, L. J. (1988). *Five perspectives on validity argument*. En Wainer, H & Braun, H. (Eds.), Test validity (pp. 3—17). Princeton: IEA.

Crooks, T. J., M. T. Kane y A. S. Cohen (1996). Threats to the Valid Use of Assessments. Assessment in Education, Vol. 3 (3): 265-285.

Cureton, E. E. (1951). Validity. En E. F. Lindquist (Ed.). *Educational Measurement* (1st ed., pp. 621-694). Washington: American Council on Education.

Feldt, L. S. y R. L. Brennan (1989). Reliability. En R. L. Linn (Ed.). *Educational Measurement* (3rd ed., pp. 105-146). New York: ACE & Macmillan.

Haertel, E. H. (2006). Reliability. En R. Brennan (Ed.). *Educational Measurement* (4th ed., pp. 65-110). Westport: ACE & Praeger.

Journal of Educational Measurement, 50 (1), Spring 2013. Special Issue on Validity.

Kane, M. (2001). Current Concerns in Validity Theory. J*ournal of Educational Measurement* Vol. 38 (4): 319-342.

Kane, M. (2006). Validation. En R. Brennan (Ed.). *Educational Measurement* (4th ed., pp. 17-64). Westport: American Council on Education & Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educationa*l *Measurement*, 50 (1): 1—73.

Lissitz, R. (Ed.). (2009). *The concept of validity. Revisions, New Directions and Applications. Charlotte*: Information Age Publ.

Messick, S. (1989). Validity. En R. L. Linn (Ed.). *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education & Macmillan.

Messick, S. (1998). Test Validity: A Matter of Consequence. *Social Indicators Research,* 45(1-3), 35-44.

Michell, J. (2000). Normal Science, Pathological Science and Psychometrics. Theory & Psychology, Vol. 10 (5): 639-667.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, 23: 5-12.

Moss, P. (2008). A critical review of the validity research agenda of the NBPTS at the end of its first decade. En L. Ingvarson y J. Hattie (Eds.). *Assessing teachers for professional certification: the first decade of the National Board for Professional Teaching Standards* (pp. 257—312). Oxford: Elsevier.

Newton, P. E. (2013). *Does it matter what 'validity' means?* Presentación en el Departamento de Educación de la Universidad de Oxford, febrero 4.

Sireci, S. G. (2013). Agreeing on Validity Arguments. *Journal of Educational Measurement*, 50 (1): 99—104.

Stanley, J. C. (1971). Reliability. En R. L. Thorndike (Ed.). *Educational Measurement* (2nd ed., pp. 356-442). Washington: American Council on Ed.

Starr, D. (2012). False eyewitness. Who are you going to believe? Me or your lying eyes? *Discover*, N° 11 (November), pp. 38-64.

Thorndike, R. L. (1951). Reliability. En E. F. Lindquist (Ed.). *Educational Measurement* (1st ed., pp. 560-620). Washington: American Council on Ed

CAPÍTULO 4
EL ANÁLISIS DE LA INFORMACIÓN

CONTENIDO

Introducción

Para responder las preguntas de investigación, y llegar a conclusiones, es necesario analizar la información obtenida con uno o varios de los acercamientos presentados en el capítulo anterior. En este veremos algunas técnicas para dicho análisis, y en especial las que se pueden emplear para analizar datos cuantitativos.

Algunos cursos de posgrados de investigación buscan que los estudiantes alcancen un buen nivel de dominio de estas técnicas, lo que supone una base mínima de estadística. Muchos alumnos, sin embargo, no tienen buena formación matemática, pero incluso los que cuentan con ella no necesariamente tienen la comprensión conceptual indispensable para interpretar los resultados. Por ello, aunque con la ayuda de una computadora personal y un paquete de software puedan utilizar estadísticas complejas, no es raro que sus interpretaciones tengan errores graves.

Según el ateniense que conversa con Clinias en el *Diálogo de Las Leyes* de Platón, los ciudadanos deben aprender *la ciencia de los números y la que mide la longitud, anchura y profundidad de las cosas*, pero no es necesario que todos tengan un conocimiento exacto de estas ciencias. Como a mí, sin embargo, le preocupaba:

> [...] la general ignorancia que suele haber al respecto, pero me preocupa más habérmelas con los que se han dedicado al estudio de estas ciencias, pero las han estudiado mal. La ignorancia absoluta no es el mayor de los males ni el más temible; una vasta extensión de conocimientos mal digeridos es mucho peor [...] (Las Leyes, Libro VII, 817-819)

Este capítulo busca ayudar a evitar eso, entendiendo que no se puede esperar que el lector llegue a dominar las técnicas presentadas simplemente por su lectura. Es ne-

cesario un proceso largo y laborioso que incluya estudiar textos especializados, revisar trabajos en que se apliquen ciertas técnicas, y, aplicarlas uno mismo. Sólo así se podrá llegar a dominar algunas en el nivel de experto. Siempre ayudarán, desde luego, buenos cursos metodológicos conducidos por investigadores con experiencia en el uso de las técnicas de que se trate, en el marco de programas de posgrado.

Los apartados 1 y 2 (fundamentos y técnicas básicas) presentan unos y otras con bastante detalle, reduciendo a un mínimo la formalización matemática y enfatizando la comprensión conceptual. El inciso 3 (técnicas avanzadas), en cambio, ofrece solo una escueta descripción con una idea general, y refiere a literatura especializada. Esto se debe a dos razones: una, que no tengo los conocimientos necesarios para presentar de manera detallada esas técnicas avanzadas; y otra que, a mi juicio, el dominio de todas esas técnicas es un propósito que rebasa el alcance de cualquier posgrado. Creo más realista dar solo una visión general, dejando que quien necesite dominar alguna deberá hacer un esfuerzo suplementario para conseguirlo.

El análisis de la información siempre debe tener en cuenta la distinción entre *niveles de medición* de la información (*cfr*. Cap. 3), con la distinción entre el nivel nominal, el ordinal y el cardinal o métrico, de intervalo o de razón. Esto lleva a reflexionar sobre el paso que lleva de la obtención de datos al análisis: la codificación, para lo que hay que analizar el sentido del término *datos*.

Algunas personas rechazan el término acusándolo de ser positivista o cuantitativo. Etimológicamente se refiere a algo "dado", cuando en realidad, para poder recolectarse, los *datos* se generan o construyen. En general, el término denota:

> [...] la materia prima básica que estudiamos. Se refiere a percepciones o pensamientos que simbolizamos de alguna forma, con palabras, números o imágenes, y con los que planeamos hacer algo más, analizarlos. Sinónimos razonables de los términos "datos-análisis" son "evidencia-estudio". Decir "estudiar la evidencia" o "analizar los datos" parece principalmente cuestión de gusto. Pero sean lo que sean, los datos no hablan por sí mismos. Nosotros tenemos que hablar por ellos. (Vogt, Vogt, Gardner y Haeffele, 2014: 2)

Para analizar datos es indispensable traducir lo captado (interrogando a los sujetos, observándolos, o revisando rastros) a símbolos que corresponden a categorías o aspectos del objeto de estudio (variables), y pueden consistir en palabras, números (clasificatorios, ordinales o cardinales) o imágenes, que se analizarán de manera cualitativa, cuantitativa o gráfica, o combinando esas formas. Las categorías pueden ser definidas antes de

obtener la información, con base en estudios previos, o después de dicha obtención, pero son inevitables. Pretender que no existen es aceptar unas de manera no consciente. Es posible también partir de categorías que se afinarán a medida que avanza el proceso de recolección, e incluso de análisis.

Fundamentos

Análisis descriptivo y exploratorio

Una investigación no suele limitarse a este nivel elemental, sino que pretende al menos relacionar unas variables con otras, pero incluso en este nivel puede haber hallazgos interesantes. Además, describir una por una las variables consideradas es un primer paso indispensable de cualquier análisis complejo, que permite dar sentido a las relaciones que se encuentren y evitar interpretaciones erróneas.

Distribuciones de frecuencias

Lo primero que podemos hacer con datos de cierto número de sujetos es simplemente contar cuántos presentan cada valor de las variables consideradas, según su nivel de medición. Por ejemplo, cuántos hombres y cuántas mujeres se entrevistaron; cuántos sujetos manifestaron alto, medio o bajo gusto por las matemáticas o la lengua; o bien cuánto mide o pesa cada uno, o qué puntaje alcanza en una prueba.

El conjunto de valores de los sujetos constituye una *distribución de frecuencias*, que se puede hacer valor por valor, o agrupando valores, *v. gr.* los de los sujetos cuya estatura se sitúa entre 1.70 y 1.75 m. También es posible construir distribuciones de frecuencias *acumuladas*, en las que se agrupa el número de los sujetos que presentan cierto valor, junto con los que tienen valores inferiores (o superiores) a ese.

La Tabla 4.1 presenta el número de escuelas primarias en las entidades federativas de la República Mexicana en el ciclo escolar 2015-2016. Una primera columna presenta el número de escuelas entidad por entidad (frecuencia simple); la segunda presenta la frecuencia acumulada, o sea el número de escuelas en la entidad a la que se refiere el renglón, más las de todas las entidades que la preceden en la tabla.

TABLA 4.1. ENTIDADES FEDERATIVAS POR NÚMERO DE PRIMARIAS, 2015-2016.

Lugar	Entidad	Frecuencia simple	Frecuencia acumulada
1	Veracruz	9 606	9 606
2	Chiapas	8 505	18 111

3	México	7 813	25 994
4	Jalisco	5 859	31 783
5	Oaxaca	5 626	37 409
6	Michoacán	5 222	42 631
7	Puebla	4 624	47 255
8	Guerrero	4 618	51 873
9	Guanajuato	4 479	56 352
10	San Luis Potosí	3 299	59 651
11	Hidalgo	3 254	62 905
12	Ciudad de México	3 201	66 106
13	Chihuahua	2 834	68 940
14	Nuevo León	2 740	71 680
15	Sinaloa	2 601	74 281
16	Durango	2 573	76 854
17	Tamaulipas	2 422	79 276
18	Tabasco	2 105	81 381
19	Sonora	1 885	83 266
20	Zacatecas	1 875	85 141
21	Coahuila	1 840	86 981
22	Yucatán	1 397	88 378
23	Baja California	1 650	90 028
24	Querétaro	1 511	91 539
25	Nayarit	1 197	92 736
26	Morelos	1 184	93 920
27	Quintana Roo	860	94 780
28	Tlaxcala	795	95 575
29	Campeche	778	96 353
30	Aguascalientes	708	97 061
31	Colima	497	97 558
32	Baja California Sur	446	98 004
	Nacional	98 004	98 004

FUENTE: INEE (2017). *PANORAMA EDUCATIVO DE MÉXICO* 2016. ANEXO, TABLA ED04-A5.

El número de escuelas de cada entidad tiene que ver, desde luego, con su población total, pero el conocimiento del contexto de las entidades permite sospechar que otros factores influyen en ese número, en particular el que haya una proporción mayor de

población rural, lo que implica, por una parte, más niños en edad de asistir a la primaria y, por otra, menos ciudades medianas y grandes y más localidades pequeñas y dispersas, lo que implica menos escuelas grandes, con varios grupos por grado, y más escuelas chicas, con un solo grupo por grado o multigrado.

En la Tabla 4.2 las entidades se ordenan según el número total de sus habitantes, y las otras dos ofrecen los datos del total de alumnos de educación primaria, y el de las escuelas de este nivel educativo.

TABLA 4.2. HABITANTES, ALUMNOS Y ESCUELAS PRIMARIAS DE LAS ENTIDADES

Entidad federativa	Habitantes	Alumnos	Escuelas
México	16 187 608	1 936 448	7 813
Ciudad de México	8 918 653	879 568	3 201
Veracruz	8 112 505	901 785	9 606
Jalisco	7 844 830	933 684	5 859
Puebla	6 168 883	797 201	4 624
Guanajuato	5 853 677	718 506	4 479
Chiapas	5 217 908	781 031	8 505
Nuevo León	5 119 504	561 296	2 740
Michoacán	4 584 471	562 396	5 222
Oaxaca	3 967 889	531 074	5 626
Chihuahua	3 556 574	429 694	2 834
Guerrero	3 533 251	478 919	4 618
Tamaulipas	3 441 698	388 428	2 422
Baja California	3 315 766	385 387	1 650
Sinaloa	2 966 321	332 491	2 601
Coahuila	2 954 915	337 794	1 840
Hidalgo	2 858 359	355 796	3 254
Sonora	2 850 330	319 637	1 885
San Luis Potosí	2 717 820	327 585	3 299
Tabasco	2 395 272	296 411	2 105
Yucatán	2 097 175	230 400	1 397
Querétaro	2 038 372	249 687	1 511
Morelos	1 903 811	212 545	1 184
Durango	1 754 754	217 139	2 573
Zacatecas	1 579 209	196 740	1 875

Quintana Roo	1 501 562	176 865	860
Aguascalientes	1 312 544	157 794	708
Tlaxcala	1 272 847	155 152	795
Nayarit	1 181 050	137 067	1 197
Campeche	899 931	102 279	778
Baja California Sur	712 029	80 875	446
Colima	711 235	78 751	497
Nacional	119 530 753	14 250 425	98 004

FUENTE: INEGI ENCUESTA INTERCENSAL 2015; INEE (2017). *PANORAMA EDUCATIVO DE MÉXICO 2016*. ANEXO, TABLA ED04-A5.

Cuando se ofrecen cifras desglosadas por entidad federativa sobre diversos aspectos de la vida nacional es frecuente que las más altas sean las del Estado de México, sea que se trate de homicidios o autos robados, pacientes atendidos en clínicas del IMSS, o estudiantes de cualquier nivel educativo.

Algunas personas se sorprenden, lo que muestra simplemente que desconocen que esa entidad federativa es la más poblada del país, con casi el doble de habitantes en comparación con las dos que la siguen (la Ciudad de México y el Estado de Veracruz), y más de veinte veces más que los habitantes de Baja California Sur o Colima, las dos entidades menos pobladas del país.

La columna sobre número de alumnos de primaria de la Tabla 4.2 muestra que el Estado de México tiene también más del doble de niños en ese nivel que las dos entidades siguientes, pero que en este rubro Veracruz supera a la Ciudad de México, pese a tener menos habitantes.

Si se revisa la columna sobre el número de escuelas se advierte que la Ciudad de México tiene un número que parece pequeño en relación con su población.

Esto simplemente indica que casi todas las primarias son grandes y pocas pequeñas; de hecho, en la Ciudad de México no hay un solo plantel de educación comunitaria, mientras que en el Estado de México hay bastantes, y más en Veracruz. Chiapas que, con un tercio de la población del Estado de México tiene más escuelas; tiene también una cifra cercana a la de Veracruz, y mucho más alta que la de la Ciudad de México.

Para poder comparar con sentido datos de realidades de un tamaño tan diferente como las entidades federativas de México, es necesario usar cifras proporcionales, en particular porcentuales.

Conviene revisar este punto porque, por sorprendente que resulte, no pocos universitarios tienen dificultad para comprender tan elemental herramienta.

▪ **Datos absolutos o porcentuales**

Una distribución de frecuencias puede hacerse con cifras relativas, como porcentajes, que permiten comparar datos de tamaño desigual, lo que en términos absolutos puede fácilmente interpretarse mal. Veamos el total de primarias en las que al menos un docente atendía más de un grado, en el ciclo escolar 2003-2004, en varios estados de México, en orden descendente según el número absoluto de este tipo de planteles.

TABLA 4.3. PRIMARIAS MULTIGRADO EN ALGUNOS ESTADOS. CIFRAS ABSOLUTAS

Entidades	Escuelas multigrado	Entidades	Escuelas multigrado
Chiapas	4,950	Tabasco	1,098
Veracruz	4,792	Chihuahua	1,096
Oaxaca	2,432	Estado de México	1,044
Michoacán	2,301	Baja California Sur	128
San Luis Potosí	1,723	Colima	125
Durango	1,362	Tlaxcala	123

FUENTE: INEE (2004: 181). TABLA 4.7.

Si se considera que el número de primarias multigrado es indicador de carencias de las escuelas, la situación de Chiapas y Veracruz parece dos veces más precaria que la de Oaxaca y Michoacán, y peor que la de San Luis Potosí y Durango.

Por su parte Tabasco, Chihuahua y el Estado de México tendrían una situación muy similar, en tanto que la de Baja California Sur, Colima y Tlaxcala sería muy favorable. Pero si en vez de cifras absolutas se analizan porcentajes, que muestran la proporción de las escuelas multigrado en el total de primarias, la perspectiva cambia.

TABLA 4.4. PRIMARIAS MULTIGRADO EN ALGUNOS ESTADOS (CIFRAS ABSOLUTAS Y %)

Entidades	Escuelas Multigrado	Total de Primarias	% de multigrado
Chiapas	4,950	8,461	58.5
Durango	1,362	2,594	52.5
Tabasco	1,098	2,161	50.8
San Luis Potosí	1,723	3,474	49.6
Veracruz	4,792	9,800	48.9
Oaxaca	2,432	5,656	43.0

Michoacán	2,301	5,781	39.8
Chihuahua	1,096	2,892	37.9
Baja California Sur	128	407	31.5
Colima	125	494	25.3
Tlaxcala	123	755	16.3
Estado de México	1,044	7,406	14.1

FUENTE: INEE (2004: 181). TABLA 4.7.

En la Tabla 4.4 los estados se ordenan según el porcentaje en el total de escuelas con situación multigrado, que aparece en la última columna, con las cifras absolutas en que se basa el porcentaje en las dos columnas anteriores. Se puede apreciar así que, aunque el total de primarias multigrado de Durango es casi cuatro veces menor al de Chiapas, la situación de ambos estados en términos porcentuales es similar: más de la mitad de las primarias (58.5% y 52.5%) tienen situación multigrado, ya que Durango tiene también un total de primarias mucho menor que Chiapas. Algo similar pasa con Veracruz y San Luis Potosí, cuyos números absolutos son muy distintos, pero el porcentaje de multigrados es similar: 48.9% y 49.6%.

La tabla muestra también que Tabasco, Chihuahua y el Estado de México tienen un número similar de multigrados (1098, 1096 y 1044), pero como el total de primarias es diferente, el porcentaje es mayor en Tabasco (50.8%), intermedio en Chihuahua (37.9%), y bajo en el Estado de México (14.1%). Veracruz y San Luis Potosí tienen una proporción similar de escuelas multigrado en relación con el total de primarias (48.9% y 49.6%), aunque Veracruz tenga cerca del triple de multigrados, ya que tiene también cerca del triple de primarias. Baja California Sur, Colima y Tlaxcala, con poca población y primarias, tienen pocos multigrados (128, 125 y 123), pero la proporción que representan en el total de las primarias de cada entidad (31.5, 25.3 y 16.3) es mayor a la de los 1,044 planteles multigrado, que es solo 14.1% del total de 7,406 primarias del Estado de México, la entidad más poblada del país.

Un ejemplo más, que muestra la importancia del cambio de perspectiva que implica comparar cifras porcentuales o absolutas, se refiere al análisis de las carencias que rodean a buena parte de las poblaciones indígenas del país, y en particular las de sus regiones más pobres.

Es sabido que los resultados que obtienen en pruebas en gran escala los alumnos de escuelas indígenas suelen ser bastante bajos, y la atención suele centrarse en estados como Chiapas y Oaxaca, donde la población indígena es importante, y los indicadores

socioeconómicos las sitúan entre las entidades de menor desarrollo en México, frente a estados como Chihuahua, Jalisco, Durango y Nayarit, que tienen también población indígena, pero en menores proporciones, y gozan también, en promedio, de niveles de desarrollo más altos. Sin embargo, un análisis más fino que permite hacer la Tabla 4.5 muestra que las regiones y localidades de esos cuatro estados en las que hay escuelas indígenas tienen indicadores socioeconómicos particularmente negativos. (*cfr*. Martínez Rizo, 2006)

TABLA 4.5. INDICADORES DE LOCALIDADES INDÍGENAS

Estados	Población indígena	Localidades con 40% y más de indígenas		
		N° de Localidades	Población total	Viviendas sin electricidad (%)
Chihuahua	136,589	1,789	87,410	92.5
Durango	39,545	448	25,108	89.7
Jalisco	75,122	269	13,480	82.0
Nayarit	56,172	421	38,482	57.9
Chiapas	1,117,597	3,590	1,134,617	21.5
Oaxaca	1,648,426	4,478	1,521,462	19.6
Puebla	957,650	1,532	818,226	16.4
Yucatán	981,064	1,069	894,966	7.7

FUENTE: SERRANO, EMBRIZ Y FERNÁNDEZ (2003).

Chiapas y Oaxaca. La Tabla 4.5 muestra que alrededor de 20% (21.5 y 19.6%) de los indígenas de estas entidades que viven en localidades con 40% o más de ese tipo de población habitan en viviendas sin electricidad. Como el total es superior a 2.6 millones (1,134,617 + 1,521, 462), 20% es más de medio millón de personas.

Puebla y Yucatán. En estas entidades la población de localidades con 40% o más de indígenas es también numerosa (818,226 + 894,966 ~ 1.7 millones), pero el porcentaje de viviendas sin electricidad es de solo 16.4 y 7.7% respectivamente.

Chihuahua, Durango, Jalisco y Nayarit. En estas cuatro entidades, en cambio, la población indígena en localidades con 40% o más de indígenas es mucho menos numerosa (en total algo más de 150 mil personas), pero la proporción de quienes carecen de electricidad en sus viviendas es altísima (92.5, 89.7, 82 y 57.9%).

La consideración de las cifras porcentuales confirma la opinión de Embriz y Ruiz:

> La región cora-huichol-tepehuana de Durango, Nayarit y Jalisco, es la región más pobre de la nación mexicana. La sierra tarahumara, con rarámuris, guarijíos, pimas y tepehuanos, ocupa el segundo lugar en atraso de servicios en sus viviendas [...]. (Embriz y Ruiz, 2003)

Todo esto parece demasiado elemental, y lo es, en comparación con modelos estadísticos avanzados que se pueden utilizar en la investigación. Pese a ello, con sorpresa he constatado, una y otra vez, que algunos estudiantes de posgrados que buscan formar investigadores educativos tienen problemas para entender los porcentajes y emplearlos correctamente. El que las calculadoras de bolsillo más baratas tengan una tecla especial para calcular porcentajes es un indicio inequívoco de que no son raros los casos de personas en tal situación.

▪ Representación gráfica de distribuciones de frecuencias

Las distribuciones de frecuencias se pueden representar gráficamente, lo que facilita la visualización de algunas características, por ejemplo, si los valores se concentran alrededor de algún valor o están muy dispersos, si se distribuyen a uno y otro lado de cierto valor de manera simétrica o asimétrica, entre otros aspectos. Las gráficas pueden enriquecer el nivel básico, descriptivo o exploratorio, del análisis de la información, tanto si esta se refiere a una sola variable, o bien a dos o más.

En las gráficas más simples, con datos de una sola variable, cada barra se refiere a diversos valores o rangos de valor de la misma. En gráficas más complejas todas las barras dan información de una variable, pero los datos de cada barra se refieren a ciertos valores de otra variable, lo que permite hacer comparaciones. Como ocurre en las tablas, la información se puede presentar en cifras absolutas o relativas.

Unas gráficas muy comunes son las circulares o de pastel, en las que se representan los valores de una variable en términos relativos. La proporción de la superficie de cada segmento en relación con el total, que es el área del círculo, representa la proporción (%) de cada valor de la variable de que se trate respecto al total.

La Tabla 4.6 presenta el número de escuelas de educación básica (preescolares, primarias y secundarias) que había en las entidades federativas de la República Mexicana al inicio del ciclo escolar 2005-2006. Se ofrece el dato para cada una de las ocho entidades en que había más escuelas, seguido por el subtotal de esas entidades. En seguida, en un solo renglón, se indica el total de las escuelas que había en las otras 24 entidades, y por último el gran total. En una columna se dan las cifras absolutas y en otra los porcentajes correspondientes.

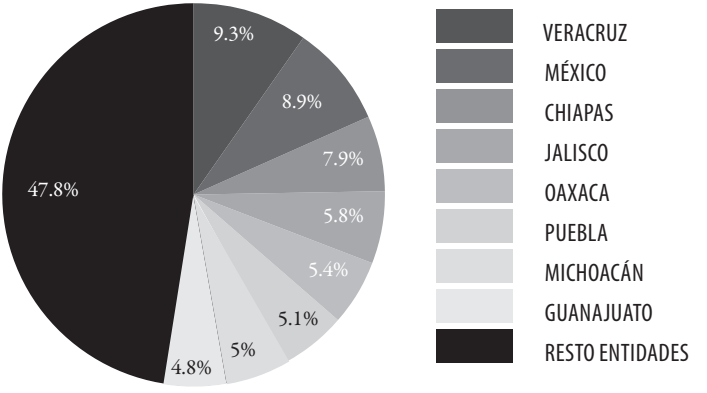
TABLA 4.6. ESCUELAS DE EDUCACIÓN BÁSICA POR ENTIDAD

Entidades	Escuelas N° absoluto	Escuelas porcentaje
Veracruz	19 883	9.27 %
Estado de México	19 148	8.93 %
Chiapas	17 039	7.95 %
Jalisco	12 350	5.76 %
Oaxaca	11 631	5.42 %
Puebla	11 037	5.15 %
Michoacán	10 705	4.99 %
Guanajuato	10 213	4.76 %
Subtotal	112 006	52.24 %
Otras 24 entidades	102 388	47.76 %
TOTAL	214 394	100.0 %

FUENTE: INEE (2006: 45). TABLA 1.14.

El renglón con el subtotal de las ocho entidades con más escuelas de educación básica muestra que en ellas se concentraba más de la mitad del total de ese tipo de planteles, mientras que las 24 entidades restantes juntas tenían menos escuelas que las ocho primeras, menos de la mitad del total que había en el país. Una gráfica circular permite apreciar esto visualmente.

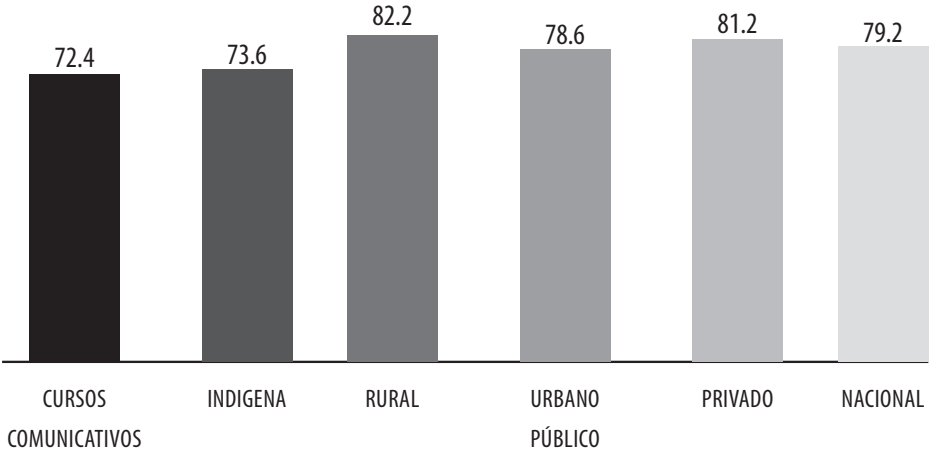
GRÁFICA 4.1. CONCENTRACIÓN DE ESCUELAS DE EDUCACIÓN BÁSICA EN LAS ENTIDADES



FUENTE: INEE (2006: 46). GRÁFICA 1.3.

Otras gráficas conocidas son las de barras, o *histogramas*. En ellas la altura de las barras representa el valor de cierta variable, según la escala del eje vertical. Estas gráficas pueden ofrecer información de más de una variable. La Gráfica 4.2 muestra el valor porcentual de la permanencia del docente en un grupo durante un ciclo escolar (variable común); además, cada barra se refiere a un tipo de servicio educativo, una segunda variable, con categorías nominales, según la modalidad de servicio.

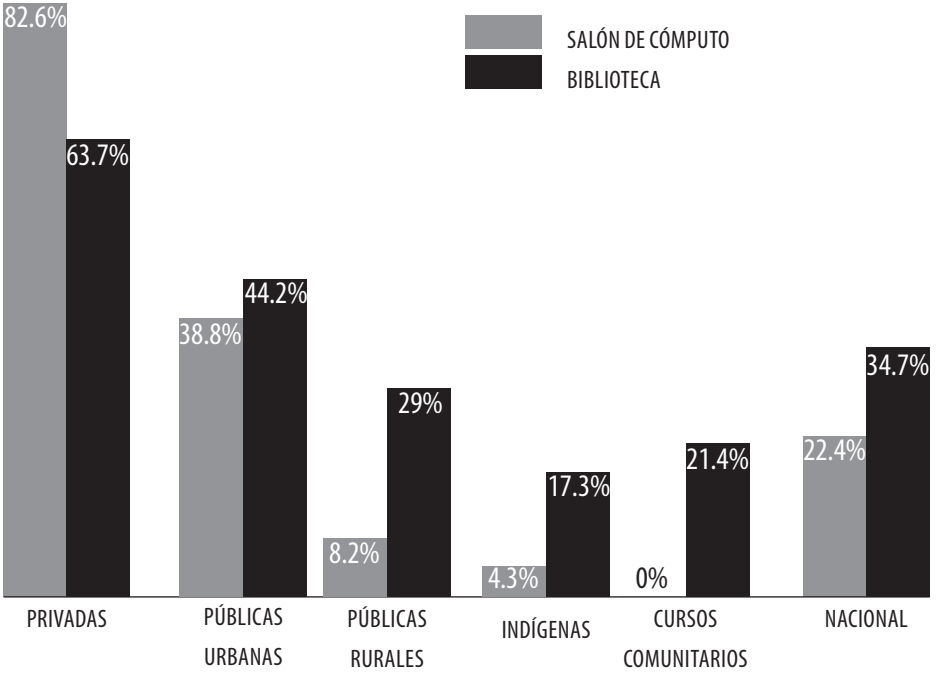
GRÁFICA 4.2. PORCENTAJE DE PROFESORES DE 3° DE PRIMARIA QUE PERMANECIERON EN EL MISMO GRUPO DESDE EL INICIO DEL CICLO ESCOLAR



FUENTE: INEE (2007: 26). FIGURA 1.3.

La Gráfica 4.3 presenta pares de barras con datos sobre dos variables (la existencia de aula de cómputo y biblioteca) en varios tipos de servicio (primarias privadas, públicas urbanas y rurales, indígenas y cursos comunitarios, así como en el total.

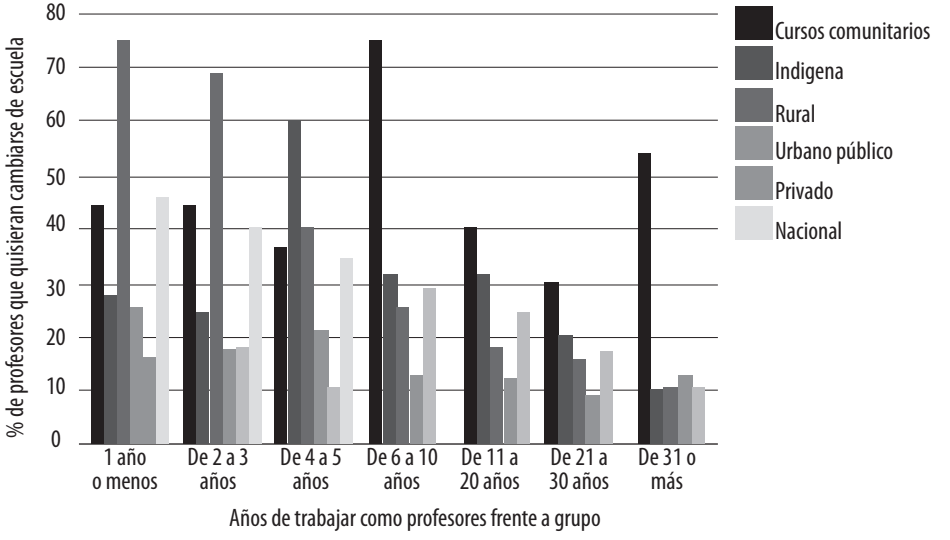
GRÁFICA 4.3. % DE PLANTELES, POR MODALIDAD, CON SALÓN DE CÓMPUTO Y BIBLIOTECA



FUENTE: INEE (2006: 109). GRÁFICA 3.1.

La Gráfica 4.4 incluye siete grupos de barras simples, que informan del porcentaje de maestros que expresaron el deseo de cambiar de escuela, y cada uno de los grupos distingue la antigüedad del maestro. El sombreado distingue tipo de escuela del informante, lo que permite presentar mucha información en forma sintética.

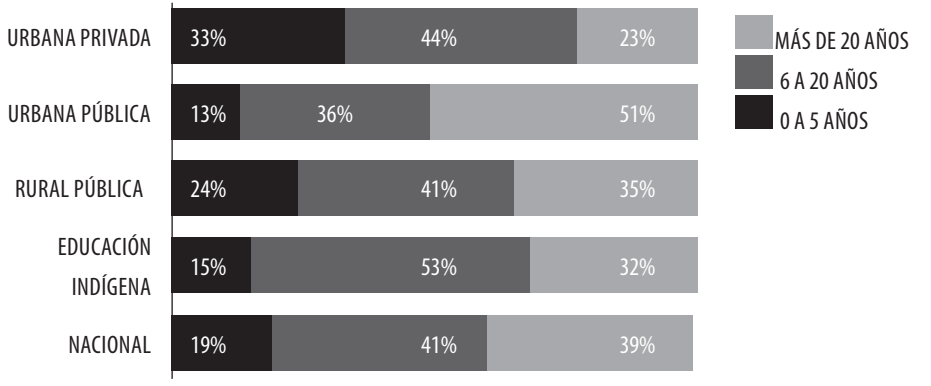
GRÁFICA 4.4. % MAESTROS QUE DESEAN CAMBIAR ESCUELA POR ANTIGÜEDAD Y ESTRATO.



FUENTE: INEE (2007). (2007: 26). GRÁFICA 1.3.

Otra forma de incluir más información en las gráficas de barras consiste en dividir cada barra en segmentos que se refieren a un segundo aspecto. En el ejemplo siguiente cada barra se refiere a los docentes que trabajan en cierto tipo de escuela, y los segmentos de cada barra a sus años de experiencia.

GRÁFICA 4.5. AÑOS DE EXPERIENCIA DOCENTE POR MODALIDAD DE PRIMARIA



FUENTE: INEE (2006: 128). GRÁFICA 3.13.

▪ **Uso inadecuado de las gráficas**

Pese a la impresión de claridad que da la visualización, la información que presenta una gráfica puede ser engañosa. Las gráficas de barras, por ejemplo, pueden dar una falsa impresión de la diferencia entre dos valores, con una escala que exagera o minimiza la diferencia, como puede apreciarse en el caso siguiente.

La Tabla 4.7 presenta las medias del desempeño en la escala global de Matemáticas de las pruebas PISA 2012, en algunas entidades de la República Mexicana.

TABLA 4.7. MEDIAS DE DESEMPEÑO DE ALGUNAS ENTIDADES, MATEMÁTICAS, PISA 2012

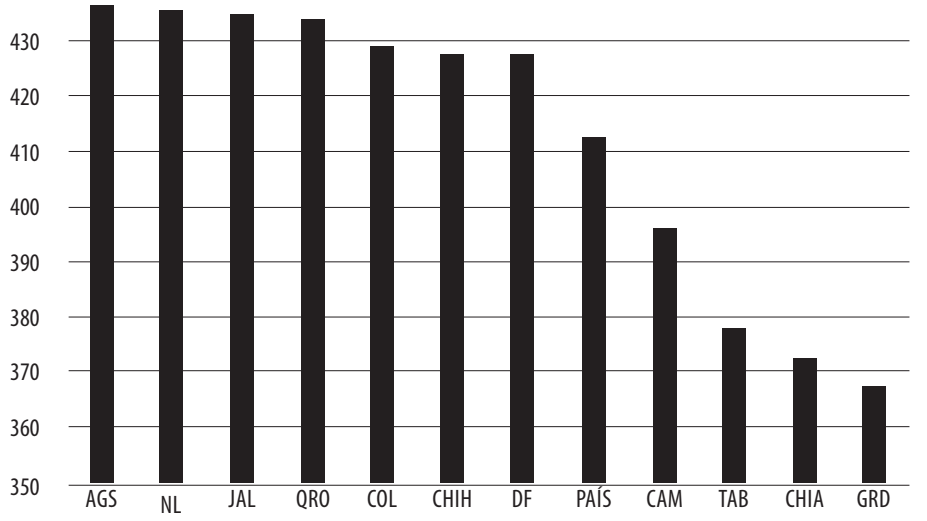
Entidad	Media	Entidad	Media
Aguascalientes	437	Media nacional	413
Nuevo León	436	Diez entidades más	412 – 402
Jalisco	435	Campeche	396
Querétaro	434	Tabasco	378
Colima	429	Chiapas	373
Chihuahua	428	Guerrero	367
Distrito Federal	428		
Ocho entidades más	424 – 414		

FUENTE: INEE (2013: 43). GRÁFICA 2.3.

La parte izquierda de la tabla presenta datos de siete entidades en las que los jóvenes evaluados obtuvieron un resultado significativamente superior a la media nacional, y en un solo renglón datos de ocho entidades más con resultados un poco superiores a la media nacional, en que la diferencia no alcanza a ser significativa. El primer renglón de la derecha ofrece el dato de la media nacional, los de diez entidades con resultados un poco inferiores con diferencias que no alcanzan a ser significativas respecto a la media nacional, y de cuatro entidades significativamente inferiores a dicha media.

Las dos gráficas siguientes, basadas en la Tabla 4.7, presentan esa información, pero la escala vertical de las barras de la Gráfica 4.6 está truncada, con valores que van solo de 350 a 450, mientras la escala de la Gráfica 4.7, correctamente, va de 250 a 750, lo que corresponde mejor a los valores máximos y mínimos que pueden alcanzar las medias en las pruebas PISA.

GRÁFICA 4.6. MEDIAS DE DESEMPEÑO DE UNAS ENTIDADES, MATEMÁTICAS, PISA 2012



FUENTE: INEE. (2013: 43). GRÁFICA 2.3.

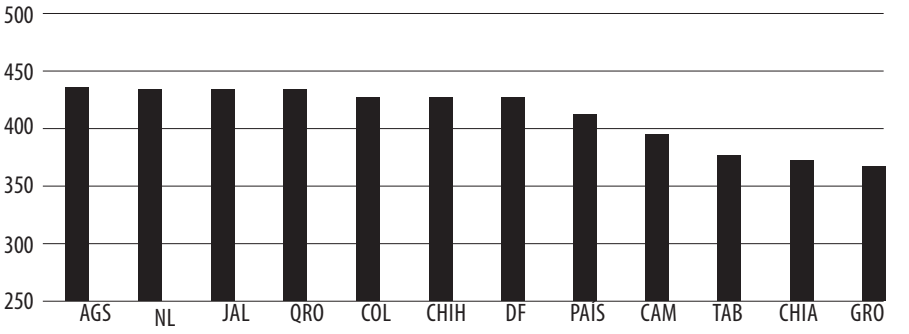
La escala truncada de la Gráfica 4.6 da una impresión distorsionada de la importancia de las diferencias entre entidades. Las distancias que parecen separar las medias de Guerrero, Chiapas y Tabasco, e incluso Campeche, parecen grandes, y claramente inferiores a la media nacional. Las medias más altas de Aguascalientes, Nuevo León, Jalisco y Querétaro parecen cercanas entre sí, como las de Colima, Chihuahua y el Distrito Federal (Ciudad de México), pero incluso las de las cuatro primeras parecen ligera, pero claramente superiores a las de las tres últimas.

Las mismas diferencias, con la escala correcta de la Gráfica 4.7, parecen pequeñas, como lo son realmente. Al explicar el contenido de la Tabla 4.7 se indicó que solamente las medias de las entidades mencionadas en el párrafo anterior, que son aquellas cuyo nombre aparece en la tabla, son significativamente mayores o menores a la media nacional, en tanto que las medias de las entidades que se agrupan en un solo renglón no alcanzan a ser significativamente distintas de la nacional.

Lo anterior se debe a que las medias de las pruebas PISA se basan en resultados de muestras de alumnos, por lo que son estimaciones que tienen un margen de error, como se verá en el punto 1.3 de este mismo capítulo. Por ello las diferencias respecto a la media nacional de los resultados de las entidades cuyo nombre no aparece en la Tabla 4.7 (ocho con medias que van de 424 a 414 puntos, y diez con

valores entre 412 y 402), en realidad no difieren significativamente de la media del país.

GRÁFICA 4.7. MEDIAS DE DESEMPEÑO DE UNAS ENTIDADES, MATEMÁTICAS, PISA 2012



FUENTE: INEE (2013: 43). GRÁFICA 2.3.

Si se presentan los resultados de los países participantes en PISA 2012 con la escala de la Gráfica 4.7 se podría ver que incluso los estados mexicanos con medias más altas están abajo de la media de la OCDE (494 en Matemáticas en PISA 2012), y muy lejos de los puntajes que alcanzan los mejores sistemas educativos del mundo, como Singapur (573), Hong Kong (561), Corea del Sur (554) o Finlandia (519).

Por otra parte, hay que advertir sobre otros riesgos que puede traer consigo el hacer gráficas visualmente atractivas. Introducir un efecto tridimensional, por ejemplo, *viola el principio central de las imágenes estadísticas (proporcionalidad del espacio y las cifras) y entrega una impresión distorsionada, exagerada e inexacta* (Best, 2009: 49). El uso de paquetes de software facilita esto, y para efectos publicitarios es atractivo, pero no es aceptable en informes académicos que pretendan ser rigurosos.

Me atrevo a sostener que el uso de gráficas vistosas es un recurso utilizado por investigadores poco consistentes para impresionar a lectores ingenuos.

Medidas sintéticas o resumen

Dos rasgos de una distribución de frecuencias, que una gráfica permite apreciar de manera general, pueden ser captados de manera más precisa por algunas *medidas sintéticas*: los puntos en que se concentra la mayor parte de los valores, y el grado en que la distribución muestra concentración o dispersión.

Uno de los más importantes precursores de la moderna metodología de investigación en ciencias sociales, John Stuart Mill, afirmaba: *una constante no es susceptible de explicación,* lo que quiere decir que *todo trabajo científico comienza con la observación de la variación*. (Weisberg, 1992: v)

Si no hubiera variación, si sólo hubiera uniformidad, no habría ciencia. De hecho, no habría conocimiento alguno, ni vida. Esta comienza con la existencia de diferencias, de múltiples realidades interconectadas entre sí, pero diferentes; de innumerables facetas de la realidad indistintas en sí, pero distinguibles para la mente humana. La ciencia comienza con la identificación de algunos de esos aspectos indistintos pero distinguibles. En otras palabras, con la definición de variables y su estudio. Pero... ¿qué se puede estudiar de las variables?

Las diferencias y la tipicidad. Las particularidades y las generalidades. Lo individual y lo común. Dos caras, indisociablemente unidas, de la realidad. Cada ser es diferente, individual, único en cierto sentido. Pero, a la vez, todos tienen algo en común, todos comparten rasgos, participan de generalidades más amplias. La conocida expresión que dice que, a veces, *los árboles no dejan ver el bosque* se refiere a esta dicotomía: si se presta demasiada atención a los individuos particulares -los árboles- se presenta el riesgo de perder de vista el conjunto, la globalidad, el bosque. Pero también puede ocurrir lo contrario: que por captar el conjunto se pierdan de vista los detalles particulares, que sea el bosque el que impida ver los árboles.

El problema del estudio riguroso de ciertos aspectos de la realidad, del análisis científico de ciertas variables, consiste en buscar la forma de tener una visión del todo (de un todo relativo, ciertamente) sin perder de vista las peculiaridades de sus partes. De ver el bosque sin olvidar los árboles, y a estos últimos sin perder de vista aquél.

En términos más técnicos, al estudiar cada variable nos preocuparemos por estudiar ciertas *expresiones numéricas o medidas sintéticas* que presenten en forma resumida, pero representativa, al todo (medidas de tendencia central o promedios) y otras que reflejen la homogeneidad o heterogeneidad, el grado de variación que se da en el seno del todo que estemos analizando (medidas de dispersión).

Un promedio da una aproximación de la tendencia mayoritaria de un conjunto de valores, resumiendo lo *general* de la distribución de una variable. Si se dice que el promedio de la estatura de unas personas es de 170 cm (o 1.70 m), basta para saber que no se trata de niños de primaria ni de jugadores profesionales de basketbol. Sin embargo, varias distribuciones pueden tener el mismo promedio y ser diferentes. Un promedio de estatura de 1.70 m. puede resultar de un grupo *homogéneo*, cuyos inte-

grantes tienen todos una estatura muy similar y cercana a 1.70 m. El mismo promedio puede encontrarse en un grupo *heterogéneo*, con integrantes muy altos y muy bajos, que en promedio miden también 1.70 m. Esto lleva a preguntarse si es posible estimar también la homogeneidad o variabilidad de una distribución.

La respuesta es afirmativa, pero la estimación de una medida de tendencia central, o de homogeneidad o dispersión, se deberá hacer de manera diferente según el nivel de medición de las variables de que se trate, nominal, ordinal o métrico.

En las décadas de 1960 y 1970, impulsado por John W. Tucker (1962, 1977), se desarrolló un tipo de análisis de datos estadísticos que subraya la importancia de la etapa descriptiva, para explorar características generales de las distribuciones, antes de hacer análisis más complejos, con herramientas gráficas y estadísticas simples, pero robustas, que constituyen el Análisis Exploratorio de Datos (*Exploratory Data Analysis*, EDA) (*Cfr*. Palmer Pol, 1999).

▪ Medidas sintéticas para variables nominales

Tendencia central

Cuando se trabaja con variables medidas a nivel nominal, o sea con categorías en las que se clasifica a los sujetos según algún rasgo, pero que no tienen entre sí una relación de orden, y menos de diferencia proporcional de su tamaño, no hay mucho que hacer, aunque convencionalmente se designen las categorías con números, como al codificar con el número uno los sujetos varones, y con cero las mujeres. En esta situación una medida de tendencia central responde la pregunta ¿de cuál *valor* o categoría hay más casos? Y la respuesta es un valor que se denomina modal o **moda.** Siguiendo con el ejemplo, si se pregunta si hay más varones que mujeres o viceversa, la respuesta es simplemente el número de sujetos del sexo mayoritario.

Dispersión

Si la medida de tendencia central es la moda, la medida de dispersión más sencilla para variables nominales es, simplemente, la proporción de casos no modales:

- Razón de Variación: 1 - (f modal/N).

Otras medidas (Weisberg, 1992: 67-73):

- Índice de diversidad (D): se obtiene sumando el cuadrado de la proporción de casos que hay en cada valor de la distribución, y restando el total de 1.

- Índice de variación cualitativa (IQV): es una variante normalizada del anterior para que su valor sea 1 cuando hay la máxima dispersión posible; se obtiene dividiendo D entre el valor máximo que es igual al número de categorías menos 1 sobre el número de categorías k-1/k.

▪ Medidas sintéticas para variables ordinales

Tendencia central

Con medidas ordinales es posible ordenar sujetos de mayor o menor, o viceversa, y en este caso es posible preguntarse qué *valor* divide a la población en dos partes iguales. El valor que responde esta pregunta se denomina la **mediana.**

En un ordenamiento es posible identificar los valores que dividen a los sujetos en cuatro partes, llamadas **cuartiles**. El primer cuartil (Q1) es el valor superior a 25% de los casos de la distribución, e inferior a 75% restante. El tercer cuartil (Q3) es el valor superior a 75% de los casos, e inferior a 25%. El segundo cuartil (Q2) es lo mismo que la mediana, ya que al ser superior a 50% e inferior al 50% de los casos, divide a la distribución en dos partes iguales. De manera similar, en una distribución de valores ordinales se pueden identificar los que la dividan en cinco partes iguales (quintiles); en diez (deciles), en cien (centiles, percentilas) u otro número (cuantiles).

Medidas de tendencia central para datos ordinales (Weisberg, 1992: 38-40):

- Promedio de valor mínimo y máximo (*midextreme-midrange*): mínimo + máximo/2
- Promedio de cuartiles 1 y 3 (*midhinge*): Q1 + Q3/2
- Promedio de cuartiles y mediana (*Best Easy Systematic Estimate*): Q1 + 2Q2 + Q3/4
- Promedios de otros cuantiles (*midsummaries*): Q° superior + Q° inferior/2

La estimación de cualquiera de los promedios anteriores puede arrojar un valor que no coincida exactamente con uno de la distribución, sino que se sitúe entre dos de ellos. Al calcular la mediana, por ejemplo, si el número de casos es impar, habrá un valor que divide en dos partes iguales la distribución, pero si el total de casos es par, la mediana se situará entre dos valores, el superior de la mitad inferior de la distribución, y el inferior de la mitad superior.

Dispersión

No suelen calcularse medidas de dispersión de datos ordinales pero puede hacerse, y dado que hay un orden entre los valores, consisten básicamente en res-

ponder la pregunta de a qué distancia se encuentran entre sí valores más o menos alejados del centro, definido por la mediana, lo que se denominan *rangos o recorridos:*

- Rango simple: valor máximo - valor mínimo.
- Rango intercuartílico (IQR): Valor del cuartil 3 menos valor del cuartil 1 = Q3 - Q1.
- Rango semi-intercuartílico o desviación cuartílica (QD): IQR/2.
- Coeficiente de variación cuartílica (CQV): Q3 - Q1/ Q3 + Q1.

▪ Medidas sintéticas para variables de nivel cardinal o métrico

Tendencia central

La medida de tendencia central más conocida, a la que remite la palabra *promedio* si no se precisa otra cosa, es la *media aritmética:* el resultado de dividir la suma de valores de todos los sujetos de una distribución entre el número total de sujetos.

La moda y la mediana también pueden usarse con variables medidas a nivel de intervalo o de razón (cardinales o métricas), pero la media aritmética es más precisa, y tiene la ventaja de que permite utilizar procedimientos estadísticos que no pueden usarse con variables nominales ni ordinales.

El uso de la media aritmética puede presentar también problemas, en especial por su sensibilidad a valores extremos, a los que la mediana y la moda no son sensibles. Si se trata de un número considerable, y la distribución es simétrica y sin casos extremos, la media, la mediana y la moda serán idénticas o muy similares; pero un solo caso sumamente alejado del resto de los valores basta para que la media cambie bastante, mientras la mediana y la moda no se ven afectadas.

Imaginemos un grupo de estudiantes de licenciatura, de los que tres tienen 18 años; cinco 19; siete 20; cinco 21 y tres 22. En total tenemos 23 sujetos, con un total de 460 años y una media aritmética de 20 (460/23). La media sería exactamente igual a la mediana (20 años), la edad del alumno 12 del ordenamiento, que tiene 11 antes y 11 después; la moda, el valor más frecuente, sería también 20 años.

Pero si uno solo de los estudiantes del grupo, el de más edad, tuviera 45 años en vez de 22, el total de años del grupo sería de 483, y la media pasaría a ser de 21 años, en tanto que la mediana y la moda seguirían siendo de 20 años, ya que la edad del 12° estudiante del ordenamiento no se habría modificado, ni la edad modal; en ambos casos seguirían siendo de 20 años.

El ejemplo anterior muestra la importancia de usar medidas robustas o resistentes, lo que se refiere a su insensibilidad *a la presencia de datos extremos o, incluso, errores altos de medición,* o bien como *la insensibilidad de un estadístico ante el cambio de una parte pequeña del conjunto de datos.* (Escobar, 1999: 8, 32)

Ante la fragilidad de la media, se proponen otros promedios, que tienen en común eliminar o modificar unos valores extremos (Weisberg, 1992: 40-41):

- Media de la mitad central de los datos (*midmean*): que simplemente elimina 25% de los casos de los dos extremos de la distribución.
- Media recortada o despuntada (*trimmed mean*): que elimina proporciones menores de casos extremos, por ejemplo 2% más alto y más bajo, 5%, etc.
- Media de Windsor (*winsorized mean* por el apellido de su autor, Charles P. Winsor): un promedio en que los valores de una proporción de datos extremos (2%, 5%) no se eliminan, como en la media despuntada, sino que se modifican, igualándolos a los valores de los datos de un porcentaje similar (2%, 5%) inmediatamente anterior.

Dispersión

La operación básica para medir la dispersión es promediar la diferencia de cada valor respecto a la media, pero hay variantes:

- Desviación media (*mean deviation, average deviation*): promedio de los valores absolutos de las diferencias respecto a la media, que de esta manera no se anulan.
- Varianza ($\sigma 2$): promedio de los cuadrados de las desviaciones respecto a la media, con lo que tampoco se anulan, pero el valor resultante no está en las unidades de los valores originales, por haber elevado al cuadrado las diferencias.
- Desviación estándar ($\sigma = \sqrt{\sigma 2}$): raíz cuadrada de la varianza, con lo que el valor vuelve a estar en unidades originales y dificulta la comparación de distribuciones diferentes: $\sigma = 100$ es grande si se trata de centímetros de estatura, pero pequeña si se trata de pesos de ingreso mensual de los sujetos.
- Coeficiente de variación (CV, σ/x): al dividir la desviación estándar entre la media se obtiene un valor normalizado, o sea que no depende de los valores originales y permite comparar la dispersión de diferentes distribuciones.

▪ Variables dicotómicas reales o ficticias (dummy)

Ya se mencionó el caso de variables dicotómicas, en que la presencia de uno de los valores implica ausencia del otro. En este caso es usual codificar con 1 la presencia de la característica en cuestión, y con 0 su ausencia, y el porcentaje de valores positivos en el total es la media de la distribución.

Una variable con mayor número de valores (politómica) puede transformarse en otras tantas variables dicotómicas ficticias (*dummy*), cada una con presencia o ausencia de uno de los valores de la variable politómica. Si en una encuesta hay 32 valores posibles de la variable lugar de nacimiento, considerando las 32 entidades federativas de México, se pueden construir 32 variables dicotómicas *dummy* del tipo "Nativo de Aguascalientes: SI-NO"; "Nativo de Baja California: SI-NO", etc.

De esta forma se transforman datos de una variable con n valores de nivel nominal en n variables dummy con valores 0 y 1, lo que permite aplicar estadísticas que suponen un nivel de medición cardinal o métrico.

En este caso la varianza es el producto del porcentaje de casos en que la variable está presente (valor p) por la de casos en que está ausente (valor q); esos porcentajes son complementarias, sumando obviamente 100% en todos los casos: $\sigma 2 = p(1-p)$ o bien p.q. La desviación estándar σ es la raíz cuadrada de la varianza tal como se acaba de definir. La dispersión es máxima cuando la proporción de presencias y ausencias es 50% respectivamente. En este caso la media será de .5 (o 50%), la varianza de .25 (.5 x .5) y la desviación estándar de .5 ($\sqrt{.25}$).

▪ Representaciones gráficas

Tendencia central y dispersión son características básicas de una distribución, pero no las únicas. También puede preguntarse si los valores de la distribución se agrupan simétricamente a los dos lados de la media, o si hay más a uno u otro lado, o bien qué tan grande es la proporción de valores concentrados alrededor de la media, que hacen que el histograma que representa la distribución parezca más apuntado o plano. Hay medidas sintéticas que captan la simetría (sesgo) y el apuntamiento (curtosis) de una distribución, pero su uso es raro, y es más práctico apreciar visualmente esos rasgos de las distribuciones aprovechando algunas gráficas.

▪ Histogramas

Al inicio del capítulo se presentaron las gráficas de barras. Ahora se amplía esa presentación, para tener en cuenta las características que distinguen una dis-

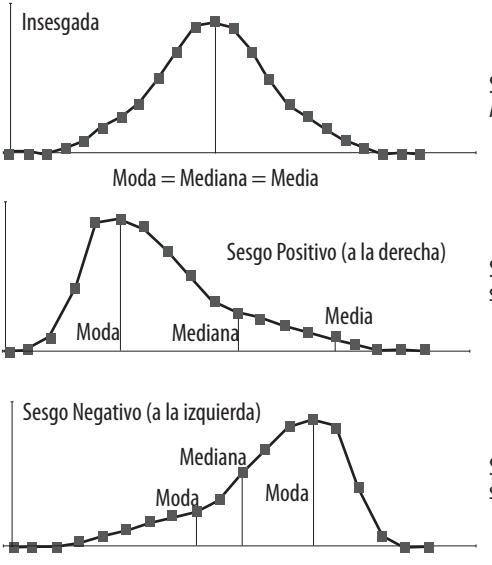
tribución, además de la tendencia central y la dispersión, o sea el sesgo y el apuntamiento.

Debe añadirse que el referente implícito o explícito de toda distribución es la curva llamada normal que —como se presentará con mayor detalle más abajo— se caracteriza por ser simétrica (sin sesgo) y no demasiado apuntada ni achatada (*mesocúrtica*). En la medida en que se alejen de dicha "normalidad" tendremos distribuciones sesgadas a la derecha o a la izquierda, o bien más apuntadas (*leptocúrtica*) o más achatadas (*platicúrtica*).

Las gráficas siguientes son ejemplos de estos tipos de distribuciones.

GRÁFICA 4.8. DISTRIBUCIÓN SIMÉTRICA (INSESGADA) Y CON SESGO POSITIVO Y NEGATIVO

TIPOS DE ASIMETRÍAS EN LA DISTRIBUCIÓN



Si la distribución es simétrica o insesgada A_s será igual a 0

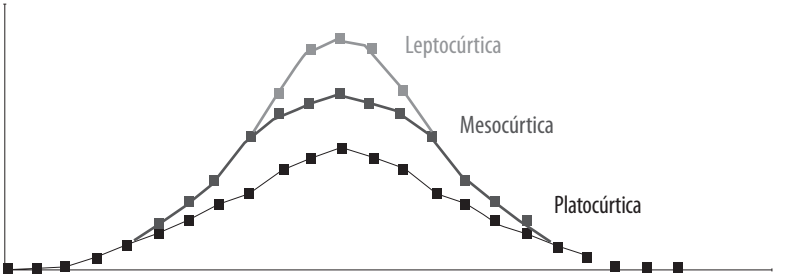
Si la distribución es asimétrica o con sesgo positivo, A_s será mayor que 0

Si la distribución es asimétrica o con sesgo negativo, A_s será menor que 0

MSC. ECON. ALEXANDER NUÑEZ MEDIDAS DE DISTRIBUCIÓN: ASIMETRÍAS Y CURTOSIS

FUENTE: HTTPS://BIT.LY/2TOUPQo.

GRÁFICA 4.9. DISTRIBUCIÓN APUNTADA, NORMAL Y ACHATADA
MEDIDAS DE APUNTAMIENTO O CURTOSIS



SI LA DISTRIBUCIÓN ES NORMAL (MESOCÚRTICA), EL ÍNDICE VALE 0
SI LA DISTRIBUCIÓN ES LEPTOCÚRTICA, EL ÍNDICE ES MAYOR QUE 0
SI LA DISTRIBUCIÓN ES PLATOCÚRTICA, EL ÍNDICE ES MENOR QUE 0

MSC. ECON. ALEXANDER NUÑEZ MEDIDAS DE DISTRIBUCIÓN: ASIMETRÍAS Y CURTOSIS

FUENTE: HTTPS://BIT.LY/2U1O5SG.

▪ Diagramas de dispersión unidimensionales

Una forma sencilla de visualizar los valores de una distribución consiste en representar cada valor mediante un símbolo (usualmente un pequeño círculo), que se sitúa en una gráfica de una sola dimensión (generalmente horizontal) que contiene una escala cuyos extremos coinciden con los de los valores de la distribución. Estas gráficas solo funcionan bien con un número reducido de valores, ya que con números grandes hay muchos valores que se traslapan, haciendo confusa la imagen. Esto puede minimizarse desplazando ligeramente los puntos en el eje perpendicular al de la escala (lo que se designa con la expresión *jittering*). Por esta razón este tipo de gráfica es poco utilizado.

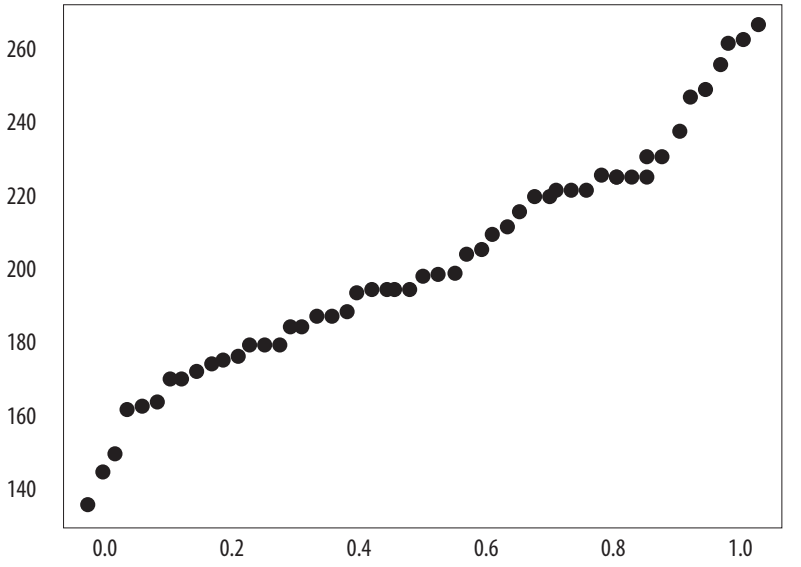
▪ Gráficas de cuantiles

Estas gráficas permiten visualizar una distribución univariada de valores ordinales en un espacio de bidimensional, en el que la escala del eje vertical representa los valores de los datos, y la del eje horizontal los cuantiles. La imagen es una serie de puntos, en orden siempre creciente (monótona), que grafican la distribución empírica acumulada (*empirical cumulative distribution function*, ECDF) de la variable.

El ejemplo siguiente presenta la distribución del puntaje de calidad que obtuvieron en el programa Medicaid en 1986 los 50 estados norteamericanos y el Distrito de

Columbia. Los puntajes (eje vertical) fueron de un mínimo de 133 puntos a un máximo de 264, y el valor p en el eje de cuantiles fue de .010 a .99.

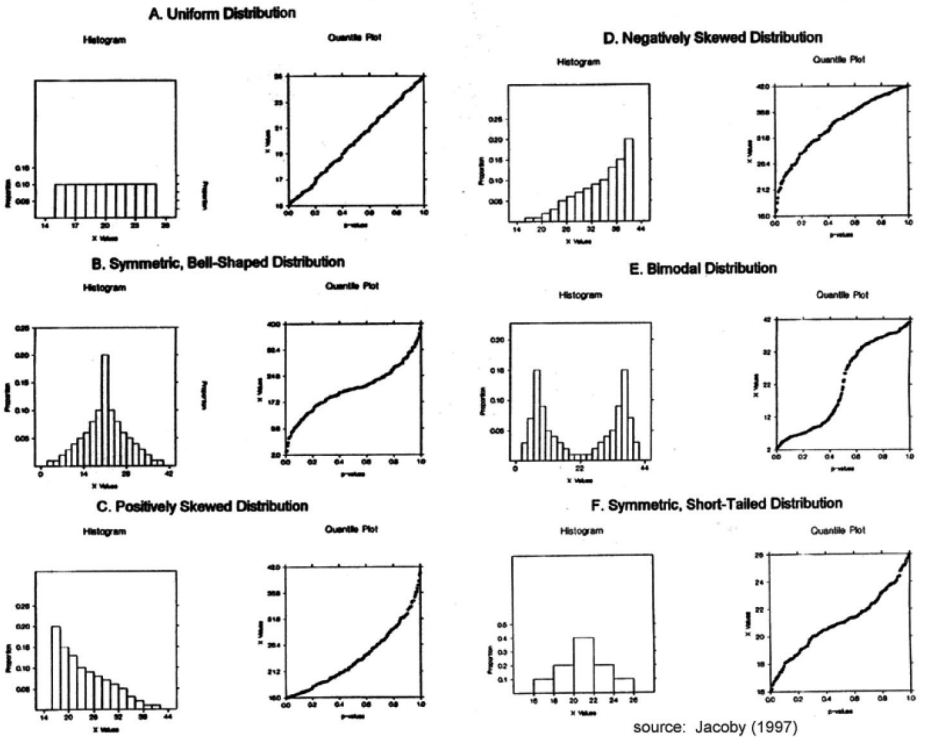
GRÁFICA 4.10. GRÁFICA DE CUANTILES DE LOS PUNTAJES DE CALIDAD MEDICAID, 1986



FUENTE: JACOBY, (1997: 39). FIGURE 2.11.

La Gráfica 4.11 tiene ejemplos hipotéticos de varias distribuciones (uniforme, simétrica campaniforme, con sesgo positivo o negativo, bimodal, simétrica con extremos cortos), mostrando el histograma y la gráfica de cuantiles correspondiente; se puede apreciar la forma característica que toma en cada caso la distribución empírica acumulada.

GRÁFICA 4.11. COMPARACIÓN DE HISTOGRAMAS Y GRÁFICAS DE CUANTILES DE VARIAS DISTRIBUCIONES DE DATOS QUE PRESENTAN DISTINTA FORMA

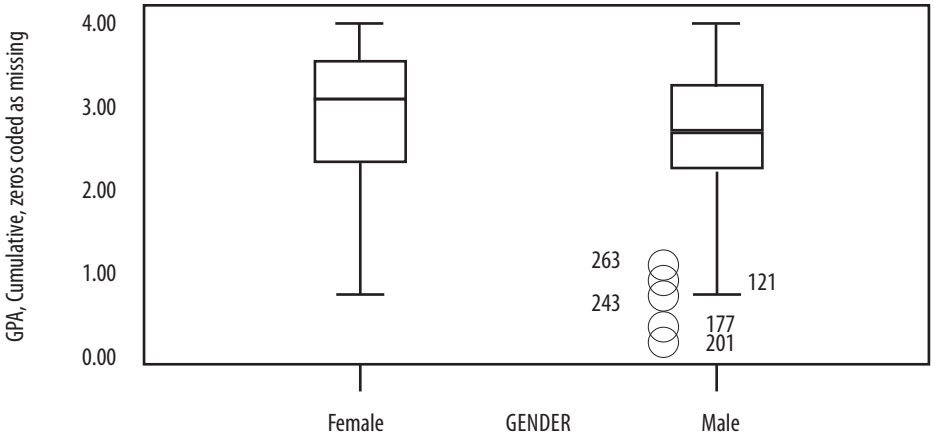


FUENTE: JACOBY, (1997: 34-35). FIGURE 2.10.

▪ **Gráficas de caja (box & whishkers plots)**

Las gráficas de caja (o de caja y bigotes) permiten apreciar visualmente la tendencia central y la dispersión de una distribución de datos ordinales. Estas gráficas muestran la ubicación de la mediana (Q2) y de los cuartiles 1 y 3 (Q1, Q3). Estos cuartiles se marcan con líneas perpendiculares a la escala, conformando un rectángulo (la caja) cuyos extremos son los cuartiles 1 y 3 y la línea central la mediana. La escala permite apreciar fácilmente el rango intercuartílico. Dos líneas que inician en los extremos de la caja (Q1 y Q3) constituyen los *bigotes*, y la pequeña línea horizontal en que terminan (las cercas o fences) marcan valores situados hasta 1.5 veces el valor del recorrido intercuartílico (IQR) por debajo de Q1 y por encima de Q3. Los puntos entre 1.5 y 3 veces o más el recorrido intercuartílico (IQR) por debajo de Q1 o por encima de Q3 representan valores atípicos y extremos (*outliers* & *extreme outliers*) de la distribución.

GRÁFICA 4.12. PROMEDIO DE CALIFICACIONES POR GÉNERO



FUENTE: VOGT, VOGT, GARDNER Y HAEFFELE, (2014: 214). FIGURA 6.3.

Puede apreciarse que la distribución del promedio de calificaciones (*Grade Point Average*, GPA) de hombres y mujeres es similar, pero no idéntica. La mediana de mujeres es mayor que la de varones. En ambos casos la mediana no está al centro de la caja definida por Q1 y Q3 (percentiles 25 y 75), indicando que las distribuciones no son simétricas. El recorrido intercuartílico de las mujeres es mayor que el de los varones. Las "cercas" se sitúan en puntos similares, pero en las mujeres no hay valores atípicos (*outliers*), mientras que en el caso de los varones hay varios en la parte inferior de la distribución. (*cfr*. Vogt, Vogt, Gardner y Haeffele, 2014: 214-215)

Patrones y tendencias

Otro tipo de análisis descriptivo consiste en la identificación de ciertos patrones en el comportamiento de una variable, los valores que toma lo largo del tiempo. Puede haber una tendencia regular o irregular; plana, si no hay cambios y los valores de la variable son iguales o muy similares; creciente o decreciente, con valores cada vez más grandes o más pequeños; ondulatoria u oscilatoria, si aumentan y disminuyen.

Es posible detectar patrones de este tipo observando las gráficas de tipo cartesiano que se pueden construir con los valores que toma una variable a lo largo del tiempo, representando el tiempo se representa en abscisa (eje horizontal, de las x), y los valores de la variable de que se trate en ordenada (eje vertical, de las y). Algunas gráficas típicas de este tipo son:

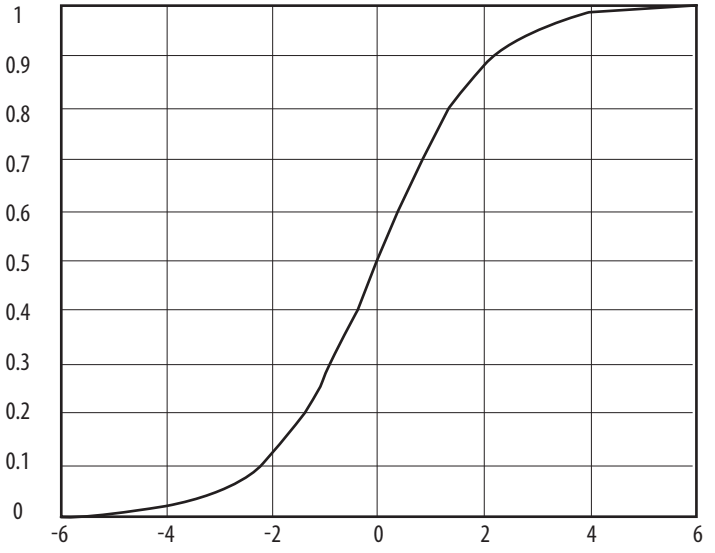
- Una línea horizontal refleja una tendencia plana, de una variable con valores iguales una y otra vez a lo largo del tiempo.
- Una línea recta con inclinación ascendente (o descedente) refleja tendencia creciente (o decreciente) de variable con incremento (o decremento) regular.
- Una semiparábola invertida muestra tendencia creciente que va acelerando; la misma gráfica en sentido horizontal muestra desaceleración.

Un ejemplo más complejo es el de fenómenos que se explican por contagio, como las epidemias. La velocidad a la que se extiende una epidemia al principio es lenta, aumenta hasta alcanzar un máximo, y luego disminuye. Esto se explica porque dicha velocidad de propagación es el producto del número de sujetos ya infectados por el número de los aún no infectados, pero susceptibles de serlo.

Al inicio de una epidemia con un primer caso (digamos 1%), aunque haya muchos (99%) susceptibles de contagiarse (no inmunizados), la velocidad de propagación no puede ser muy rápida porque solo hay un caso activo. La velocidad de propagación aumenta a medida que lo hace la proporción de contagiados y los sujetos por contagiar disminuyen. Si la proporción de contagiados es p, y la de casos por contagiar 1 — p (q), el producto p x q es máximo cuando p = q. El producto p x q aumenta (1 x 99, 2 x 98, 5 x 95, 10 x 90, 20 x 80...) hasta 50 x 50; luego disminuye 51 x 49, 60 x 40, 70 x 30, 90 x 10, 95 x 5, 98 x 2, 99 x 1, 100 x 0. La velocidad de propagación va en aumento, hasta un máximo en que el número de contagiados y por contagiar es igual: 50%. A partir de ese punto, aunque haya muchos sujetos contagiados, hay cada vez menos por contagiar y la velocidad de propagación va disminuyendo, hasta el momento en que toda la población ha sido contagiada.

La tendencia se puede visualizar con una función logística, cuya representación gráfica (acumulada) tiene forma de letra S.

GRÁFICA 4.13. FUNCIÓN LOGÍSTICA



Esta tendencia se da en la difusión de rumores, que son un caso de contagio (Monin, Benayoun y Sert, 1973: 15).

La inferencia estadística

El conocimiento humano no se reduce a percepciones; siempre hay un componente de inferencia para pasar de la información sensorial a la interpretación. Un tipo de inferencia, llamada estadística, se refiere a la generalización a una población mayor de lo que se encontró en un subconjunto de ella, en una *muestra*.

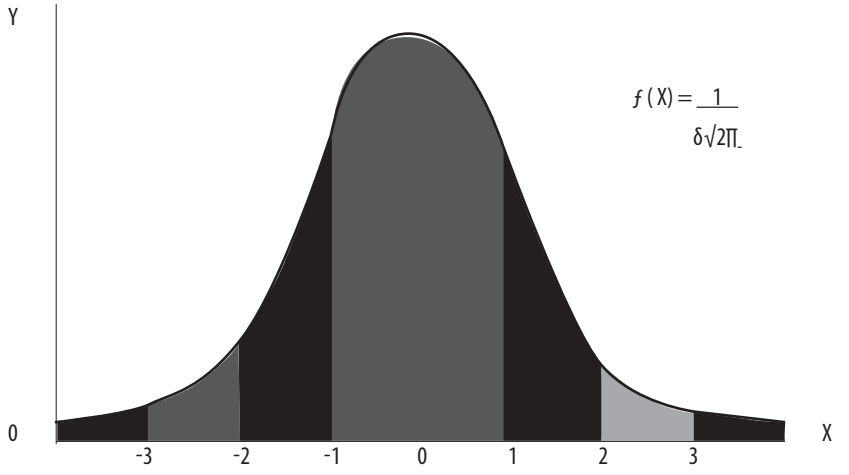
Los valores obtenidos a partir de una muestra no serán tan precisos como los que resultarían de una medición en toda la población; es posible que una muestra arroje valores cercanos a los de la población, en tanto que otra produzca valores bastante distintos. De allí la necesidad de estimar la probabilidad de uno u otro resultado, lo que exige aplicar nociones derivadas de las propiedades de las distribuciones de frecuencias, en especial las de la distribución normal, para estimar la probabilidad de que coincidan o no con los de la población, de que sean o no *significativamente distintos* de ellos; puede también estimarse la importancia del tamaño del efecto, el margen de error de los valores inferidos y el intervalo de confianza.

La curva normal y su interpretación

La *curva de Gauss* es la representación gráfica de una distribución de valores que tiene forma de campana; con un eje central definido por la media; a cuyos lados se distribuyen los valores en forma simétrica respecto a la media; y asintótica en sus extremos respecto al eje horizontal, según la desviación estándar.

La Gráfica 4.14 muestra un ejemplo de curva normal, así como la función matemática que la define, en la que se puede apreciar que, además de las constantes π y e, la ecuación solo incluye dos variables: la media μ y la desviación estándar σ. En el eje de las x se marcan los valores de 1, 2 y 3 σ a un lado y otro de la media, a las que corresponden las áreas sombreadas bajo la curva.

GRÁFICA 4.14. REPRESENTACIÓN GRÁFICA Y ECUACIÓN DE LA CURVA NORMAL



En la distribución normal, entre la media y una desviación estándar, a un lado u otro, se agrupa 34.13% de los valores. Entre una desviación estándar más y una menos hay 68.26% del total de casos. Una σ adicional añade 13.59% valores, por lo que en el espacio entre dos σ arriba y dos bajo la media hay 95.44% de ellos. Otra σ añade 2.14% a cada lado, y una cuarta 0.13%; en el rango de 3 s arriba y abajo de la media cae 99.72% de los casos, y en el de cuatro s arriba y abajo 99.98%.

La diferencia de un valor de la distribución respecto a la media se puede expresar en las unidades en que esté medida la variable, o en desviaciones estándar. Si la media

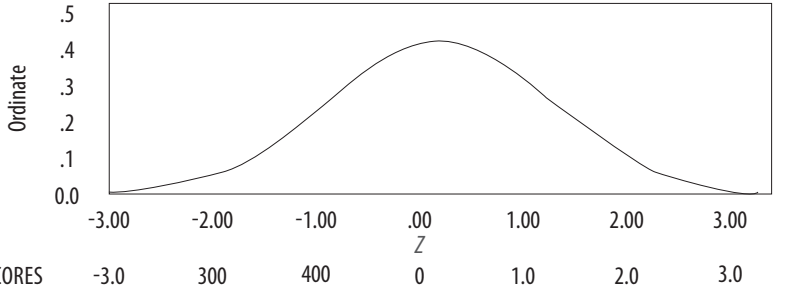
de estatura de un grupo es 170 cm, y la s 15 cm, la posición de un sujeto que mida 185 cm se sitúa 15 cm, o una desviación estándar, a la derecha de la media. Esta expresión en términos de la s se conoce como *valor z*, que es una medida de posición análoga a los cuantiles de una distribución de valores ordinales.

Con base en la curva normal pueden construirse diferentes puntajes estandarizados definiendo arbitrariamente un valor para la media y otro para la desviación estándar.

Los valores del Cociente Intelectual (IQ) se estandarizaron con media de 100 y s de 15. Un valor de tres σ por arriba de la media (145 puntos) o más solo será obtenido por entre uno y dos de cada mil sujetos. Por ello se considera genios a personas con un IQ de 145 o más. Las pruebas SAT y GRE del *Educational Testing Service* y el College Board se estandarizan con media de 500 y desviación estándar de 100, por lo que los resultados varían entre 200 a 800 puntos. Las pruebas del *American College Testing* tienen media de 20 y desviación estándar de 5. Las pruebas PISA, al igual que las llamadas PLANEA del Instituto Nacional para la Evaluación de la Educación, que sustituyeron a las pruebas Excale del INEE, y a las ENLACE que aplicaba la SEP, usan también escalas de media 500 y desviación estándar 100.

La gráfica siguiente presenta algunos puntajes estandarizados con base en la curva normal, mostrando la equivalencia con los puntajes z, y también con una medida de posición usual con una escala ordinal, los percentiles.

GRÁFICA 4.15. PUNTAJES ESTANDARIZADOS BASADOS EN LA CURVA NORMAL



Z. SCORES	-3.0	300	400	0	1.0	2.0	3.0
SAT & GRE	200	30	40	500	600	700	800
T SCORE	20	70	85	50	60	70	80
IQ	55	10	15	100	115	130	145
ACT	5	2	16	20	25	30	35
PERCENTILES				50	84	98	

FUENTE: VOGT, (2007: 24). FIGURA 2.1.

Además de entender las características matemáticas y la representación gráfica de la curva normal, para comprender su sentido sustantivo, que la hace tan usada en muchas áreas de investigación, hay que entender que permite modelar fenómenos que tienen un carácter aleatorio, o sea que no son el resultado de la influencia de uno o pocos factores bien identificados, sino de un número indefinido de factores no identificados e independientes, cuya influencia se mezcla en forma no sistemática.

Ejemplo de distribución normal: distancia al blanco de gran número de disparos. Un número reducido de ellos impacta en el centro del blanco o muy cerca de él, otro pequeño número muy lejos, y la mayoría ni tan cerca ni tan lejos, impacta a una distancia media, lo que refleja el efecto de muchos factores independientes que se combinan de manera aleatoria: cantidad exacta de pólvora de cada proyectil y su peso, fatiga del que apunta, el viento, etc. Si un factor particular cambia —por ejemplo, un tirador distinto, o una nueva caja de municiones de otra marca— el resultado será una nueva distribución de los impactos que será también normal, por la influencia aleatoria de muchos factores, pero el cambio en la tendencia general no será debido al azar, sino al efecto de la variable particular que se modificó.

Desde principios del siglo XIX, con Quetelet, la curva normal comenzó a usarse para representar la distribución de la estatura y, en general, de ciertos rasgos físicos de las personas, e incluso de disposiciones y actitudes. La estatura, en efecto, puede verse afectada sistemáticamente por factores como edad, género, estado de salud, nutrición o ciertos rasgos étnicos. Los niños tienen una estatura menor que los adultos, las mujeres que los varones, las personas mal alimentadas que las bien nutridas, y la de cierto grupo étnico en comparación con la de otros grupos. Pero tratándose de una población homogénea en cuanto a esos y otros factores precisos, la estatura depende de un número desconocido, sin duda grande, de factores de tipo genético y/o ambiental, cada uno de los cuales influye poco, y en conjunto los hacen de manera no sistemática, *aleatoria*, por lo que se distribuye según la curva normal. Con un grupo no homogéneo, por ejemplo, si se mezclan adultos de sexo masculino y femenino, la distribución no será normal, sino bimodal. Las dos "jorobas" de la distribución muestran que, en realidad, se trata de dos poblaciones, y que la estatura de cada una se distribuye normalmente.

La estadística inferencial parte de la idea de que, si los valores de una variable de cierta población se distribuyen normalmente, es posible calcular la probabilidad de que extrayendo al azar cierto número de casos los valores que se obtengan coincidan con los de la población, con cierto margen de error y cierta probabilidad.

Pruebas de hipótesis o de significatividad estadística

En el campo de la psicología, hace ya casi un siglo, se intentó determinar la solidez de una generalización de resultados obtenidos de una muestra al conjunto de la población de la que forma parte, mediante la llamada *prueba de hipótesis*.

Si se quiere conocer la estatura de los niños de un sistema educativo, o su dominio de aritmética, se puede medir la variable de interés en una muestra, partiendo del supuesto de que es representativa de la población, que los valores de la muestra y de la población son iguales, no difieren, la diferencia es cero (*hipótesis nula*).

Una vez hecha la medición en una muestra aleatoria y obtenido cierto resultado, la estadística inferencial permite estimar la probabilidad (*valor p*) de que ese resultado, con una muestra de ese tamaño, coincida con el de la población (o que los dos no sean distintos, que su diferencia sea igual a cero).

Si la probabilidad de que tal cosa ocurra es alta (*no se puede rechazar la hipótesis nula*), lo más seguro es que el hecho de que el valor obtenido a partir de la muestra coincida con el de la población se debe a simple coincidencia, al azar. Pero si la probabilidad de que los valores de muestra y población coincidan es baja, *se puede rechazar la hipótesis nula*: un valor p muy bajo implica que en la muestra hay algo no debido al azar que la hace distinta de la población en ese aspecto.

Estas formulaciones, con dos o más negativas, son complicadas: la hipótesis es que **NO** hay diferencia entre el valor derivado de la muestra y el de la población (*hipótesis nula*, *Ho*), y se calcula la probabilidad de que eso **NO** ocurra.

- Si la probabilidad de que tal cosa ocurra es baja se rechaza la hipótesis nula, lo que lleva a pensar que *sí hay diferencia*. Nótese que la expresión usual es confusa: *se rechaza* la hipótesis de *que no hay* tal diferencia (NO x NO = SÍ).
- Si la probabilidad es alta, *no* se puede *rechazar* la hipótesis nula de *que no hay* diferencia significativa, lo que en forma menos confusa equivale a decir que probablemente no la hay. (NO x NO x NO = NO).

Se pueden cometer dos tipos de error al poner a prueba una hipótesis nula:

- Tipo I: rechazar una H0 que en realidad es verdadera. Se trata de los casos llamados *falsos positivos*: concluyo que el paciente tiene cierta enfermedad cuando en realidad no la tiene.

- Tipo II: aceptar una H0 que en realidad es falsa. Los *falsos negativos*: concluir que el paciente no tiene la enfermedad cuando en realidad sí la tiene.

Muchas veces se prefiere evitar el riesgo de errores Tipo I, lo que en el caso de un diagnóstico médico implicaría multiplicar los estudios previos, con altos niveles de exigencia y márgenes de error reducidos, que llevarían a interpretar como datos de la presencia de la enfermedad incluso indicios leves de ello. Así será difícil que un enfermo real pase desapercibido (error Tipo I), pero aumenta el riesgo de considerar enfermas a personas sanas (error Tipo II).

Es deseable evitar o minimizar la probabilidad de cometer ambos tipos de error, pero con un nivel de precisión de las mediciones y otras circunstancias iguales no es posible minimizarlos simultáneamente con una muestra de cierto tamaño; esto puede lograrse con una muestra mayor, con implicaciones de costo y otras, pero lo que siempre se puede es mejorar la calidad de las mediciones, del diseño de investigación y otras circunstancias que no tiene por qué ser siempre iguales.

Por otra parte, se ha hablado de probabilidad alta o baja, sin precisar estos términos, lo que no es sencillo. En el caso de un riesgo muy menor se puede aceptar una probabilidad no tan pequeña de que ocurra, pero tratándose de un riesgo grave solo se aceptará correrlo con una probabilidad mínima.

El uso del valor $p < .05$, el más utilizado en la investigación social como límite para rechazar una hipótesis nula, se extendió cuando no se contaba con computadoras que permiten estimarlo con total precisión. Es necesario subrayar que se trata de un convencionalismo que no tiene sustento sólido y, sobre todo, que depende en buena parte del tamaño de la muestra, por lo que diferencias mínimas entre el valor de una muestra muy grande y el de la población resultarán significativas con una probabilidad mucho menor a .05; con muestras pequeñas, en cambio, diferencias que podrían ser muy relevantes no resultarán significativas estadísticamente.

Ningún valor p es un umbral significativo por sí mismo. Con una muestra de decenas de miles como ahora abundan, uno de .01, o incluso de .001, puede referirse a una diferencia irrelevante, mientras que con una muestra de un centenar un valor p de .1, o aún mayor, podría ser indicio de una diferencia sustantivamente relevante.

De lo anterior se desprende, por una parte, que ninguna técnica estadística puede sustituir el buen juicio de un investigador que conoce el fenómeno que estudia y puede cuidar de varias formas la calidad de la información que obtiene y la solidez de las conclusiones a las que llega; por otra parte, que hay otras formas

de estudiar la importancia de las diferencias entre valores de varias muestras y de la población.

Los límites de las pruebas de hipótesis y el riesgo de sacralizar el valor $p < .05$ han sido señalados hace más de medio siglo (Cohen, 1994). En el año 2000 la *American Psychological Association* y la *American Educational Research Association*, las agrupaciones profesionales más importantes de sus campos, adoptaron el criterio de exigir para publicar un texto que no tuviera solo resultados de pruebas de hipótesis, sino también de otros análisis, como los del poder de una prueba para descartar errores Tipo II (*Power Analysis*, *cfr*. Norton y Strube, 2001) información sobre Tamaño del Efecto (*Effect Size*) e intervalos de confianza (Vogt, 2007: 97).

La cuestión se sigue discutiendo, lo que llevó a la *American Statistical Association* a formar un grupo de trabajo que en 2016 difundió un posicionamiento oficial en el mismo sentido (Wasserstein y Lazar, 2016). Al respecto puede verse también Nuzzo, 2014; y Gigerenzer, Krauss y Vitouch, 2004).

Tamaños del efecto, márgenes de error e intervalos de confianza

Además de saber qué tan probable es que una diferencia entre dos valores se deba al azar o no, importa saber qué tan importante es la diferencia. Esto es muy claro en estudios con los que se quiera saber qué tan importante es el efecto de una variable experimental, como un tratamiento médico o una intervención pedagógica.

El valor p depende en parte del tamaño del efecto de la variable independiente, y en parte del tamaño de la muestra usada; por eso es necesario contar con medidas que estimen solo el tamaño del efecto, de manera independiente del tamaño de la muestra (Coe, 2002). Así es posible identificar efectos que, en términos de valor p parecen pequeños, pero en términos sustantivos pueden ser muy importantes (Prentice y Miller, 1992).

Hay diversas medidas del tamaño del efecto, que son apropiadas según el nivel de medición de la variable independiente y de la dependiente.

En un experimento se pueden comparar los resultados obtenidos en el grupo en el que se aplicó el tratamiento o se llevó a cabo la intervención (grupo experimental) con los del grupo de control. Este es un caso en el que la variable independiente es categórica (dicotómica), y la dependiente es métrica, suponiendo que se cuente con una medición de ese nivel. Si las mediciones en los grupos se hicieron con diferente escala, para poder comparar los resultados hay que usar unidades estandarizadas, como la σ, como se hace con los *valores z*. La medida usual para estimar el Tamaño del Efecto en un

caso como el anterior es la diferencia entre las medias de los dos grupos, estandarizada por la desviación estándar del conjunto: la d de Cohen.

La comparabilidad que permite el uso de valores estandarizados es particularmente importante en estudios cuyo propósito es, precisamente, comparar resultados de muchas investigaciones que estudiaron las mismas variables, utilizando medidas distintas: los metaanálisis que se presentan en el apartado 3 de este capítulo.

Si tanto la variable independiente como la dependiente son métricas, la medida del Tamaño del Efecto será simplemente el coeficiente de correlación r de Pearson. Y si las dos variables son dicotómicas, la medida apropiada será una *razón de momios* (*odds ratio*), que se presentará en el punto 3 de este capítulo.

En cuanto a márgenes de error e intervalos de confianza, consideremos el caso de una prueba aplicada a una muestra aleatoria de 100 estudiantes, con el resultado de una media de 445 puntos y una desviación estándar de 50. Una inferencia sobre el probable puntaje promedio de la población de la que se sacó la muestra implica calcular el margen de error de la medición hecha en la muestra (*error estándar*) y, a partir de ello, el rango en el que, con cierta probabilidad, podrá situarse realmente el puntaje promedio de la población, el *intervalo de confianza*.

Con base en el estudio de las distribuciones muestrales, se determina que el *error estándar de la medición* se obtiene dividiendo la desviación estándar del puntaje obtenido en la muestra, entre la raíz cuadrada del tamaño de la muestra. En la muestra del ejemplo: $50/\sqrt{100} = 50/10 = 5$. Con base en las propiedades de la curva normal, el *intervalo de confianza*, con 95% de probabilidad, será igual a dos desviaciones estándar por arriba y por debajo de la media obtenida en la muestra, lo que quiere decir que el valor de la población se ubicará, con 95% e probabilidad, entre 435 y 455 puntos (445 ± 10).

Conviene reiterar que los paquetes de software calculan en segundos los valores de muchas pruebas estadísticas, pero no pueden sustituir al buen juicio del investigador para interpretar los valores p que resulten, con tamaños del efecto, márgenes de error e intervalos de confianza.

Es claro, además, que muchos estudios no buscan llegar a inferencias estadísticas sobre una población a partir de hallazgos muestrales, sea porque cubren a toda la población (es de tipo censal), sea porque se trata de *estudios de caso,* que cubren un número reducido de sujetos, e incluso solo uno, y solo pretenden decir algo de ese o esos casos, sin pretensiones de generalizar.

Muestreo

La posibilidad de llegar a conclusiones sobre una población grande a partir de la información obtenida de un subconjunto mucho menor hace atractivo el uso de diseños que usen muestras. Sin embargo, el muestreo es un campo complejo, y la formación al respecto de muchos investigadores es limitada; por ello en este inciso se presentan algunas ideas básicas, comenzando con un poco de historia.

Uno de los cinco primeros libros de la Biblia (el *Pentateuco*), el llamado los *Números*, comienza diciendo:

> [...] Yahvé habló a Moisés en el desierto del Sinaí, en la Tienda de la Reunión, el primer día del segundo mes del segundo año después de la salida de Egipto, y le dijo: harás un censo de toda la comunidad de los hijos de Israel [...]

El libro sigue con prolijos detalles de la forma en que debía organizarse el censo y de sus resultados, según los cuales *los varones de 20 años y más, aptos para hacer campaña*, de las doce tribus de Israel, sumaban 603,550.

Rivalizando en antigüedad con el censo bíblico, en China se hacía este tipo de ejercicios hace decenas de siglos. Hace mil años, el *Domesday Book* recogió los datos del censo de Inglaterra ordenado por Guillermo el Conquistador, y en 1790 se levantó el primer censo en los recién independizados Estados Unidos. En México el primer censo general de población tuvo lugar en 1895.

Desde mediados del siglo XIX, por otra parte, personas preocupadas por la situación de las familias obreras comenzaron a realizar estudios sobre el tema; los trabajos de Henry Mayhew y Charles Booth en Inglaterra, y los de Le Play en Francia, son ejemplos destacados de trabajos pioneros que fueron seguidos, en las primeras décadas del siglo XX, por otros como los de Benjamín Rowntree y Arthur Bowley en el Reino Unido, o los de Paul Lazarsfeld, Marie Jahoda y Hans Zeisel sobre los desempleados de Marienthal. (*Cfr*. Bradburn y Sudman, 1988: 15-16)

Según Rea y Parker, la idea de hacer inferencias sobre ciertas características una población a partir de la observación de un subconjunto de la misma (*una muestra*), la tuvo por primera vez un inglés llamado W. S. Gosset, que firmaba sus trabajos de estadística con un seudónimo que se ha vuelto familiar para muchas personas que comienzan a estudiar esa disciplina: *Student*. Es curioso el contexto en el que una idea tan importante para la ciencia fue concebida: su autor trabajaba en la cervecería Guinness, y buscaba una forma de probar el producto de la planta, para verificar su calidad, sin consumirlo o,

al menos, sin consumir una proporción considerable. Para resolver la cuestión, Gosset-Student desarrolló las bases teóricas en que se sustenta la posibilidad de hacer inferencias sólidas —esto es, con un margen de error reducido— relativas a una población grande, con base en la observación de una muestra pequeña de ella. (Rea y Parker, 1992: 6)

En ámbitos relacionados con las ciencias sociales, en particular la politología y los estudios de mercado, Bradburn y Sudman (1988) señalan que, desde la primera mitad del siglo XIX, algunos periódicos y revistas comenzaron a hacer las encuestas conocidas como *straw polls*, con el propósito de predecir el resultado de elecciones próximas a desarrollarse.

Se acepta que la primera de estas encuestas fue hecha por el periódico *Harrisburg Pennsylvanian*, con un resultado de 355 votos a favor de Andrew Jackson, 169 por John Quincy Adamas y números menores por otros dos candidatos. Los periódicos imprimían una boleta de votación y pedían a los lectores que la recortaran, llenaran y enviaran por correo. Con el siglo XX un número creciente de medios de adoptó esa práctica. La revista *Literary Digest,* comenzó a hacerlo en 1916.

Los estudios de mercado comenzaron en Estados Unidos en 1879, en la agencia de publicidad Ayer & Son. En 1911 J. George Frederick fundó el primer despacho de estudios de mercado (*Business Bourse*), que fue el primero en contratar en forma permanente entrevistadores residentes en diferentes localidades. En la misma fecha se creó en la Universidad de Harvard el *Bureau of Business Research*, que lanzó el primer estudio de lectores de revistas y comenzó el departamento de investigación comercial de la empresa Curtir Publishing Co., que llegó a tener alrededor de 1,200 entrevistadores. (Bradburn y Sudman, 1988: 12-14)

Más tarde aparecieron empresas dedicadas expresamente a hacer encuestas para algunos clientes. En julio de 1935 Elmo Roper, que trabajaba para un despacho de estudios de mercado, comenzó a hacer encuestas para la revista *Fortune*. En octubre de ese año George Gallup fundó el *American Institute of Public Opinion*, para hacer encuestas semanales sobre temas de interés comercial o político para un grupo de 35 periódicos, así como para otros clientes.

La superioridad de los resultados de una encuesta basada en una muestra pequeña pero bien hecha quedó espectacularmente establecida en 1936, con ocasión de la elección que enfrentó a Alf Landon por el Partido Republicano y Franklin Roosevelt por el Demócrata. El *Literary Digest*, cuyas previsiones habían acertado en todas las elecciones presidenciales desde 1920 predijo amplia victoria de Landon. Con su método tradicional, la revista envió por correo diez millones de boletas para simular la votación

a los hogares enlistados en los directorios telefónicos y los registros de automóviles de todo el país, recibiendo alrededor de 2.4 millones de respuestas.

En cambio, tres encuestas hechas por las empresas de Gallup, Roper y Crossley predijeron un holgado triunfo de Roosevelt. Bradburn y Sundman citan una nota que permite entender por qué la superioridad de esas encuestas respecto al trabajo del *Literary Digest* se mostró de manera tan contundente. En la edición del *Saturday Evening Post* del 21 de enero de 1939, W. Rich escribió:

> Landon fue nombrado candidato el 11 de junio de 1936. El día 12 Gallup advirtió a sus suscriptores que los anticuados métodos del Literary Digest se equivocarían en sus predicciones sobre el resultado de la elección [...] y mencionó que esos métodos predecirían alrededor de 56% para Landon y 44% para Roosevelt. Como la encuesta tendría lugar seis semanas después, el editor del Literary Digest, Wilfred J. Funk se indignó, y dijo: "Nunca antes se había atrevido alguien a predecir lo que nuestra encuesta mostraría antes incluso de que comenzara. Nuestro buen amigo estadístico debería saber que el Literary Digest hará su encuesta con esos anticuados métodos que, en el pasado, han arrojado predicciones correctas el 100% de las veces". (Citado en Bradburn y Sundman, 1988: 21-22)

Una vez realizada, la encuesta del *Literary Digest* predijo una victoria de Landon, con 57% de la votación, con base en 2.4 millones de respuestas. Las tres encuestas predijeron el triunfo de Roosevelt. La encuesta de Gallup, con una muestra de 5,000 personas seleccionadas aleatoriamente según cuotas de edad, género y zona de residencia, predijo que Roosevelt ganaría con ±60% de la votación. Roosevelt obtuvo 61% de los votos. El *Literary Digest*, que ya tenía problemas financieros, tuvo que cerrar, y la superioridad de las encuestas basadas en muestras calculadas con base en las novedosas teorías estadísticas, se impuso *casi de la noche a la mañana,* según la expresión de Bradburn y Sudman. (1988: 22)

El desarrollo de la teoría del muestreo ha refinado los planteamientos iniciales de la época de Gallup, con lo que hoy pueden emplearse muestras aún menores, y también es posible saber en qué condiciones hay mayor o menor riesgo de error.

Nociones básicas

Las pruebas PISA se aplican cada tres años a muestras representativas de una población de jóvenes de 15 años de edad, inscritos en cualquier grado escolar a partir del primero de secundaria. México tiene más de 130 millones de habitantes, más de dos millones de

jóvenes de 15 años, y ± 1.3 millones inscritos en secundaria o educación media superior. La muestra mínima que se pide para que los resultados que obtengan los estudiantes a los que se aplique la prueba sean representativos de la situación del grupo objetivo de PISA es de unos 5,500 estudiantes.

Es posible preguntarse de qué tamaño será la muestra mínima que se pida a países como Estados Unidos que tiene más de 300 millones de habitantes; Brasil, más de 200; Alemania, con unos 80; Finlandia, con 5.7; Uruguay, con 3.5; o Luxemburgo, cuyo total de habitantes es poco superior a medio millón. A muchos sorprenderá saber que el tamaño de la muestra requerida para participar en PISA es el mismo en todos estos casos. Para entender la razón, sin necesidad de ser especialista en muestreo, es necesario tener claras algunas nociones básicas:

- Universo o población total: todos los miembros del conjunto de que se trate.
- Población efectiva: la considerada para el estudio, descartando por razones expresas una parte de la población total o universo.
- Población marco: la que incluye el mejor listado disponible de los miembros de la población efectiva; el listado constituye el marco muestral.
- Población conseguida: los integrantes del marco muestral a los que se pudo tener acceso efectivamente.
- Censo: estudio con todo el universo.
- Muestra: subconjunto de la población que se considera representativo.

Para tener datos sobre ciertos aspectos de un grupo, se puede buscar a cada uno de sus miembros. Eso es fácil con grupos pequeños, no grupos grandes. Por ello los estudios exhaustivos (*censos*) se hacen pocas veces y sobre pocos aspectos, y es frecuente estudiar subconjuntos de una población grande (*muestras*), cuidando que sean representativos del total. Esto lleva a otra noción fundamental:

- Representatividad: propiedad de una muestra que consiste en que, por su tamaño y por la forma en que se extrajo, los resultados de ella pueden ser generalizados al conjunto de la población de que se sacó, con la precisión y la significatividad estadística que se consideren adecuadas.

Se deben cuidar dos elementos para que una muestra sea representativa: por una parte, aspectos cuantitativos, incluyendo el tamaño y otros; por otra, lo cualitativo, en

particular la forma en que se obtenga. Suele prestarse atención al tamaño de la muestra, con la idea simplista de que mientras más grande mejor; suelen descuidarse otros aspectos que pueden influir.

La muestra representativa

Intuitivamente es razonable pensar que el tamaño de la muestra es importante para la solidez de las inferencias respecto de la población. Es claro que una conclusión sobre un universo grande basada en la observación de un solo caso no puede ser muy sólida, pero hay que preciar muchos puntos.

¿Será lo mismo una muestra de 10 casos, o de 100 casos, para hacer generalizaciones respecto de una población de 100 o de 1000 sujetos? ¿Importará el tamaño absoluto o el relativo? ¿Serán de igual calidad dos muestras que sean la misma proporción de su población, una muestra de 10 de una población de 100, una de 100 de una población de 1,000, o una de 1,000 de una población de 10,000?

▪ **Aspectos cuantitativos que influyen en la representatividad**

El tamaño de una muestra representativa depende de cuatro factores:

- El tamaño de la población.
- La homogeneidad de la población.
- La precisión con que se quieran estimar los valores de la población.
- La probabilidad de error que se acepte (significatividad estadística).

Suele pensarse que el tamaño de una muestra depende principalmente del de la población. Esta idea es solo parcialmente correcta. En realidad, los cuatro factores mencionados importan todos, pero de distinta forma.

Tamaño de la población. Importa, desde luego, pero tratándose de poblaciones pequeñas; si el tamaño de la población aumenta su importancia es cada vez menor.

Homogeneidad. Una población menos homogénea hace necesarias muestras más grandes que una más homogénea. Para entender por qué, recordemos que *para muestra basta un botón*. ¿Cuántos botones debo sacar de un cajón en el que hay mil, para saber de qué color son, si estoy seguro de que todos son del mismo color? Solo uno. Pero si todos los botones son de distinto color deberé sacarlos todos. El viejo dicho citado expresa justamente la idea de que, para poder inferir algo sobre una población a partir de la información obtenida de una muestra de ella, el tamaño de la muestra necesaria es

menor con una población homogénea, pero debe aumentar si se trata de una población más heterogénea.

Precisión. Si quiero saber cuál es, en promedio, la estatura de los niños de primaria de México, a partir de una muestra, el tamaño de esta deberá ser mayor si quiero saber dicha estatura aceptando una precisión de ± 1 cm., que si me conformo con conocerla con una precisión de ± 5 cms. Mientras más precisas deban ser las estimaciones de los valores de cierta variable en la población, a partir de las obtenidas de una muestra, mayor deberá ser esta.

Significatividad estadística. La muestra necesaria para estimar el valor de cierta variable en una población deberá ser mayor si acepto que la probabilidad de que el resultado de la muestra no se deba al azar sea, a lo más, de 5 % ($p < .05$); pero si quiero que esa probabilidad sea menor (*v.gr.* $p < .01$) la muestra deberá ser mayor.

El tamaño de la población y su homogeneidad o heterogeneidad no dependen del investigador; el primero suele conocerse con razonable exactitud; la segunda puede conocerse por estudios previos, o estimarse. Los otros dos factores, la precisión y la significatividad estadística, si son definidos por los investigadores.

Cada factor influye en el tamaño que deba tener la muestra para dar resultados representativos de los valores de la población, y es necesario tenerlos todos en cuenta, aplicando los principios de la inferencia estadística. Los cálculos necesarios pueden ser bastante complejos cuando se trata de muestras estratificadas, con pesos diferentes de subgrupos de la población, entre otros aspectos.

La Tabla 4.8 presenta sintéticamente el tamaño que debe tener una muestra para dar resultados representativos de la población, teniendo en cuenta los cuatro factores que se acaba de presentar.

- Los renglones corresponden a poblaciones de distinto tamaño, con seis casos de 100, 1,000, 10,000, 100,000, 1,000,000 y 10,000,000 de sujetos.
- Las ocho columnas con datos numéricos de la tabla presentan las combinaciones de dos casos imaginarios de cada uno de los otros tres factores:
- Las cuatro columnas de la izquierda tratan de poblaciones más homogéneas (de menor desviación estándar, $\sigma = 5$), y las cuatro de la derecha de poblaciones más heterogéneas (de mayor desviación estándar, $\sigma = 10$).
- En cada bloque de cuatro columnas, las dos de la izquierda se refieren a casos en que se pide más precisión (se acepta un error de ± 1%); en las dos de la derecha se acepta menos precisión, aceptándose un error de ± 5%.

- En cada uno de los ocho pares de columnas, en la de la izquierda se acepta un nivel de significatividad estadística menor ($p < .05$); en la de la derecha se pide un nivel mayor de significatividad estadística ($p < .01$).

La Tabla 4.8 permite apreciar la importancia de cada uno de los cuatro factores, en lo que se refiere a la definición del tamaño que debe tener una muestra para que sus resultados reflejen razonablemente los de la población de la que fue extraída, en cada una de las situaciones imaginarias que definen los factores mencionados.

Todas las casillas de las cuatro columnas de la parte derecha de la tabla tienen cifras mayores que sus similares de la parte izquierda, lo que confirma que una población menos homogénea implica mayor tamaño de muestra. Las muestras de casos en que se pide precisión de ± 5% son mucho menores que las de los casos similares en que se pide precisión de ± 1%, lo que refleja que si no se busca mucha precisión basta una muestra bastante chica. En forma similar, las muestras de los casos en que se acepta una significatividad estadística menor ($p < .05$) son menores a las de los casos en que se requiere una mayor ($p < .01$).

Y lo que más debe llamar la atención: en todos los casos, con todos los niveles de homogeneidad de la población, de precisión y de significatividad estadística,

- Con una población pequeña (100 sujetos) la muestra comprende una proporción de la población que va desde sumamente alta en el caso más exigente (99/100 con una población poco homogénea y exigencias altas de precisión y significatividad), hasta poco menos de la mitad del total en el caso menos exigente (42/100 con una población muy homogénea y exigencias bajas de precisión y significatividad).
- En todos los supuestos el tamaño de la muestra necesaria aumenta con el de la población, pero cada vez menos. La muestra necesaria para poblaciones de 1,000 o 10,000 sujetos son bastante mayores que la necesaria con una población de solo 100, pero el incremento no es proporcional. En el caso más exigente la muestra está lejos de representar 99% de la población: es solo algo mayor a la mitad (5,564); y en el caso menos exigente, sorprendentemente, la muestra necesaria para una población de 10,000 sujetos es solo de 72. A partir de 100,000 sujetos, aun con poblaciones más y más grandes, hasta de 10 millones, el tamaño de la muestra necesaria cambia muy poco, con todas las combinaciones de los otros tres factores.

TABLA 4.8. MUESTRA NECESARIA PARA DAR RESULTADOS REPRESENTATIVOS DE POBLACIÓN DE DISTINTO TAMAÑO Y HOMOGENEIDAD, CON DIFERENTE PRECISIÓN Y PROBABILIDAD DE QUE EL RESULTADO NO SE DEBA AL AZAR (SIGNIFICATIVIDAD ESTADÍSTICA)

Tamaño de la población	Mayor homogeneidad σ = 5				Menor homogeneidad σ = 10			
	Precisión		Precisión		Precisión		Precisión	
	± 1%		± 5%		± 1%		± 5%	
	Significatividad		Significatividad		Significatividad		Significatividad	
	p < .05	p < .01	p < .05	p < .01	p < .05	p < .01	p < .05	p < .01
100	95	97	42	56	99	99	74	83
1,000	645	758	68	111	879	926	225	334
10,000	1,537	2,387	72	124	4,207	5,564	282	478
100,000	1,783	3,040	73	125	6,770	11,145	290	499
1'000,000	1,812	3,126	73	125	7,210	12,387	290	501
10'000,000	1,815	3,135	73	125	7,257	12,527	290	502

FUENTE: ELABORACIÓN PROPIA.

No es esperable que todo investigador domine inferencia estadística y muestreo en el grado necesario para calcular el tamaño de muestras, en especial complejas, y para hacer inferencias con base en los resultados muestrales; de hecho, en estudios de grandes dimensiones es usual involucrar a especialistas en muestreo, distintos de los investigadores responsables, para que se encarguen de estos aspectos.

El lector de este apartado que se tienen en mente es un investigador en formación que, al diseñar su proyecto de tesis, en ocasiones trata de definir el tamaño de la muestra que necesitará sin saber que, tratándose de poblaciones chicas, como es muchas veces el caso, ese tamaño deberá ser cercano al de la población. Es más sencillo tratar de cubrir a toda la población, a sabiendas de que siempre quedarán fuera algunos casos, y no intentar trabajar con una muestra bien calculada, con las dificultades que supondrá luego que sea aleatoria, con casos no localizados que se deben sustituir por otros definidos también aleatoriamente. Esto lleva a considerar de los aspectos cualitativos que inciden en la representatividad de una muestra.

▪ La forma de obtener la muestra

Para que la muestra dé resultados generalizables a una población de cierto tamaño y homogeneidad, y lo haga con cierto margen de precisión y cierta significatividad, es

esencial que sea estrictamente aleatoria, sin confundir aleatoriedad con cualquier forma no sistemática de seleccionar los sujetos de la muestra.

Supongamos que un estudio requiere una muestra de 1,000 sujetos de un total de 15,000 estudiantes de una universidad. Se pueden pensar dos formas de muestrear

- Se escoge a los alumnos que el investigador encuentra casualmente en la biblioteca de la institución, o bien en la cafetería, en la parada del transporte urbano más cercana, o en el estacionamiento del plantel.
- Se escogen sorteando sus nombres en una lista de todos los inscritos.

Ir a cierta área de la institución y entrevistar a las personas que encuentre no es un procedimiento aleatorio, ya que el hecho de que una persona esté en uno u otro lugar no se debe al azar, sino a factores precisos, como el compromiso académico de los sujetos (biblioteca vs cafetería), o su nivel socioeconómico (parada de autobús vs estacionamiento. Una muestra aleatoria implica contar con un listado lo más completo que se pueda de los sujetos que componen la población (un buen marco muestral), y un procedimiento realmente azaroso de seleccionar los sujetos.

▪ Tipos de muestra

Las muestras simples tienen un solo nivel, pero a veces se necesitan muestras más complejas, para poblaciones en que hay subgrupos de características peculiares (*estratos*), que pueden ser de tamaño diferente, por lo que su *peso* es distinto. Hay muestras estratificadas, ponderadas, o polietápicas. Por otra parte, no siempre hay listados razonablemente completos de los integrantes de la población (marcos muestrales), y se pueden hacer muestras intencionales, en cascada, entre otras.

Una muestra es aleatoria si todos los sujetos de la población tienen probabilidad conocida y definida, aunque no necesariamente igual, de ser escogidos. Por ello, para que una muestra sea realmente aleatoria se necesita:

- Un buen *marco muestral*: un listado que incluya de manera muy completa, si no exhaustiva, a todos los miembros de la población de que se trate.
- Una forma de escoger a los sujetos que dé a todos la misma probabilidad (o la que se haya definido) de resultar seleccionado.

Definida la muestra hay que recoger la información, evitando el error de omitir los sujetos que sea difícil localizar, privilegiando a los más accesibles, lo que hace que la muestra pierda representatividad, ya que los sujetos difíciles de localizar tendrán en general características diferentes de los más accesibles. Por ello al definir la muestra se suele incluir un número de *remplazos*, que los encuestadores deberán buscar cuando no encuentren a alguno(s) del listado inicial.

Generalización a partir de datos muestrales

Suele creerse que un estudio censal siempre es mejor que uno muestral (*censo mata muestra*), o que una muestra grande siempre es mejor que una chica. No es así. Las amenazas a la calidad de la información aumentan a medida que crecen las dimensiones de la aplicación. No se puede cuidar igual la calidad de la aplicación de un instrumento a una muestra de 5,000 sujetos, que la de un censo de millones. Si no se necesitan resultados individuales, un buen estudio muestral tiene ventajas respecto a un censo, no sólo por costo, sino también por la calidad de la información que se obtiene. Así lo señala un sistema de evaluaciones del rendimiento escolar reconocido mundialmente por su solidez:

> El NAEP aplica pruebas a una proporción relativamente pequeña de la población de alumnos de interés, con métodos probabilísticos de muestreo. Reducir el número de alumnos a los que se aplican pruebas permite dedicar más recursos a garantizar la calidad de éstas y su aplicación, produciendo estimaciones considerablemente mejores que las que se obtendrían si se aplicaran pruebas a todos los alumnos en condiciones menos controladas. El uso de muestras reduce la carga que se impone a los alumnos, a los estados y a las localidades, en comparación con un programa de pruebas que las aplique a proporciones importantes de los alumnos del país. (NAEP, 1999)

En un estudio muestral puede haber dos tipos de errores: *de muestreo (sampling errors),* y *no derivados del muestreo (non sampling errors)*. Un censo evita el error de muestreo, pero implica aumentar los no derivados del muestreo. La inferencia estadística permite minimizar el error de muestreo, pero minimizar los errores no derivados del muestreo implica otros cuidados, distinguiendo dos subtipos:

- Errores de medición, debidos a limitaciones de los instrumentos y/o de los procedimientos de obtención de datos. Si las preguntas de un cuestionario inducen

cierta respuesta, los datos obtenidos tendrán un error de medición que no tiene que ver con el muestreo; lo mismo ocurrirá si el instrumento es bueno, pero unos aplicadores dan más tiempo para responder a los sujetos y otros les dan menos tiempo, o si se orienta a los sujetos en cierto sentido.

- Errores por otros factores, sistemáticos o aleatorios, como los que pueden presentarse al capturar los resultados del trabajo de campo. Si quien captura datos derivados de un trabajo de campo no lo hace con el debido cuidado, si la captura se hace con un lector óptico y el aparato o el software que utiliza tienen una falla, o si ocurre algo similar en la fase de procesamiento de los datos, estamos ante otros ejemplos de este subtipo de error.

Análisis de la calidad de la información

La importancia de trabajar con información de la mejor calidad posible, junto con la variedad de las amenazas que atentan contra dicha calidad, hacen que se hayan desarrollado técnicas específicamente orientadas a analizar diversos aspectos de la noción misma de calidad de la información.

En el Cap. 3 se presentan conceptualmente las cualidades fundamentales de una buena medición: confiabilidad y validez. Aquí se presentan técnicas básicas para el análisis de esas cualidades, y para el de un supuesto de ambas, la dimensionalidad de un instrumento de medición. Se hace alusión a técnicas complejas para estudiar estos aspectos, que se describen brevemente en el apartado 3.1.

Dimensionalidad

Para elaborar un instrumento de obtención de información o medición —cuestionario, escala, guía de entrevista o de observación visual— se debe comenzar definiendo con precisión el aspecto de la realidad que se quiere medir.

Como se dijo en el Cap. 1, construir el objeto de estudio implica acotar, precisar un tema que inicialmente suele ser muy amplio. Ese acotamiento o delimitación debe ser empírico y teórico, y este último quiere decir precisar el aspecto o aspectos de la realidad que se quiere estudiar, ya que la realidad es inmensa, inabarcable; el entendimiento humano solo puede captar aspectos particulares. Un instrumento de obtención de información o medición no puede captar la totalidad, sino solo uno o unos aspectos de la misma, ciertas *dimensiones* o *variables*.

En consecuencia, para obtener información de buena calidad con un instrumento hay que asegurarse de que está diseñado para que atienda realmente el aspecto o los

aspectos que se quiere medir. Esto es lo que se busca con el proceso de operacionalización de los aspectos de que se trate (dimensiones, variables, indicadores o marcadores de conducta, etc.), para que cada pregunta o ítem del instrumento se refiera a una variable precisa.

En el Cap. 3 se mostró que la diferencia que distingue una escala de un cuestionario convencional es que cada pregunta de este último pretende medir una variable distinta, mientras que con el conjunto de los ítems de una escala se quiere medir una sola variable (actitud, opinión) y se espera conseguirlo de manera más confiable con varias preguntas o reactivos, y no solo con uno. Sin embargo, la manera en que los sujetos a quienes se aplica un instrumento entienden las preguntas no siempre coincide con lo que esperaban sus autores; la consecuencia es que las dimensiones de la información que efectivamente se obtiene con el instrumento pueden ser diferentes de las que se plantearon al diseñarlo, y que por ende no basta el cuidado *a priori* de la unidimensionalidad que se busca con la operacionalización, sino que es necesario cuidarla también *a posteriori*, con un análisis especial para ello.

En un estudio sobre las actitudes de maestros de educación media superior (*high school*) respecto a la forma correcta de vestir de los estudiantes, los investigadores encontraron que las siete preguntas de la escala usada formaban dos grupos, uno sobre el uso de uniformes y el otro sobre lo apropiado de la demás ropa. Pese a ello la confiabilidad de la escala era aceptable (Vogt, 2007: 116).

Las técnicas para analizar las dimensiones medidas con un instrumento, verificando si corresponden a la dimensión o constructo latente que se quiere estudiar (si son realmente sus indicadores, componentes o factores) son el análisis de componentes principales y el análisis factorial. Por ser técnicas relativamente complejas se deja su presentación para el tercer apartado de este capítulo.

Una vez precisada la variable(s) que se mide(n) se pueden hacer análisis de la confiabilidad y la validez de la información recabada, con las herramientas analíticas a que se refieren los puntos siguientes.

Confiabilidad

Se entiende por confiabilidad la *consistencia* de una medición, que hace obtener resultados iguales o muy parecidos si se hace una y otra vez (Brennan, 2001).

La consistencia de la información obtenida se puede analizar de varias formas según el instrumento de que se trate: cuestionarios o escalas que tienen cierto número de ítems; protocolos de observación para valorar prácticas videograbadas, o guías de

análisis de evidencias, en los dos últimos casos con la intervención de cierto número de observadores o calificadores.

Los procedimientos estadísticos para calcular la confiabilidad de la información obtenida, los *coeficientes de confiabilidad* son coeficientes de correlación, cuyo valor va de 0 (completa inconsistencia) a 1 (consistencia directa perfecta) o a — 1 (consistencia inversa perfecta). Según el tipo de instrumento de que se trate, y las formas de utilizarlo, se pueden distinguir varios tipos de medidas de confiabilidad:

- Entre dos aplicaciones (*test-retest reliability*).
- Entre formas paralelas (*parallel forms reliability*).
- Consistencia interna (*internal consistency reliability*), incluyendo medidas de confiabilidad de mitades (*split-half reliability*).
- Entre jueces (*inter-rater reliability*). (Vogt, 2007:114-115; Salkind, 2007: 307)

Las medidas de confiabilidad entre dos aplicaciones y entre formas paralelas son simplemente coeficientes de correlación de Pearson.

Las medidas de consistencia interna son diferentes, ya que tratan de ver si los ítems o preguntas de un instrumento dan resultados consistentes *entre sí*. Para este propósito, los ítems se pueden dividir en partes (por ejemplo, mitades) con lo que llegamos a las medidas *split-half*.

La más conocida de estas, y la más usada de todas las medidas de confiabilidad, es el coeficiente alfa (α) de Cronbach, que conceptualmente se define como la media de las medidas de confiabilidad obtenidas entre todas las mitades que sea posible formar con los ítems de un instrumento. (Vogt, 2007: 115)

Aunque el concepto parece complicado, la fórmula para calcular el alfa de Cronbach es sencilla, y cualquier paquete de cómputo la incluye (Salkind, 2007: 310-311):

$$\alpha = \frac{k}{k-1} \times \frac{s^2y - \sum s^2i}{s^2y}$$

Las letras de la fórmula significan lo siguiente

- k el número de ítems del instrumento de que se trate.
- s^2y la varianza asociada al puntaje total que arrojan en conjunto los ítems.
- $\sum s^2i$ la suma de las varianzas asociadas a cada uno de los ítems.

La primera parte de la fórmula implica simplemente dividir el número de ítems del instrumento entre la misma cifra menos uno. El numerador de la segunda parte es la diferencia que resulta de restar la suma de las varianzas asociadas a cada ítem de la varianza asociada al puntaje total; y el denominador es esta misma varianza asociada al puntaje total.

La tabla siguiente muestra un ejemplo con los datos necesarios para calcular un coeficiente α, en el caso de un instrumento de cinco ítems que se habría aplicado a 10 sujetos. En la tabla se presentan los valores obtenidos por cada sujeto en cada uno de los ítems y la varianza asociada a cada uno. En la última columna se da el puntaje total de cada sujeto, la varianza asociada a ese puntaje, y en la última casilla la suma de las varianzas asociadas a los cinco ítems.

TABLA 4.9. CÁLCULO DE UN COEFICIENTE α.

Sujetos	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Puntaje total
1	3	5	1	4	1	14
2	4	4	3	5	3	19
3	3	4	4	4	4	19
4	3	3	5	2	1	14
5	3	4	5	4	3	19
6	4	5	5	3	2	19
7	2	5	5	3	4	19
8	3	4	4	2	4	17
9	3	5	4	4	3	19
10	3	3	2	3	2	13
						$s^2y = 6.4$
	s^2=0.32	s^2=0.62	s^2=1.96	s^2=0.93	s^2=1.34	$\sum s^2i = 5.17$

FUENTE: SALKIND, 2007: 311.

Aplicando la fórmula a los datos de la tabla anterior tenemos

$$\alpha = \frac{5}{5-1}\left(\frac{6.4 - 5.17}{6.4}\right) = 1.25 \times 0.1922 = 0.24$$

El coeficiente alfa en este caso es igual a 0.24. ¿Cómo interpretarlo? Suelen darse recomendaciones en el sentido de que, por ejemplo, un instrumento sería adecuado si

el valor del α que representa su confiabilidad fuera igual o mayor a 0.75. Como otras recomendaciones similares, esta debe tomarse con cuidado, ya que las implicaciones de dar por buena cierta medición pueden ser de importancia muy distinta, lo que llevaría a ser más o menos exigentes para considerar aceptable cierto valor del coeficiente alfa.

Cuando se construye un instrumento se puede medir la confiabilidad que resulta si se utilizan todos los ítems que se están probando, o sólo algunos. Es posible ver el efecto que tiene sobre la confiabilidad eliminar cada uno de sus ítems. Por lo general la confiabilidad es mayor utilizándolos todos, y disminuye al quitar cualquiera, pero en algunos casos la confiabilidad aumenta al quitar cierto ítem, lo que indica que se trata de uno cuya correlación con el resto es negativa. Suprimirlo aumenta la confiabilidad y hace más corto el instrumento. Es posible tener niveles aceptables de confiabilidad con un número reducido de ítems de buena calidad (4-5).

Es importante añadir que un valor alto del coeficiente α no se debe interpretar como evidencia de que la escala o instrumento mida una sola dimensión. El estudio sobre actitudes de maestros de media superior antes referido es un ejemplo de que es posible tener una confiabilidad aceptable con un instrumento que mide dos o más dimensiones, si estas se correlacionan positivamente. La unidimensionalidad es una característica diferente del instrumento, que se da por supuesta cuando se estudia su confiabilidad, no es un rasgo que se pueda sustentar con un coeficiente α.

En el caso de otros tipos de instrumento, como una guía de observación, o una escala de valoración aplicada por jueces, como en una competencia gimnástica, una medida usual de la confiabilidad de las puntuaciones de dos jueces hechas en forma independiente (*inter-rater reliability*) es simplemente la proporción del número de acuerdos obtenidos por los jueces entre el total de acuerdos posibles (Salkind, 2007: 312-313). Otras medidas toman en cuenta la probabilidad de que una parte de los acuerdos que se observen se deba simplemente al azar, como el coeficiente Kappa de Cohen, que se aplica en el caso de variables categóricas.

Un protocolo muy conocido para observar prácticas docentes es el denominado *Classroom Assessment Scoring System* (CLASS). Un estudio sobre una variante para secundaria, (*Classroom Assessment Scoring System for Secondary Classrooms,* CLASS-S) considera los mismos *dominios* de la práctica docente: organización de la clase, apoyo emocional y apoyo instruccional que se ofrece a los alumnos. Análisis empíricos sobre la confiabilidad de la información que proporciona el CLASS reflejan diferencias en las medidas de estos tres dominios: alta para organización; media para apoyo emocional; baja para apoyo instruccional:

- La confiabilidad más alta se encuentra en aspectos que se identifican contando acciones puntuales, o con la frecuencia de interacciones simples en el aula. En general se observaron mejores prácticas de organización y gestión de la clase, con pocos casos de mal comportamiento o clima negativo, y estudiantes ocupados en las tareas de clase.
- En apoyo emocional, los menores puntajes se refirieron a autonomía, liderazgo, expresión de ideas por los estudiantes, o que se buscara que los contenidos fueran relevantes para ellos.
- Los puntajes más bajos de todos fueron relativos a apoyo instruccional, con pocas evidencias de apoyo *al desarrollo de habilidades complejas como análisis resolución de problemas, razonamiento, creación mediante la aplicación de conocimientos* (Gitomer *et al.*, 2014).

El texto citado señala que esta *mediocridad* de la práctica docente ha sido identificada por autores como Goodlad (1984), y añade:

> La confiabilidad de los puntajes no es simplemente una molesta exigencia psicométrica; es un indicador de que hay una comprensión compartida sobre constructos importantes, y sobre lo que constituye un desempeño exitoso respecto a tales constructos. Si no hay tal comprensión es difícil imaginar cómo se podrá apoyar eficazmente tanto a los futuros docentes en etapa de formación, como a los que ya están en servicio [...] Una forma de mejorar la confiabilidad es limitar el alcance de los sistemas de observación a los aspectos de la enseñanza que se pueden valorar más fácilmente de manera confiable [...] al hacerlo se puede sesgar la definición de enseñanza en una forma que privilegiará rasgos que se pueden observar o contar fácilmente [...] lo que puede llevar a ignorar otros cruciales [...] la buena organización de la clase es condición necesaria pero no suficiente de una buena enseñanza [...] Una enseñanza efectiva se distingue por interacciones como las incluidas en los dominios de apoyo emocional e instruccional. Los hallazgos indican que la tendencia histórica a dar calificaciones altas e indiferenciadas al maestro puede no ser simplemente cuestión de voluntad institucional o administrativa, como sugieren reportes recientes [...] Si los observadores no se sienten seguros en cuanto a sus juicios sobre aspectos instruccionales, y centran la atención en puntos relativos a la organización de la clase, es probable que evalúen correctamente la enseñanza como más exitosa que lo que pasaría si también se tomaran en cuenta juicios precisos sobre las dimensiones propiamente instruccionales de la enseñanza. (Gitomer *et al.*, 2014)

Además del CLASS-S, el estudio incluyó un cuestionario de autoevaluación (CLASS-T) para que los docentes dieran su propia valoración de los mismos dominios y dimensiones de sus prácticas. Los resultados reflejan dificultad para valorar prácticas de apoyo instruccional, y facilidad para hacerlo en cuanto a las de organización de la clase, de manera similar a lo encontrado de las observaciones, lo que parece corroborar la dificultad de tener comprensiones compartidas sobre los aspectos de las prácticas docentes que son más relevantes. (Gitomer *et al.*, 2014)

Las medidas de confiabilidad mencionadas para este tipo de instrumentos, como la proporción del número de acuerdos obtenidos por los jueces entre el total de acuerdos posibles, o coeficientes como el Kappa de Cohen, tienen limitaciones que hacen preferible otras técnicas para estimar la confiabilidad.

Las técnicas mencionadas consideran que la falta de confiabilidad se debe a una sola fuente de error, supuesto que no corresponde a la realidad, como observó desde la década de 1960 Cronbach que propuso el principal desarrollo del análisis de confiabilidad, con la *Teoría de la Generalizabilidad* (TG), aplicando el análisis de varianza a los datos de la medición, por considerar que las distintas fuentes del error de medición deben estudiarse por sí misma cada una.

La TG, junto con la Teoría de Respuesta al Ítem (TRI) son desarrollos de la Teoría Clásica de las Pruebas (*Clasical Tests Theory*, CTT). Por su complejidad se deja también su presentación para el tercer apartado del capítulo.

Validez

En el Cap. 3 se presentaron conceptualmente varias acepciones de la noción de validez, desde la que la entiende como la cualidad que consiste en medir realmente lo que se pretende medir, hasta las formulaciones recientes que, con Messick, la definen como "un juicio evaluativo integral del grado en que la evidencia empírica y los argumentos teóricos apoyan lo adecuado y apropiado de inferencias y acciones basadas en puntajes de pruebas y otras formas de evaluación [...]". (1989: 13)

Si se quiere valorar alguna variante o faceta de la validez, como se hacía antes de la propuesta de unificación de Messick, y se sigue haciendo todavía, se entiende que en cada caso se requiera de algún tipo particular de análisis.

En forma similar, y a partir de la idea de que el concepto de validez es unitario, los *Estándares para pruebas en educación y psicología* de las principales organizaciones profesionales de estos campos (AERA-APA-NCME, 2014) consideran que para sustentarla es necesario recoger evidencias de varias fuentes: de contenido, de estructura interna, de

relaciones de los puntajes del instrumento con otras variables, de procesos de respuesta, y de consecuencias de la evaluación.

En cuanto a contenido, las evidencias consistirán en información que sustente la idea de que los ítems del instrumento de que se trate son una muestra adecuada del universo de conductas consideradas en el dominio a evaluar.

La estructura interna se sustentará con datos sobre la planeación del instrumento, la claridad y consistencia de las especificaciones, entre otros.

La relación de los puntajes del instrumento con otras variables coincide con la llamada validez de criterio. En la variante concurrente implica analizar la correlación entre los resultados obtenidos con el instrumento de medición de que se trate y los que se recojan paralelamente con otro instrumento que busque medir los mismos rasgos. En la variante predictiva se analizará la correlación entre los resultados obtenidos en un momento dado con el instrumento en cuestión (como el puntaje obtenido por los aspirantes en una prueba de admisión a la universidad), y los de una medición posterior de la variable criterio que se esperaría predecir, como las calificaciones obtenidas por los admitidos en el primer año de su carrera.

La fuente relativa a los procesos de respuesta atiende a la consideración de que la forma en que los sujetos a quienes se aplica un instrumento interpretan el contenido de los ítems, y la forma en que los responden, puede ser diferente a lo que suponían los diseñadores, por factores de naturaleza cultural, lingüística, socioeconómica u otros, que no tienen que ver con el conocimiento de los constructos de interés. En este caso suelen usarse entrevistas cognitivas —con técnicas de pensamiento en voz alta, por ejemplo— para aproximarse al proceso real de respuesta que emplean los sujetos al responder y ver si coincide o no con el esperado.

Y en cuanto a consecuencias, deberán aportarse evidencias de que desde el diseño del instrumento se buscó identifica posibles efectos negativos de su aplicación, tomando precauciones para evitar que ocurrieran, y que se vigiló también si se presentaban consecuencias negativas inesperadas, para corregir lo necesario en el diseño y las aplicaciones futuras.

En el Cap. 2 se presentó el *enfoque a la validez basado en argumentos* (*Argument-based approach to validity, ABAV*) que implica sostener las interpretaciones de los resultados obtenidos con un instrumento, y los usos que se piensa hacer de ellos, mediante un conjunto de argumentos, interpretativos y de validez.

Cuidar la calidad de la información que se obtiene con un instrumento cualquiera no necesariamente implica el uso de técnicas complejas; aún sin ellas es posible evitar errores importantes, estando siempre alerta a lo que pueda hacer que los sujetos a los

que se aplica entiendan una pregunta o un estímulo en forma distinta a la que preveían los responsables de su diseño.

Unos ejemplos bastan para mostrar las diferencias considerables que puede haber en cuanto al sentido de ciertas preguntas, y no únicamente en escalas tipo Likert y en relación con *constructos latentes*, sino también en relación con cuestiones de tipo factual, en cuestionarios simples, debido a diferencias en el contexto correspondiente. Si se tienen en cuenta esas diferencias del contexto es posible entender resultados aparentemente paradójicos.

Es razonable pensar, por ejemplo, que el que los padres de familia (o los maestros) dediquen tiempo a apoyar a sus hijos o alumnos para realizar las tareas escolares contribuirá a que obtengan mejores resultados, pero en algunos estudios se ha encontrado una relación negativa entre el rendimiento escolar de unos niños y el tiempo que sus padres o su maestro dedican a apoyarlos. Una posible explicación de lo anterior es que, en algunos casos, los padres y los maestros dedican más tiempo a apoyar a los alumnos que presentan problemas, y no a todos.

En forma similar, un estudio de la OCDE sobre los maestros de secundaria dio como resultado que, en México, los alumnos de los grupos que reciben más visitas del director de la escuela o del supervisor, son los que tienen peores resultados, y no mejores como se podría esperar. Probablemente la explicación sea que, en nuestro país, los directores y supervisores no suelen visitar las aulas, y lo hacen más frecuentemente cuando se han presentado problemas de disciplina u otros.

El cuidado de la calidad de la información que se recaba ha dado lugar al desarrollo de técnicas complejas, cuya aplicación es importante tratándose de estudios que pueden tener fuerte impacto en decisiones de política educativa, como ocurre con las pruebas en gran escala. Ejemplos ya mencionados de estas herramientas son la Teoría o los Modelos de Respuesta al Ítem y la Teoría de la Generalizabilidad; puede mencionarse también el análisis de los sesgos de las respuestas de los sujetos a los que se aplican instrumentos estandarizados.

En general, el dominio de estas herramientas rebasa el propósito de la formación metodológica de los posgrados de investigación educativa, por lo que solo se tratarán brevemente en el tercer apartado de este capítulo.

Técnicas básicas

Diferencias y asociación

El análisis uno a uno de los aspectos (variables) de interés es solo el primer paso de una investigación. Generalmente se querrá seguir indagando hasta llegar a explicar los

fenómenos y entender por qué ocurren, qué factores inciden en ellos, pues lo interesante es llegar a explicar, en términos causales, los cambios de una variable, lo que implica indagar si hay otra u otras a cuya influencia se deban los cambios que interesa explicar. Un primer paso en esta dirección es indagar simplemente si hay algún tipo de relación entre las variables involucradas. El caso más sencillo es el de la relación entre solo dos variables, una de las cuales pudiera incidir en la otra. Hay dos formas básicas de plantear preguntas en este sentido:

- Si hay o no *diferencia* entre las distribuciones de una variable en dos o más grupos de sujetos definidos con base en otra variable.
- Si dos variables se comportan en forma similar, si aumentan o disminuyen en paralelo, o sea si hay *asociación* entre ellas.

Estas dos formas de plantear la pregunta son en el fondo la misma (Vogt., 2007: 132), pero cada una da lugar a una familia de técnicas particular.

Estudio de las diferencias entre dos variables

La primera familia, que aplica los principios de las pruebas de hipótesis, comprende pruebas de significatividad como la t de Student, con la que se puede indagar qué tan probable es que la diferencia que se encuentre entre las medias de una variable en dos o más muestras de sujetos se deban o no al azar. Como en toda medición hay márgenes de error, si se quiere hacer conclusiones sobre una población a partir de una muestra hay que aplicar los principios de la inferencia estadística para estimar la probabilidad de que un resultado sea significativo o no, dentro de cierto intervalo de confianza.

La prueba conocida como *t de Student* (seudónimo del estadístico William Gosset, que la propuso en 1908) se basa en una distribución de probabilidad con la que se puede estimar la media de una población distribuida normalmente cuando el tamaño de la muestra es pequeño. Cuando la muestra es grande la distribución de Student es casi idéntica a la normal, pero con muestras pequeñas puede diferir bastante.

Para estimar la probabilidad de que la diferencia entre las medias de dos muestras independientes se deba o no al azar hay que calcular el valor t, con una fórmula que relaciona la diferencia entre las dos medias X_1 y X_2 (en el numerador) con la proporción de la variación que hay entre los dos grupos y dentro de cada uno de ellos (en el denominador); esto último implica el cálculo de las varianzas s_{21} y s_{22} que son los

promedio de los cuadrados (*mean square*) de las desviaciones respecto a la media (ver el inciso 1.2.). (Salkind, 2007: 192-199)

$$t = \frac{X_1 - X_2}{\sqrt{\left[(n_1 - 1)\, s^2_1 + (n_2 - 1)\, s^2_2\right]\left[n_1 + n_2\right]}{n_1 + n_2 - 2n_1 n_2}}$$

Para saber la probabilidad de que el valor t así obtenido pueda deberse al azar, se contrasta con el valor que corresponda de la distribución *t de Student*, teniendo en cuenta el número de valores que pueden variar libremente (*grados de libertad*, que dependen en general del tamaño de la muestra), y en una prueba de diferencia de medias es igual al tamaño de cada muestra menos 1 (n1 — 1 + n2 — 1). Además, hay que usar la distribución para pruebas "de dos colas" (*two-tailed*), ya que no tenemos una estimación a priori de en qué sentido se dará la diferencia, si es que la hay.

Antes de la difusión de las computadoras esa verificación se hacía consultando una tabla con los valores de t según los grados de libertad y el nivel de probabilidad que el investigador definiera como aceptable, por lo general 0.10, 0.05 y 0.01. Hoy los paquetes de software permiten calcular en segundos promedios y varianzas, así como el valor t, y también saber la probabilidad exacta de que un valor t encontrado al contrastar dos medias se deba o no al azar.

Un ejemplo: se aplica una prueba a dos muestras de 100 sujetos. En una muestra la media del puntaje fue 555 y la varianza 2025 (s = 45); en la otra la media fue 445 puntos y la varianza s2 1225 (s = 35). La diferencia entre las dos medias fue, pues, de 110 puntos (555 — 445). Para aplicar la prueba t de Student calcularemos:

$$t = \frac{555 - 445}{\frac{99x2025+99x1225}{198} \times \frac{200}{10000}} = \frac{110}{\frac{321750}{198} \times 0.02} = \frac{110}{32.5} \qquad t = 3.385$$

Según la distribución t para una prueba de dos colas, con 198 grados de libertad, el valor necesario para rechazar la hipótesis nula con probabilidad 0.01 es ~2.6, menor al valor de t en el ejemplo (3.385) por lo que se puede concluir que la diferencia que se observó entre las medias de las dos muestras del ejemplo no se debe al azar.

Con muestras no independientes (*v. gr.* un grupo al que se aplica dos veces una prueba) para calcular el valor t hay que dividir la suma (S) de todas las diferencias entre

grupos y dividirla por la raíz cuadrada del cociente que resulta de dividir la suma de diferencias al cuadrado, menos el cuadrado de la suma de diferencias, entre el número de pares de observaciones menos 1: $T = \Sigma D / \sqrt{\Sigma D^2 - (\Sigma D)^2 / n - 1}$.

En este caso se usará la distribución para pruebas "de una cola" (*one-tailed*), porque puede hipotizarse, por ejemplo, que los resultados en la prueba posterior serán mejores que en la prueba previa. (Salkind, 2007: 211-217)

Estudio de la asociación entre dos variables

La segunda familia de técnicas para explorar la relación entre dos variables utiliza medidas de asociación. Esta manera de abordar la cuestión tiene la ventaja de que hay técnicas particulares para los diversos niveles de medición de las variables. En los incisos de este apartado se presentan técnicas para estudiar la asociación entre variables nominales, ordinales y métricas, recomendaciones para la interpretación de los coeficientes resultantes y representaciones gráficas de la asociación.

Luego regresaremos a la cuestión de la relación entre las dos familias de técnicas, con la extensión de la prueba de diferencia de medias en el Análisis de Varianza, y de las medidas de correlación con los diversos modelos de regresión.

▪ Asociación en el caso de variables dicotómicas

Imaginemos un estudio sobre una población cuyos integrantes tienen un valor u otro de dos variables dicotómicas, nominales como el sexo masculino-femenino, o bien ordinales o métricas, pero tratadas como dicotómicas. Su relación se puede expresar con una tabla de dos columnas y dos renglones (*tabla de contingencia*) cuyas columnas corresponderán a valores de una variable y los renglones a la otra.

TABLA 4.10. TABLA DE CONTINGENCIA VACÍA

Variable dependiente	Variable independiente		Totales de renglón
	Valor A	Valor B	
Valor C			
Valor D			
Totales de columnas			Total General

FUENTE: ELABORACIÓN PROPIA.

En cada una de las cuatro casillas se podrá anotar el número de casos que presente la combinación correspondiente de valores de las dos variables. Si los valores de la variable registrada en columna son A y B, y los de la registrada en renglón son C y D, las combinaciones será AC, AD, BC y BD. Además, podemos conocer los totales de columna y renglón (la cantidad de sujetos que tienen los valores A, B, C y D) sin saber cuántos tienen cada una de las cuatro combinaciones. Un ejemplo imaginario:

TABLA 4.11. TABLA DE CONTINGENCIA CON TOTALES MARGINALES

Variable dependiente	Variable independiente		Totales
	Valor A	Valor B	
Valor C			50
Valor D			50
Totales	50	50	100

FUENTE: ELABORACIÓN PROPIA.

Si conocemos los totales marginales (que no tienen por qué ser iguales como en la Tabla 4.11), si se conoce el valor de cualquiera de las cuatro casillas interiores que en la tabla están en blanco, los valores de las otras tres casillas automáticamente quedan definidos. Esto es lo que se quiere decir la expresión de que una tabla de 2 x 2 tiene solo un *grado de libertad*.

Una estimación de la fuerza de la asociación entre dos variables dicotómicas es el estadístico llamado X^2 (ji o chi cuadrada). Para obtenerlo es necesario calcular primero los valores que irían en cada una de las cuatro casillas de la tabla de contingencia si se definieran en forma proporcional a los totales marginales. En una tabla con total general de 100 y totales marginales de 50 y 50 tanto en las columnas como en los renglones, si se distribuyen en forma proporcional (lo que se denomina la frecuencia esperada, fe) las cuatro casillas tendrían 25 casos cada una. Luego hay que calcular la diferencia entre los casos reales de cada combinación (la frecuencia observada, fo) y la esperada (fe). Si los valores reales fo de cada casilla fueran 40, 10, 10, 40 (en negritas en Tabla 4.12), las diferencias entre frecuencias observadas y esperadas de las cuatro casillas serían 15, -15, 15 y -15:

TABLA 4.12. TABLA PARA EL CÁLCULO DE LA X^2

$\chi^2 = 36$

Variable dependiente	Variable independiente		Totales
	Valor A	Valor B	
Valor C	**40** — 25 = 15	**10** - 25= -15	50
Valor D	**10** - 25 = -15	**40** - 25= 15	50
Totales	50	50	100

FUENTE: ELABORACIÓN PROPIA.

La ji cuadrada se define como la suma, elevada al cuadrado, de las diferencias entre las frecuencias observadas y las esperadas, divididas cada vez por la frecuencia esperada: $X^2 = \Sigma (fo - fe)^2 / fe$. En la Tabla 16 las cuatro cantidades a sumar, a partir de los datos de cada casilla, son iguales a 9: 15 o -15 al cuadrado = 225 entre 25. La suma, o sea el valor de la X^2 que corresponde a esos datos, es 36.

Esto es aplicable a tablas de contingencia con más de dos columnas y/o renglones, con cualquier combinación de variables politómicas. A partir de X^2 (prueba del ajuste entre una distribución empírica u observada y una teórica o esperada, se puede construir una medida de la fuerza de la relación entre dos variables. S entiende que, si 100 casos se distribuyen por igual en las cuatro casillas, con 25 para cada una de las cuatro combinaciones, significa que tener uno u otro valor en una variable no tiene que ver con el valor de la otra, *que no hay relación entre las variables*:

TABLA 4.13. EJEMPLO DE TABLA QUE MUESTRA AUSENCIA DE RELACIÓN

Variable dependiente	Variable independiente		Totales
	Valor A	Valor B	
Valor C	25	25	50
Valor D	25	25	50
Totales	50	50	100

FUENTE: ELABORACIÓN PROPIA.

En cambio, si los casos se concentran en dos de las cuatro casillas, con 50 en dos de las cuatro combinaciones en dos de las esquinas de la tabla, y o en las otras dos esquinas, eso querrá decir que tener un valor en una variable implica tener cierto valor en la otra, o sea que hay una relación perfecta entre las dos variables:

TABLA 4.14. EJEMPLO DE TABLA QUE MUESTRA RELACIÓN PERFECTA

Variable dependiente	Variable independiente		Totales
	Valor A	Valor B	
Valor A	50	0	50
Valor B	0	50	50
Totales	50	50	100

FUENTE: ELABORACIÓN PROPIA.

La X^2 aumenta si hay asociación entre las variables, pero también con el número de casos (N). Si la calculamos para la tabla 19, similar a la 16 con datos diez veces mayores, la c2 es también 10 veces mayor, aunque la intensidad de la asociación entre las dos variables sea la misma: las cifras a sumar a partir de los datos de cada casilla son 90: 150 o -150 al cuadrado = 22,500 entre 250. La suma, la X^2, es 360.

TABLA 4.15. TABLA PARA EL CÁLCULO DE LA X^2

$X^2 = 360$

Variable dependiente	Variable independiente		Totales
	Valor A	Valor B	
Valor C	400 — 250 = 150	100 - 250= -150	500
Valor D	100 - 250 = -150	400 - 250= 150	500
Totales	500	500	1000

FUENTE: ELABORACIÓN PROPIA.

Esto lleva a pensar que si se divide la X^2 entre N (el total de casos) se puede obtener una medida de la asociación entre las variables de la tabla. En caso de relación perfecta toma el valor de 1; si la relación es nula, 0. En los demás casos oscila entre 1 y 0, más cerca de 1 cuanto mayor sea la relación: es el coeficiente de correlación $\phi 2$ de Karl Pearson, que lo propuso a fines del siglo XIX. Las tablas siguientes son ejemplos de asociación perfecta, nula, fuerte y débil, con cifras imaginarias de 100 alumnos, mitad hombres y mitad mujeres, y que también por mitad habrían o no desertado de la escuela. En cada caso se dan los valores de X^2 y $\phi 2$.

TABLA 4.16. EJEMPLOS DE ASOCIACIONES DE DIFERENTE INTENSIDAD

ASOCIACIÓN NULA

$X^2 = 0 \quad \Phi^2 = 0$

Variable dependiente	Género		Totales
	Masculino	Femenino	
Desertó	25	25	50
No desertó	25	25	50
Totales	50	50	100

ASOCIACIÓN PERFECTA

$X^2 = 100 \quad \Phi^2 = 1$

Variable dependiente	Género		Totales
	Masculino	Femenino	
Desertó	50	0	50
No desertó	0	50	50
Totales	50	50	100

ASOCIACIÓN FUERTE

$X^2 = 36 \quad \Phi^2 = 0.36$

Variable dependiente	Género		Totales
	Masculino	Femenino	
Desertó	40	10	50
No desertó	10	40	50
Totales	50	50	100

ASOCIACIÓN DÉBIL

$X^2 = 4 \quad \Phi^2 = 0.04$

Variable dependiente	Género		Totales
	Masculino	Femenino	
Desertó	30	20	50
No desertó	20	30	50
Totales	50	50	100

FUENTE: ELABORACIÓN PROPIA.

En los cuatro ejemplos es posible apreciar intuitivamente que hay una asociación nula o perfecta, fuerte o débil, pero el coeficiente ϕ^2 permite una apreciación más clara, y no depende del número de casos, como la X^2.

Hay variantes del coeficiente $\phi 2$ para tablas de contingencia que tengan más de dos columnas y/o renglones (por ejemplo, V de Kramer, T de Tschuprov, etc.). Los paquetes de software permiten también calcularlos.

Asociación en el caso de variables ordinales

Cuando se tienen datos de dos variables ordinales de un conjunto de sujetos es posible, desde luego, ordenarlos del primero al último en cada una de las variables, y luego comparar el lugar que ocupa cada sujeto (su rango) en el ordenamiento resultante de cada variable.

El coeficiente de correlación de rangos r (rho), propuesto en 1904 por Charles Sperman, se calcula elevando al cuadrado la diferencia de rangos y multiplicándolo por seis; esto se divide entre el producto del número de sujetos N por su cuadrado menos 1, y el resultado finalmente se resta de uno: $\rho = 1 - 6 \sum D^2 / (N \times N^2 - 1)$.

De esta manera el coeficiente toma el valor de 1 cuando la relación es directa y perfecta; de - 1 cuando es perfecta pero inversa; y con valores intermedios en los demás casos, tendiendo a cero mientras más débil es la relación, como se muestra en ejemplos imaginarios sobre la correlación de los resultados de diez alumnos en un examen y sus respectivos promedios, en términos ordinales.

TABLA 4.17. EJEMPLO DE ASOCIACIÓN DIRECTA PERFECTA ENTRE VARIABLES ORDINALES

$\rho = -0.01$

Caso	Rango en examen	Rango en promedio	Diferencia de rangos (D)	D^2
Abelardo	1°	1°	0	0
Bernardo	2°	2°	0	0
Carlos	3°	3°	0	0
Daniel	4°	4°	0	0
Enrique	5°	5°	0	0
Francisco	6°	6°	0	0
Gabriel	7°	7°	0	0
Héctor	8°	8°	0	0
Ignacio	9°	9°	0	0
Javier	10°	10°	0	0
			0	0

FUENTE: ELABORACIÓN PROPIA.

En este primer ejemplo los 10 sujetos tienen exactamente el mismo rango en los dos ordenamientos, por lo que todas las diferencias de rangos son iguales a cero, sus cuadrados y la suma de estos es también de cero, que dividido entre 990 (103 — 10) sigue dando cero como resultado, que al restarse de 1 lleva a un coeficiente rho que significa que hay una correlación perfecta.

TABLA 4.18. EJEMPLO DE ASOCIACIÓN INVERSA PERFECTA ENTRE VARIABLES ORDINALES

Caso	Rango en examen	Rango en promedio	Diferencia de rangos (D)	D^2
Abelardo	1°	10°	-9	81
Bernardo	2°	9°	-7	49
Carlos	3°	8°	-5	25
Daniel	4°	7°	-3	9
Enrique	5°	6°	-1	1
Francisco	6°	5°	1	1
Gabriel	7°	4°	3	9
Héctor	8°	3°	5	25

Ignacio	9°	2°	7	49
Javier	10°	1°	9	81
			0	330

FUENTE: ELABORACIÓN PROPIA.

Ahora el rango de los imaginarios alumnos en cuanto a promedio es justo el opuesto al que tienen en el examen, con lo que las diferencias son máximas, positivas y negativas. Tras elevarlas al cuadrado para que no se anulen, y multiplicar por seis la suma de esos cuadrados de diferencias de rangos (330) obtenemos 1980, que se divide entre 990 (103 - 10) lo que arroja como resultado 2, que restado de 1 lleva a un coeficiente rho que representa de nuevo una correlación perfecta, pero inversa.

En el ejemplo siguiente los rangos en las dos variables no tienen relación.

TABLA 4.19. EJEMPLO DE ASOCIACIÓN NULA ENTRE VARIABLES ORDINALES

$\rho = -0.01$

Caso	Rango en examen	Rango en promedio	Diferencia de rangos (D)	D^2
Abelardo	1°	3°	-2	4
Bernardo	2°	7°	-5	25
Carlos	3°	5°	-2	4
Daniel	4°	10°	-6	36
Enrique	5°	2°	3	9
Francisco	6°	8°	-2	4
Gabriel	7°	9°	-2	4
Héctor	8°	1°	7	49
Ignacio	9°	6°	3	9
Javier	10°	4°	6	36
			0	180

FUENTE: ELABORACIÓN PROPIA.

En este último ejemplo, al multiplicar por seis la suma de las diferencias de rangos al cuadrado (180) obtenemos 1080, que se divide entre 990 (103 — 10) lo que arroja como resultado 1.09, que restado de 1 lleva a un coeficiente rho prácticamente nulo.

Tratándose de distribuciones con suficiente número de sujetos y sin muchos casos extremos, el coeficiente rho de Spearman arroja resultados similares al que se presenta

en seguida (r de Pearson), que fue diseñado para variables medidas a nivel métrico, por lo que es frecuente que se use este último en vez de aquel.

▪ Asociación en el caso de variables métricas

Además de la prueba X^2 y del coeficiente ϕ^2, Karl Pearson, sin duda uno de los más importantes pioneros de la estadística moderna, propuso otras herramientas que siguen siendo esenciales para analizar datos cuantitativos: la desviación estándar, el coeficiente de correlación r, y las bases de la regresión lineal.

Para entender el coeficiente r, pensemos en el ejemplo de la estatura y el peso de las personas, dos variables que pueden medirse a nivel de razón y tienen una clara relación, aunque no perfecta: mientras más alta sea una persona tenderá a pesar más, pero hay personas altas y delgadas que pesarán menos que bajas y gruesas. La Tabla 4.20 muestra intuitivamente la relación entre estatura y peso en casos claros, pero Pearson construyó una medida de correlación a partir de las medidas de variación, que muestran el grado en que los sujetos se apartan de la media.

TABLA 4.20. APROXIMACIÓN INTUITIVA A LA RELACIÓN ENTRE ESTATURA Y PESO

Caso	Estatura	Peso	Relación
1	Alto	Pesado	Positiva
2	Bajo	Liviano	Positiva
3	Promedio	Alto	Nula
4	Promedio	Bajo	Nula
5	Alto	Liviano	Negativa
6	Bajo	Pesado	Negativa

FUENTE: ELABORACIÓN PROPIA.

Para ir más allá de una apreciación intuitiva tan burda como la de la tabla anterior, Pearson propuso una forma de calcular una medida de la importancia de la relación entre variables como la estatura y el peso que, a más de cien años de distancia, sigue siendo fundamental en los análisis estadísticos, y cuya ingeniosidad no deja de sorprender cuando se analiza.

El punto de partida es la idea de que, cuando los valores de dos variables están asociados, la diferencia del valor de cada variable respecto a la media de su propia distribución será similar: en ambos casos positiva para la combinación alto-pesado,

o en ambos casos negativa, en el caso de bajo-liviano. Por otra parte, sabemos que la multiplicación de dos valores positivos, al igual que de dos valores negativos, da un resultado positivo (*más por más da más, y menos por menos también da más*), mientras que el producto de un valor positivo y uno negativo es negativo.

A partir de lo anterior la ingeniosa idea que permitió la construcción del coeficiente de correlación r es que cuando haya una asociación fuerte entre dos variables, la suma del producto de los cuadrados de las diferencias (ambas positivas o ambas negativas) respecto a sus respectivas medias será grande; que cuando haya una asociación fuerte pero negativa (alto-liviano, bajo-pesado), una de las diferencias será positiva y una negativa, por lo que el producto será negativo y la suma será también grande pero negativa; y que cuando no haya una asociación clara, unas de las diferencias serán positivas y otras negativas, los productos también, y la suma tenderá a anularse.

La tabla siguiente muestra los elementos necesarios para calcular el coeficiente r:

TABLA 4.21. CÁLCULO DEL COEFICIENTE R

Caso ($_i$)	Estatura (x_i)	Peso (y_i)	(x_i -x)	(x_i -x)2	(y_i -y)	(y_i -y)2	(x_i -x) (y_i -y)
1	155	48	- 6	36	- 5.3	28.09	31.8
2	156	45	- 5	25	- 8.3	68.89	41.5
3	162	52	1	1	- 1.3	1.69	-1.3
4	167	56	6	36	2.7	7.29	16.2
5	160	57	- 1	1	3.7	13.69	-3.7
6	162	56	1	1	2.7	7.29	2.7
7	158	47	- 3	9	- 6.3	39.69	18.4
8	159	59	- 2	4	5.7	32.49	- 11.4
9	168	59	7	49	5.7	32.49	39.9
10	163	54	2	4	.7	.49	1.4
Σ	1610	533	0	166	0	232.1	136
Σ/N	x = 161	y = 53.3	---	σ^2=16.6	---	σ^2=23.21	13.6
				σ_x = 4.07		σ_y = 4.82	

FUENTE: ELABORACIÓN PROPIA.

A partir de los datos de estatura (x_i) y peso (y_i), se pueden calcular las diferencias de cada valor de estatura o peso respecto a su respectiva media (x_i-x y y_i-y) y luego los

cuadrados de estas diferencias $(x_i-x)^2$ y (y_i-y^2); se puede calcular también el producto de las diferencias (x_i-x) (y_i-y).

Una primera medida de asociación, la *covarianza*, es simplemente el promedio de estos productos de las diferencias de los valores del mismo sujeto en las dos variables respecto a la media correspondiente.

$COVxy = \Sigma (x_i - x) (y_i - y) / N$. En el ejemplo, 136/10 = 13.6

Mientras más grande es este valor, la relación entre las dos variables es más fuerte, pero como ocurría con la X^2, los valores son difíciles de interpretar porque no tienen un valor máximo que sirva como referente. Por ello Pearson propone una forma de estandarizar la covarianza para que adopte valores máximos de + 1 y -1 cuando la asociación entre las variables en cuestión es perfecta, directa o inversa, y un valor de 0 cuando no hay asociación. Esa estandarización se consigue dividiendo la covarianza entre el producto de las desviaciones estándar de las dos variables, y el resultado es precisamente el coeficiente de correlación $Rxy = COVxy / \sigma_x \times \sigma_y$.

En el ejemplo r = 13.6/4.07 x 4.82, o sea 13.6/19.62 = 0.693

Si se eleva al cuadrado un coeficiente de correlación r se obtiene un coeficiente de determinación r^2, que representa la proporción de la varianza que comparten las dos variables. Como r suele ser menor a 1, normalmente $r^2 < r$.

▪ Interpretación de los coeficientes de correlación

Los valores extremos de un coeficiente de correlación (0 y 1) se pueden interpretar en forma clara, porque significan que no hay relación alguna entre las variables en cuestión, o bien que la relación es perfecta, o sea que siempre que una de las dos variables toma un valor arriba o debajo de la media, la otra se comporta de manera igual. Los valores intermedios entre 0 y 1 no son tan fáciles de interpretar.

Algunos textos de estadística proponen tablas para interpretar un coeficiente de correlación, en las que se dice, por ejemplo, que un valor menor a 0.20 significaría una correlación leve, casi insignificante; uno de 0.20 a 0.40, una correlación definida pero baja; de 0.40 a 0.70 correlación moderada, sustancial; de 0.70 a 0.90 una Correlación marcada, alta; y de 0.90 a 1.00 una correlación altísima.

Estas interpretaciones deben tomarse con cuidado. Cada coeficiente se comporta de modo diferente. El coeficiente r es sensible a cambios en los valores extremos, como la media y la desviación estándar, en las que se basa. Por otra parte, los fenómenos que se estudian con estas herramientas son también distintos. En ciencias sociales y educación suele encontrarse que unos valores del coeficiente r que, en una interpretación

convencional podrían considerarse bajos (*v.gr.* 0.40) o moderados (0.50), son los más altos que se encuentran en la realidad.

Un ejemplo de esto es la asociación entre los resultados obtenidos en las pruebas de admisión a los estudios superiores y las calificaciones posteriores de los sujetos aceptados con base en esas pruebas. Por eso, para interpretar un coeficiente de correlación, es indispensable conocer cuáles son los valores que se encuentran empíricamente, lo que permite formarse una idea más sólida de qué tan importante es una relación identificada. Los ejemplos siguientes, con datos imaginarios, permiten apreciar características del comportamiento del coeficiente r.

TABLA 4.22. ASOCIACIÓN ENTRE PUNTAJE EN PRUEBA DE ADMISIÓN Y PROMEDIO DE CALIFICACIONES POSTERIORES: RELACIÓN PERFECTA DIRECTA E INVERSA, Y RELACIÓN NULA

Sujetos	Puntaje en prueba	Promedio de calificaciones		
		Ejemplo 1	Ejemplo 2	Ejemplo 3
Abelardo	1300	10.00	06.00	08.00
Bernardo	1250	09.66	06.33	08.00
Carlos	1200	09.33	06.66	08.00
Daniel	1150	09.00	07.00	08.00
Enrique	1100	08.66	07.33	08.00
Francisco	1050	08.33	07.66	08.00
Gabriel	1000	08.00	08.00	08.00
Héctor	950	07.66	08.33	08.00
Ignacio	900	07.33	08.66	08.00
Javier	850	07.00	09.00	08.00
Karl	800	06.66	09.33	08.00
Luis	750	06.33	09.66	08.00
Miguel	700	06.00	10.00	08.00
Media	1000	08.00	08.00	08.00
r		+ 1.00	- 1.00	0.00

FUENTE: ELABORACIÓN PROPIA.

Los datos del Ejemplo 1 del promedio de calificaciones tienen una relación directa perfecta con los puntajes en la prueba, y el coeficiente de correlación entre ambas variables tiene en ese caso el valor +1. Los promedios de calificaciones del Ejemplo 2 tienen una relación también perfecta, pero inversa, con los puntajes de la prueba, y el valor del coeficiente r en ese caso es igual a -1. Por su parte, los promedios del Ejemplo 3, todos iguales a 8.0, son un caso evidente de ausencia completa de asociación con los puntajes de la prueba, y el coeficiente r es igual a cero.

El Ejemplo 3, en que todos los sujetos tienen un mismo valor, es un caso en que no estamos ante una variable, sino una constante, e ilustra la idea de que la correlación entre una variable y cualquier constante será siempre nula, igual a cero.

Una sencilla reflexión permite entender lo anterior: el valor de la constante no tiene nada que ver con los valores de la variable; sean estos altos o bajos, el valor de la constante es el mismo: no hay relación alguna entre la variable y la constante. Puede hacerse el ejercicio de sustituir el valor 8.0 del Ejemplo 3 por cualquier otro, siempre que sea el mismo para todos los casos (constante), y el valor del coeficiente r será siempre de 0.

La tabla siguiente presenta dos ejemplos imaginarios más, que permiten apreciar la sensibilidad del coeficiente r a cambios en valores extremos.

TABLA 4.23. ASOCIACIÓN ENTRE PUNTAJE EN PRUEBA DE ADMISIÓN Y PROMEDIO DE CALIFICACIONES POSTERIORES: EJEMPLOS CON CASOS ATÍPICOS

Sujetos	Puntaje en prueba	Promedio de calificaciones	
		Ejemplo 1	Ejemplo 2
Abelardo	1300	10.00	06.00
Bernardo	1250	09.66	09.66
Carlos	1200	09.33	09.33
Daniel	1150	09.00	09.00
Enrique	1100	08.66	08.66
Francisco	1050	08.33	08.33
Gabriel	1000	08.00	08.00
Héctor	950	07.66	07.66
Ignacio	900	07.33	07.33
Javier	850	07.00	07.00
Karl	800	06.66	06.66
Luis	750	06.33	06.33
Miguel	700	10.00	10.00
Media	1000	08.30	08.00
r		+ 0.623	+ 0.209

FUENTE: ELABORACIÓN PROPIA.

Los promedios de calificaciones del Ejemplo 1 son casi iguales a los del Ejemplo 1 de la tabla anterior, con excepción del último caso, el del imaginario Miguel, que en vez de tener la calificación más baja de todas (6.0), tendría una de 10.0.

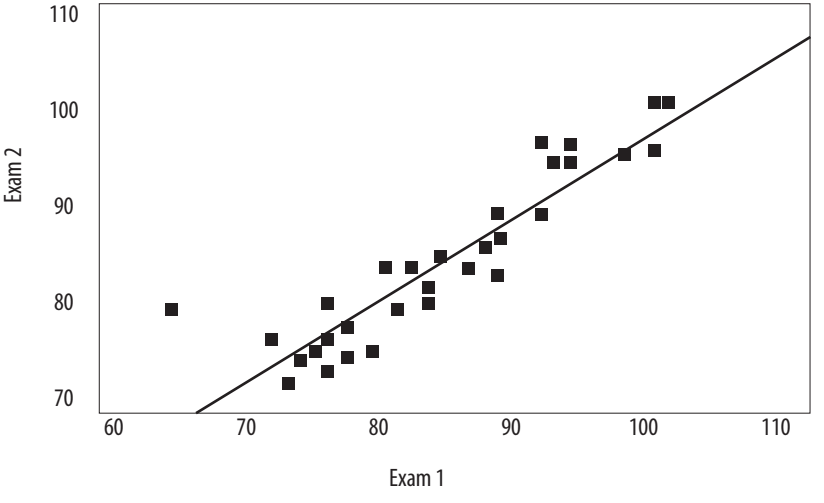
El efecto de este cambio en un solo caso extremo sobre el coeficiente r es fuerte: en vez de + 1 es de + 0.623. En el Ejemplo 2 el cambio de los dos valores extremos (Abe-

lardo con 6.0 en vez de 10.0, y Miguel con 10.0 en vez de 6.0) se refleja en la drástica reducción del coeficiente r a + 0.209.

Representación gráfica de la asociación

Las gráficas usadas para representar asociación o correlación entre dos variables son los diagramas de dispersión bidimensionales. En estos, los dos valores de cada caso de las variables se representan por la intersección de dos coordenadas, una que parte del eje horizontal, o de las x, de un plano cartesiano (abscisa), y otra del eje vertical o de las y (ordenada). Cuando la *nube de puntos* que resulta al ubicar en el plano los pares de valores de todos los casos sigue de manera aproximada una línea recta (*recta de regresión*), aunque no se conozca el valor de la correlación que correspondería a los valores se puede concluir que hay asociación o correlación fuerte entre las variables de que se trate, ya que el agrupamiento aproximadamente rectilíneo de los puntos indica que cuando un caso tiene un valor alto (o bajo, o medio) en una variable, tiende a tener también un valor alto (o bajo, o medio) en la otra, como muestra la Gráfica 15.

GRÁFICA 4.16. CORRELACIÓN FUERTE ENTRE LOS RESULTADOS DE DOS EXÁMENES

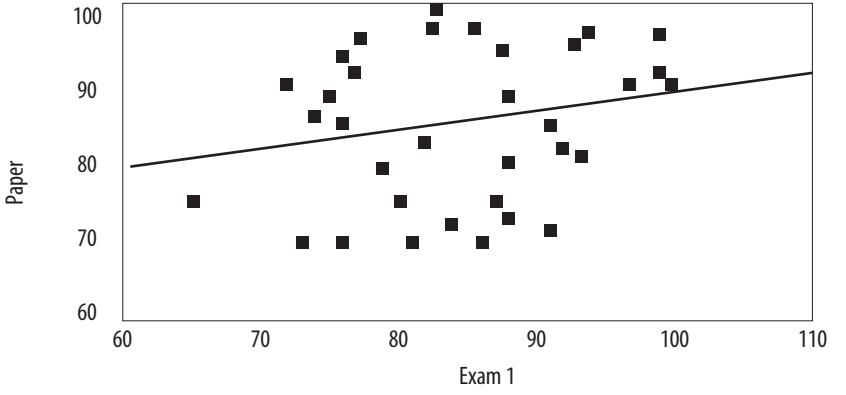


FUENTE: VOGT, 2007: 33. FIGURA 2.2.

Si la nube de puntos es dispersa, no marca con claridad una recta (aunque se puede estimar la que más se aproxima a los puntos) se puede concluir que la relación entre las

variables es débil, pues cuando hay un valor alto (o medio, o bajo) en una variable el valor en la otra no coincide con el primero.

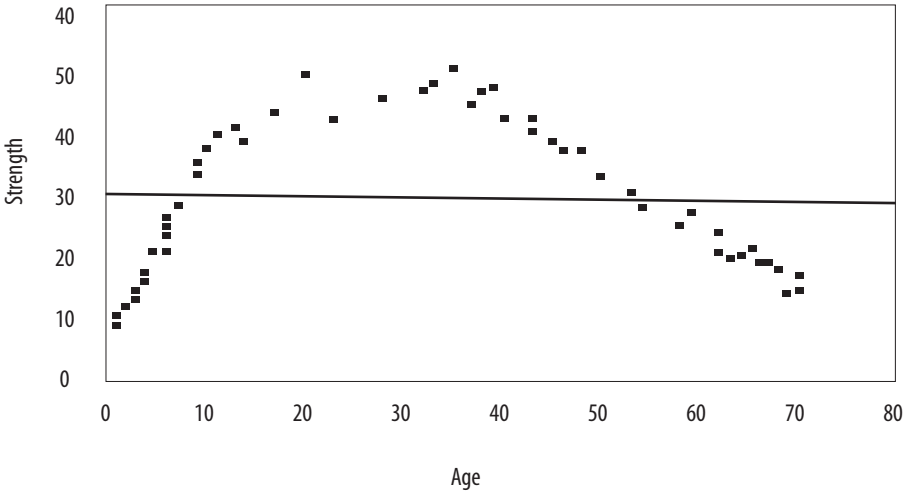
GRÁFICA 4.17. CORRELACIÓN DÉBIL ENTRE LOS RESULTADOS DE DOS EXÁMENES



FUENTE: VOGT, 2007: 34. FIGURA 2.3.

Un diagrama de dispersión bidimensional permite también apreciar si hay relación entre dos variables, pero no rectilínea. La Gráfica 17 muestra la relación entre la edad de unas personas y su fuerza física. Hay relación lineal positiva clara a partir del nacimiento hasta cierta edad (entre 30 y 40 años), lapso en el que al aumentar la edad se incrementa también la fuerza. Luego la relación sigue siendo fuerte, pero se invierte: a medida que una persona envejece su fuerza disminuye. Un coeficiente de correlación r no permite apreciar este tipo de relación, para lo que pueden servir otros coeficientes, pero la gráfica lo muestra con claridad. Vogt (2007: 69-72) señala que el coeficiente r de una distribución así es casi nulo (0.03), pero un coeficiente no lineal (regresión cuadrática) tiene un valor cercano a 1 (0.95).

GRÁFICA 4.18. CORRELACIÓN CURVILÍNEA ENTRE EDAD Y FUERZA



FUENTE: VOGT, 2007: 35. FIGURA 2.4.

Otra manera de representar gráficamente la relación entre dos variables hace uso de vectores, como la representación de una variable que tiene magnitud y dirección. En este caso, si hay relación directa perfecta entre dos variables, los vectores que las representan se orientan en la misma dirección, por lo que el ángulo que forman es igual a cero grados (0°); si la relación es perfecta pero inversa, se orientan en dirección opuesta, formando un ángulo de 180°; si la relación entre ambas es nula, forman un ángulo de 90°, un ángulo recto, por lo cuando no hay correlación alguna entre dos variables se dice que tienen una relación *ortogonal*.

Avances en busca de explicaciones

Un viejo principio del análisis estadístico, que tiene casi valor axiomático, es el que establece que correlación no implica necesariamente causalidad. Para sustentar esta idea suelen presentarse ejemplos que la ilustran.

Se cuenta que un estudio encontró alta correlación entre el número de nacimientos registrados en las poblaciones de Francia y el número de cigüeñas que llegaban a anidar en las iglesias de cada población. Se señala que la importancia de los daños producidos por los incendios ocurridos en cierta ciudad en un año mostraba alta correlación con el número de bomberos que había acudido a apagar cada incendio.

Tales correlaciones, que pueden ser altas, no se deben interpretar en términos causales: ni las cigüeñas tienen que ver con los nacimientos de niños, ni los daños de los incendios se deben a los bomberos. La interpretación es otra: las cigüeñas anidan más frecuentemente en torres de iglesias de pueblos pequeños, en los que la tasa de natalidad puede ser más alta; y la gravedad de un incendio explica tanto la importancia de los daños que produce, como el número de bomberos que acuden.

Este principio, que hoy parece tan claro, tardó mucho tiempo en entenderse, y quien lo planteó por primera vez de manera explícita y rigurosa fue Paul Lazarsfeld, en una presentación a la *American Sociological Society* en 1946, que se publicó en la famosa obra *The Language of Social Research* (Lazarsfeld y Rosenberg, 1955).

En una amplia reflexión sobre el trabajo de Lazarsfeld, Raymond Boudon señala que "ideas que hoy parecen relativamente simples tardaron prácticamente un siglo en perfilarse", desde Stuart Mill hasta autores contemporáneos de Lazarsfeld como el criminólogo Edwin Sutherland, pasando de manera destacada por Durkheim. (Boudon, 1970: 53). Según Boudon:

> Durkheim muestra algo de lo que antes de él no se tenía una idea clara: que una relación estadística no puede, salvo excepción, interpretarse en forma directa como una relación causal, ya que puede ser la consecuencia de relaciones más complejas. (Boudon, 1970: 12)

Boudon añade que curiosamente Durkheim no captó también que una correlación nula no debe interpretarse tampoco automáticamente como ausencia de relación causal, ya que esto no siempre es así, como Lazarsfeld sí mostró. (1970: 53)

Conviene añadir que tampoco se puede concluir que en ningún caso hay relación causal entre dos variables correlacionadas. Para afirmar algo respecto a una posible relación de causalidad es necesario analizar correlaciones, pero no basta. Esto es lo que se trata de mostrar en el siguiente apartado, de la mano de Lazarsfeld, y como han hecho otros. (Boudon, 1969; Cortés y Rubalcava, 1987)

Correlación espuria y control de variables

La noción clave que hay que entender es la de correlación espuria, una de las más importantes de la metodología, que no necesariamente se enseña a los estudiantes de todos los posgrados que pretenden formar investigadores, como se sugirió en la Presentación de esta obra, citando a William Spady (1970), cuando dice que "si un egresado de un posgrado en ciencias sociales termina sus estudios sin haber aprendido al menos a de-

tectar el posible carácter espurio de una correlación, debería regresar a la institución y pedir que le devuelvan la colegiatura".

Según el diccionario, el calificativo espurio se aplica a un hijo ilegítimo, pero no a cualquiera, sino a aquel cuyo padre no es el esposo de la madre sino otro varón, y del que además muchas personas piensan que sí es hijo del legítimo esposo.

Una correlación espuria, análogamente se da cuando el cambio de una variable se interpreta erróneamente, en términos causales, como efecto de una variable con la que la primera está correlacionada, cuando la causa real del efecto es otra.

Puede haber correlación entre muchas variables, pero ninguna puede interpretarse sin más en términos causales. Para hacerlo *hay que detectar su posible carácter espurio*, para lo que es necesario analizar el efecto de otras variables que puedan ser las que realmente expliquen la correlación.

Lazarsfeld fue el primero en mostrar cómo hacer esto de manera sistemática, con su *modelo de elaboración*, o de Columbia. La lógica para detectar el posible carácter espurio de una correlación consiste en introducir una tercera variable y ver qué pasa con la correlación original entre las dos variables cuando se controla la tercera.

En 1946 Lazarsfeld presentó la forma más simple de este proceso, con variables dicotómicas. Hoy las técnicas multivariadas permiten analizar la correlación entre muchas variables al tiempo que se controlan muchas más, pero la lógica es la misma, y para su comprensión es útil considerar su versión más simple, el ejemplo original, con los datos de Lazarsfeld de un estudio sobre audiencias radiofónicas.

El análisis parte de tablas de contingencia con dos columnas y dos renglones, pero en lugar de φ_2, por tratarse de variables métricas dicotomizadas se usa el coeficiente de correlación tetracórico (r_t), una aproximación del r de Pearson apropiada para tablas de 2 x 2. Lazarsfeld empleó el producto cruzado estandarizado de los valores de las cuatro casillas de cada tabla de contingencia (1955: 116)

El estudio en que se basó Lazarsfeld cubrió una muestra de 2,300 personas, de las cuales 1,000 fueron clasificadas como jóvenes y 1,300 como viejas, en una versión dicotomizada de su edad. Se preguntó también por el nivel de instrucción de los sujetos, tratado también de manera dicotómica como alto o bajo, y se encontró que 1,000 personas tenían nivel alto y las otras 1,300 un nivel bajo.

La tabla de contingencia simple 4.24 muestra la manera en que se relacionan las variables edad y nivel de instrucción.

TABLA 4.24. RELACIÓN DE LA EDAD Y EL NIVEL DE INSTRUCCIÓN

Nivel de instrucción	Jóvenes	Viejos	Total
Alto	600	400	1000
Bajo	400	900	1300
Total	1000	1300	2300

FUENTE: LAZARSFELD, 1955.

Se interrogó a los 2,300 sujetos sobre los programas radiofónicos que preferían escuchar, distinguiendo los de contenido religioso, político y musical, en forma dicotómica: escuchan o no escuchan.

Como la edad y el nivel de instrucción podrían influir sobre la preferencia por uno u otro tipo de programa, esas dos variables se pueden considerar independientes, y la preferencia por uno u otro tipo de programa como variable dependiente.

La Tabla 4.25 presenta las cifras de quienes dijeron escuchar o no cierto tipo de programa, distinguiendo jóvenes y viejos; la Tabla 4.26 hace lo mismo distinguiendo personas de nivel de instrucción alto y bajo.

TABLA 4.25. PREFERENCIA POR CIERTOS PROGRAMAS, POR EDAD

Audiencia de cierto programa	Edad		Totales
	Jóvenes	Viejos	
Progr. religiosos			
Escuchan	170	338	508
No escuchan	830	962	1792
Totales	1000	1300	2300
Progr. políticos			
Escuchan	340	585	925
No escuchan	660	715	1375
Totales	1000	1300	2300
Progr. musicales			
Escuchan	300	377	677
No escuchan	700	923	1623
Totales	1000	1300	2300

FUENTE: LAZARSFELD, 1955.

TABLA 4.26. PREFERENCIA POR CIERTOS PROGRAMAS, POR NIVEL DE INSTRUCCIÓN

Audiencia de cierto programa	Nivel de Instrucción		Totales
	Alto	Bajo	
Progr. religiosos			
Escuchan	100	400	500
No escuchan	900	900	1800
Totales	1000	1300	2300
Progr. políticos			
Escuchan	460	460	920
No escuchan	540	840	1380
Totales	1000	1300	2300
Progr. musicales			
Escuchan	400	283	683
No escuchan	600	1017	1617
Totales	1000	1300	2300

FUENTE: LAZARSFELD, 1955.

Con esto se puede analizar la relación simple entre la edad o el nivel de instrucción y la preferencia por cierto tipo de programa, pero también —y en esto reside la aportación de Lazarsfeld— la relación entre una de las variables independientes y la dependiente, *controlando la posible influencia de la otra variable independiente*.

Analizar la relación entre dos variables controlando una tercera implica construir *tablas de contingencia parciales* (para cada una de las cuales se puede calcular un coeficiente de correlación, en este caso de *correlación parcial*), lo que no es posible solo con los datos de las Tablas 4.25 y 4.26, pero los datos completos del estudio sí lo permiten, ya que permiten saber cuántas personas de cada combinación de edad y nivel de instrucción alto o bajo suelen escuchar cada tipo de programa.

A partir de la relación entre edad y preferencia por ciertos programas, plasmada en una tabla de contingencia simple, introducir el nivel de instrucción (NI) como variable de control implica hacer dos tablas parciales, en una de las cuales solo se incluyan las personas de NI alto, y en otra las de NI bajo, distinguiendo en cada una las preferencias de jóvenes y viejos, como sugiere sin datos la Tabla 4.27. Al hacerlo la relación entre edad y preferencia de cada una de las dos tablas parciales no tendrá influencia del nivel de instrucción, que no varía en cada tabla parcial, se ha vuelto constante, pues en una solo hay sujetos de NI alto, y en la otra solo de NI bajo.

TABLA 4.27. RELACIÓN ENTRE EDAD Y PREFERENCIA POR CIERTO TIPO DE PROGRAMA, CONTROLANDO EL NIVEL DE INSTRUCCIÓN

Audiencia de un programa	Nivel de instrucción				Totales
	Alto		Bajo		
	Jóvenes	Viejos	Jóvenes	Viejos	
Escuchan					
No escuchan					
Totales	600	400	400	900	2300

FUENTE: ELABORACIÓN PROPIA.

En las tablas del Apéndice del capítulo se pueden ver las tablas simples y parciales correspondientes. En este lugar solo se sintetiza el análisis que hizo Lazarsfeld con base en esas tablas.

▪ **Preferencia por programas religiosos**

La relación simple entre edad y preferencia por programas radiofónicos de tipo religioso es baja y negativa (r_t = - 0.2): los viejos parecerían tener mayor preferencia por esos programas, aunque no muy marcada: 26% dicen escuchar esos programas, frente a 17% de los jóvenes.

Al introducir la variable NI como control, sin embargo, los coeficientes de correlación de las dos tablas parciales siguen siendo negativos, pero menores, cercanos a cero (r_t -0.04 con nivel de instrucción alto, y -0.05 con nivel bajo). Eso sugiere que, a igual NI, alto o bajo, la relación entre edad y gusto por programas religiosos desaparece. Al parecer la relación entre edad y preferencia por programas religiosos se debe más bien a la influencia del nivel de instrucción, que seguramente es más alto entre las personas jóvenes que entre las de mayor edad.

Un análisis similar, partiendo de la relación entre nivel de instrucción y preferencia por programas de contenido religioso, introduciendo luego la edad como variable de control, permite entender mejor lo anterior.

La relación simple entre NI y preferencia por programas religiosos es más fuerte que la que se encontró entre edad y esos programas, y también negativa (r_t t = - 0.5): las personas de menor instrucción los prefieren más que los más instruidos.

Esta relación se confirma si se controla por edad: la correlación inicial negativa y clara (r_t = - 0.5) se mantiene casi idéntica en las dos tablas parciales: entre jóvenes y entre personas de más edad sigue habiendo relación negativa entre nivel de instrucción

y preferencia por programas de tipo religioso (r_t = - 0.49 y - 0.47). Esto refuerza la hipótesis de que es el nivel de instrucción lo que realmente influye en la preferencia por programas religiosos, y que la influencia de la edad no es relevante.

▪ Preferencia por programas políticos

La relación simple entre edad y preferencia por programas políticos es negativa y moderada (r_t = - 0.18). Esto parece confirmarse al controlar por NI: a instrucción igual, la relación entre edad y gusto por programas políticos sigue moderadamente negativa, con rt de - 0.23 entre personas de NI alto, y rt -0.26 entre las de NI bajo.

Por su parte, la relación simple entre preferencia por programas políticos y nivel de instrucción es positiva moderada (r_t = 0.17): los más instruidos prefieren ese tipo de programas un poco más que los de menor instrucción, lo que parece confirmarse al controlar por edad: la correlación sigue positiva y del mismo orden de tamaño en las tablas parciales: entre jóvenes r_t = 0.27), y r_t = 0.23 entre los de más edad.

En este caso, pues, tanto si se analiza primero la relación de la preferencia por programas políticos con la edad o con el nivel de instrucción aparecen relaciones que parecen confirmarse en las tablas parciales al controlar la otra variable.

▪ Preferencia por programas musicales

La relación entre edad y escuchar programas musicales parece nula (r_t = 0.02): al parecer no hay relación. En este caso, sin embargo, la aparente no relación oculta algo más complejo, que aparece al introducir el NI como control: en personas más instruidas hay una relación negativa (r_t = -0.32) entre edad y gusto por programas musicales: los jóvenes tienden a oír más que los viejos. En los menos instruidos, la relación sigue moderada, pero positiva (r = 0.20): los viejos escuchan más.

Por su parte, entre NI y gusto por programas musicales hay una relación simple positiva clara (r_t = 0.33): los de NI alto los oyen más que los de NI bajo. Pero la relación cambia en dos sentidos distintos al controlar por edad: entre los jóvenes la relación casi desaparece (r_t = 0.08, no hay diferencia entre jóvenes más o menos instruidos); entre personas de mayor edad la relación positiva entre NI y preferencia por programas musicales se vuelve más fuerte (r_t = 0.54).

La lección que se debe sacar de lo anterior, de la mano de Lazarsfeld, es que toda correlación simple puede ser *espuria*, por lo que no se puede interpretar sin más como evidencia de causalidad.

Para concluir que hay relación causal entre dos variables se deben cumplir varias condiciones: *concomitancia* (las dos variables se correlacionan) y *antecedencia* (una debe preceder siempre a la otra).

Pero hay una condición más, que muchas veces se olvida: hay que verificar que el efecto no se deba a ninguna otra variable, hay que descartar cualquier otra posible causa. Lo que Lazarsfeld mostró fue que, si se controla la correlación simple haciendo intervenir una tercera variable, es posible detectar el carácter espurio de la correlación inicial, y descartar su interpretación en términos causales.

La reflexión debe seguir con la consideración de que, para poder hacer con solidez tal inferencia causal, sería necesario poner a prueba el posible carácter espurio de la relación original *controlando todas las posibles variables intervinientes, y no solo una o unas cuantas*. Y como el número de esas posibles es enorme, indefinido, tal cosa no parece sencilla. Se ha dicho ya que en principio implica utilizar un diseño experimental estricto, aunque esto tiene sus límites, que dificultan también llegar a conclusiones causales totalmente sólidas.

Hay que añadir que el razonamiento que subyace el modelo de elaboración de Lazarsfeld debe aplicarse en cualquier investigación que pretenda llegar a una interpretación causal, y no solo en los estudios que emplean técnicas estadísticas. Descartar la posible influencia de otras variables es también indispensable para poder concluir que hay relación causal en una investigación cualitativa.

Casi 75 años después de que Lazarsfeld propusiera el *modelo de elaboración* se han desarrollado técnicas estadísticas muy poderosas, con las cuales es posible analizar simultáneamente la relación que guardan entre sí muchas variables, al tiempo que se controla la incidencia de todas ellas.

Las computadoras actuales, por su parte, permiten hacer en segundos lo que llevaría meses del trabajo de decenas de personas, si los cálculos implicados se tuvieran que hacer de manera manual. Ese mismo poder de cómputo, sin embargo, permite que personas sin suficiente preparación presenten trabajos con impresionantes tablas que deberían sustentar complejos análisis, que muchas veces el autor ni siquiera entiende.

Por ello creí conveniente retomar la versión embrionaria del análisis multivariado propuesta por Lazarsfeld, cuya lógica resulta fácil de comprender, para facilitar la comprensión de los modelos actuales, algunos sumamente complejos, pero cuya lógica es básicamente la misma. De esta manera será posible usar inteligentemente los paquetes de software que ahora están al alcance de cualquier estudiante.

Análisis de varianza

Cuando la variable independiente de un estudio define solo dos grupos de sujetos (como hombres y mujeres) la probabilidad de que una diferencia entre las medias respectivas se deba al azar se puede analizar con el estadístico t y la distribución de probabilidad de Student.

Pero si los grupos que define la variable independiente son más de dos, es necesario acudir a otra herramienta, como el Análisis Ordinario de la Varianza (ANOVA), con la prueba F y la correspondiente distribución de probabilidad, que George Snedecor propuso en honor de Sir Ronald Fischer, que la desarrolló en la década de 1920, con el nombre de *relación entre varianzas*.

Este nombre revela el sentido de la herramienta, que compara la varianza de los valores de una variable dependiente dentro de (*within*) un grupo de sujetos definido por el valor de la variable independiente o bien entre (*between*) todos los sujetos de la muestra. Conviene destacar que Fisher desarrolló el ANOVA para analizar datos de en experimentos estrictos, con asignación aleatoria de sujetos a los grupos.

La lógica que distingue *varianza dentro de cada grupo* (intra, *within*) y *varianza entre grupos* (*between*) es la del modelo de elaboración de Lazarsfeld para controlar una tercera variable. Si hay un grupo de alumnos de sexo masculino y otro de sexo femenino, la varianza entre ambos puede atribuirse al sexo, pero la varianza dentro de cada uno de los dos grupos solo podrá deberse a otros factores que de momento ignoramos, pero no al sexo, que no varía en cada grupo, es constante, pues cada uno está formado solo por hombres o mujeres. La lógica del análisis ordinario de la varianza es esa, con dos o más grupos definidos por la variable independiente.

Para usar la distribución F para estimar la probabilidad de que las diferencias se deban al azar hay que calcular el valor F de los datos encontrados, con la fórmula F = varianza entre grupos / varianza intra-grupos. La fórmula es sencilla, pero los cálculos que implica son laboriosos, pues requieren medias de los cuadrados de las desviaciones de cada valor respecto a la media (varianza = MS, *mean square*, σ^2).

Si la varianza entre grupos es nula y los grupos que define la variable independiente tienen resultados idénticos, esa variable no incide en los resultados; toda la varianza se dará en el interior de uno o más de los grupos, se deberá a otros factores que no conocemos o a la imprecisión de las mediciones: será ruido o error. Como la fórmula es una relación (*ratio*) entre varianzas, F será menor mientras más pequeño sea el numerador; mientras más pequeña sea la varianza entre grupos, más probable será que la diferencia se deba al azar (que no se pueda rechazar la hipótesis nula). Y mientras mayor sea el

numerador (varianza entre grupos) mayor será F, menos probable que las diferencias entre las medias de los grupos se deban al azar, más probable que se deban a la variable que los define. (Salkind, 2007: 225-237)

Los paquetes de software realizan los cálculos en fracciones de segundo, y estiman la probabilidad exacta de que el resultado se deba al azar, con cierta probabilidad, sin necesidad de utilizar las tablas respectivas, sin olvidar que siempre se deberán precisar los grados de libertad de que se trate. En este caso se deben distinguir:

- Los del numerador de la fórmula, o sea el número de grupos formados con base en los valores de la variable independiente. En el análisis de varianza estos grupos pueden ser más de dos, pero no muchos más, y en las fórmulas se les suele denotar con la letra k.
- Los del denominador, o sea el número de sujetos de cada uno de los grupos definidos, o sea el tamaño de la muestra de cada grupo, que en las fórmulas se denota convencionalmente con la letra n.

Si se usa una tabla de la prueba F para el análisis de varianza, además de precisar la probabilidad que se define como aceptable (0.1, 0.05, 0.01) se debe buscar la intersección de los grados de libertad k del numerador (usualmente de 3 a 6) y los grados de libertad n del denominador, según el tamaño de cada grupo o del total.

Análisis de regresión

El segundo grupo de técnicas para analizar la relación entre más de dos variables es el análisis de regresión, basado en medidas de asociación o correlación.

El análisis de varianza y el de regresión se desarrollaron en ámbitos de investigación diferentes: el de varianza en psicología, y la regresión en economía y sociología, por lo que a veces se entienden como si fueran técnicas opuestas, cuando en realidad son muy semejantes. (Vogt, 2007: 152-153; Licht, 1997: 19-20)

El sentido usual del término regresión no ayuda a comprender en qué consiste esta técnica de análisis, que está en la base de gran parte de las herramientas más utilizadas en la investigación social y educativa. El término fue propuesto por Sir Francis Galton, cuando encontró que la estatura de un niño dependía en parte de la de su padre, pero no perfectamente, ya que los hijos de padres particularmente altos tendían a tener una estatura un poco menor, y los hijos de padres particularmente bajos, una estatura ligeramente mayor. La expresión *regresión a la media* designa este fenómeno, que se

debe a que la estatura de una persona no depende solo de una variable, la estatura del padre, sino también de otras, como la estatura de la madre, la de los abuelos y muchas más, que es improbable que incidan todas en el mismo sentido para dar un resultado excepcional. (Vogt, 2007: 145)

La técnica llamada *regresión*, que se podría denominar mejor como *predicción* o, con algunas condiciones, *explicación*, tiene como propósito inferir el valor de una variable con base en la información que se tiene de otra relacionada (regresión simple) o de otras variables (regresión múltiple).

La expresión en inglés, que traducida literalmente suena rara, "regresamos los puntajes obtenidos en 10° grado sobre el tamaño del grupo y los puntajes obtenidos en 8° grado" (*we regressed tenth grade scores on class size and on eight grade scores*) quiere decir "intentamos explicar los puntajes obtenidos en 10° grado a partir del tamaño del grupo y los puntajes obtenidos en 8° grado". (*cfr*. Vogt, 2007: 146)

El que el análisis de regresión se base en las medidas de correlación se entiende si recordamos que la asociación se puede visualizar graficando los valores de las dos variables en cuestión en un plano cartesiano, de manera que los de la variable que se considera independiente (que se suele designar con la letra x) se representan en el eje horizontal (en abscisa), y los valores de la variable considera dependiente (y) en el eje vertical (ordenada).

Como se ha dicho, los puntos que representan cada par de valores forman una *nube* que puede ser dispersa, aproximadamente rectilínea, o curva, y se puede estimar la recta que mejor se ajuste a una nube de puntos: la recta de regresión. Para ello el método usual, el de *mínimos cuadrados* (*Ordinary Least Square,* OLS), consiste en definir una recta que pase entre la nube de puntos de forma tal que la suma de las distancias verticales de cada punto a la recta, elevadas al cuadrado, sea más pequeña que para cualquier otra recta. La distancia a la recta de cada punto se elevar al cuadrado porque si no la suma tendería a anularse, pues alrededor de la mitad de los valores serían positivos, y negativos el resto, ya que los puntos se distribuirían más o menos por igual por arriba y por debajo de la recta en cuestión.

Se indicó antes que, si la nube de puntos sigue aproximadamente una línea recta, eso quiere decir que hay correlación fuerte entre las variables de que se trate. Hay que añadir ahora que una correlación fuerte puede ser positiva o negativa:

- Positiva: los valores de una variable se asocian en forma regular con valores *similares* en la otra: altos con altos y bajos con bajos.

- Negativa: los valores de una variable se asocian en forma regular con valores *opuestos* en la otra: altos con bajos y bajos con altos.

En el primer caso la recta de regresión comenzará cerca del ángulo que forman el eje de las x y el de las y, e irá *subiendo*, a medida que se desplaza a la derecha; en el segundo la recta comenzará en un punto cercano al ángulo superior izquierdo de la gráfica, e irá *bajando*, cada vez más cerca del eje de las x. Mientras más fuerte sea la pendiente de la recta (positiva o negativa) indicará una correlación mayor. Y si la nube de puntos es muy dispersa, la recta de regresión tenderá a pasar entre ella en sentido aproximadamente horizontal, sin acercase ni alejarse al eje de las x, lo que será indicio de una correlación baja, o nula en el caso extremo.

Esta forma aproximada de estimar la fuerza de una correlación es menos precisa que la forma cuantitativa que ofrecen las medidas propuestas por Pearson, a partir de la media, la varianza y su raíz cuadrada, la desviación estándar:

- *Covarianza*: el cociente que resulta de dividir los productos de las diferencias de los valores de las dos variables en cuestión respecto a la respectiva media: $cov_{xy} = \Sigma (x_i - x^j) (y_i - y^j) / N$.
- *Coeficiente de correlación*: que no depende, como la covarianza, de la escala en que están medidas las variables y toma valores entre 0 y 1, lo que se consigue *estandarizando* la covarianza, al dividirla entre el producto de las desviaciones estándar de las dos variables $Rxy = cov_{xy} / \sigma_x \times \sigma_y$.

Tanto la covarianza (en las unidades originales), como el coeficiente r (en unidades estándar z) representan qué tanto aumenta (o disminuye) la variable y por cada unidad que aumente la variable x. En otras palabras, representan la pendiente (el ángulo) de la recta de regresión respecto al eje de las x.

Por otra parte, el punto en que la recta de regresión cruza el eje de las y cuando la variable x es igual a cero (o sea en el origen del eje de las x), es otro valor importante en el análisis de regresión: el intercepto o constante.

El lenguaje matemático, que resulta misterioso e intimidatorio a quien lo encuentra por primera vez, es una manera económica y precisa de expresar ideas que, formuladas verbalmente, pueden ser más transparentes, pero también más largas y, posiblemente, imprecisas. Por ello el uso de fórmulas como las del análisis de regresión se vuelve indispensable. Para entender la lógica de dichas fórmulas ayudará un ejemplo. (Vogt, 2007: 149)

Imaginemos un proceso de selección de alumnos de nuevo ingreso a una carrera, a partir de información de una sola variable: sus calificaciones en bachillerato. Una regresión simple permitirá predecir (*regress*) las calificaciones que un sujeto podrá obtener en la universidad (variable dependiente, y), a partir de la independiente, x.

Veamos cómo pasar de la descripción verbal a la expresión matemática, con una ecuación de regresión:

1. Calificación predicha en carrera = promedio general de los alumnos en el bachillerato + calificación del sujeto en bachillerato.
2. Variable dependiente predicha = valor constante (intercepto) + variable independiente.
3. $y = a + bx$

La letra y es la variable de cada aspirante que se quiere predecir (dependiente), y la x la que conocemos, la variable independiente, la calificación de cada uno en el bachillerato. El supuesto es que mientras más altas sean las calificaciones de un aspirante en el bachillerato, más probable será que en la universidad también saque calificaciones más altas. Y para predecir esas calificaciones futuras:

- Se toma el promedio de calificaciones de todos los aspirantes en bachillerato: la letra a, el intercepto, el punto en el que la recta cruza el eje de las y.
- Se suma el producto de su calificación en bachillerato (la letra x) por el coeficiente de regresión b que indica qué tanto aumentará y por cada unidad en que aumente x.

Cuando, a diferencia del ejemplo anterior (regresión simple), se tiene información de dos o más variables independientes, cada una de las cuales aporta algo para predecir mejor la dependiente, se trata de una regresión múltiple. El ejemplo de Vogt (2007: 149. Figura 9.1), de la selección de alumnos de nuevo ingreso a una carrera universitaria, cuya versión simplificada se vio antes, maneja de hecho dos variables independientes: las calificaciones de los aspirantes en el bachillerato, y los puntajes que obtuvieron en una prueba de admisión.

En este caso la regresión permitirá predecir la calificación que obtendrá cada alumno en la universidad (la variable dependiente, y), a partir de las variables independientes,

x_1 y x_2. Veamos cómo pasar de la descripción verbal a la expresión matemática, con una ecuación de regresión:

1. Calificación predicha en carrera = promedio de calificaciones en bachillerato + calificación del sujeto en bachillerato + puntaje en la prueba de admisión.
2. Variable dependiente predicha = valor constante (intercepto) + variable independiente 1 + variable independiente 2
3. $y' = a + b_1 x_1 + b_2 x_2$
4. $y = a + b_1 x1 + b_2 x_2 + e$

En el paso 3, el apóstrofe que acompaña la letra y indica que el valor de la variable dependiente es una estimación; en el paso 4 se elimina el apóstrofe y se añade la letra e que significa error. En ambos casos se destaca que la estimación no puede ser exacta porque siempre hay fuentes de error, debido a que las mediciones de las variables consideradas no pueden ser perfectas, y a que puede haber muchas otras variables independientes no consideradas que también influyan en la dependiente.

La letra a (intercepto o constante) es el valor que toma la variable dependiente si todas las independientes tienen un valor de cero. Los coeficientes de regresión b que acompañan a cada variable independiente representan qué tanto aumenta (o disminuye) la dependiente por cada unidad en que la independiente aumenta (o se reduce). Si hay más de una variable independiente, esa proporción de incremento o disminución de la variable dependiente relacionada con cada independiente se estima descontando estadísticamente (controlando) el efecto de las demás, con la lógica de control de variables del modelo de elaboración de Lazarsfeld. Los coeficientes de regresión b son *coeficientes parciales*.

Expresados con letras b son coeficientes expresados en las unidades originales de la variable de que se trate sin estandarizar, como la covarianza, y no se pueden compara entre sí cuando las variables están medidas en unidades diferentes, como pesos, años, metros o puntos de una escala convencional. Para compararlos hay que expresarlos en desviaciones estándar arriba o abajo de la media (*unidades z*); los *coeficientes parciales estandarizados* son coeficientes r, y se designan con la letra griega beta (β), por lo que podemos añadir un quinto paso:

5. $y = a + \beta_1 x_1 + \beta_2 x_2 + e$

Los valores de las variables independientes (x) de la ecuación son parte de la información disponible para el investigador; los valores de a y b deben estimarse, a partir de la información disponible, con cálculos similares a los explicados para los coeficientes de correlación, que hechos a mano llevarían mucho tiempo, y con las computadoras actuales implican segundos. (Vogt, 2007)

El análisis de regresión aporta elementos para responder dos preguntas:

- ¿Qué proporción de la variación de la variable dependiente es predicha o explicada por las variables independientes consideradas en conjunto? Esta es la información que da el *coeficiente de determinación múltiple* R^2, que es el cuadrado del *coeficiente de correlación múltiple R*, con el que tiene una relación similar a la de r^2 con r.
- ¿Qué tanto predice (o, con matices, explica) la variable dependiente *cada una* de las independientes, descontando el efecto de todas las demás, o sea *controlándolas* estadísticamente, en el del modelo de Lazarsfeld? Esto es lo que dice cada uno de los coeficientes parciales estandarizados de regresión.

También en este caso es posible estimar la probabilidad de que los valores que se encuentran gracias a un análisis de regresión se hayan debido al azar.

La Tabla 4.28 sintetiza resultados de un análisis de regresión múltiple para predecir-explicar el éxito en la escuela de un grupo de adolescentes (medido con el promedio de calificaciones), con base en las prácticas de los padres, controlando seis posibles variables: promedio de calificaciones de los chicos; resultados en una prueba de rendimiento, sexo, edad, estatus socioeconómico y estructura familiar. El signo (positivo o negativo) y la significatividad estadística (desde menor a 0.001 hasta mayor a .10, no significativa) de todos los coeficientes parciales de regresión son idénticos en las dos columnas, que los presentan en forma no estandarizada y estandarizada. Los valores, en cambio, son diferentes.

TABLA 4.28. PREDICCIÓN DEL ÉXITO ACADÉMICO (PROMEDIO EN 1986)

Variables predictoras	Pesos (Coef. parciales de regresión)	
	No estandarizados	Estandarizados
Prácticas autoritarias de los padres		
Aceptación	0.039*	0.127*
Apoyo a la autonomía	0.048**	0.148**

Control de conductas	0.047**	0.142**
Otras posibles variables		
Promedio de calificaciones 1985	0.367****	0.363****
Prueba de rendimiento CAT 1985	0.011***	0.300***
Sexo	-0.007	-0.004
Edad	-0.002	-0.041
Estatus socioeconómico 1	0.072	0.034
Estatus socioeconómico 2	0.221	0.118
Estructura familiar 1	0.159	0.071
Estructura familiar 2	-0.268	-0.107
Intercepto = 1.494 R = 0.787* R^2= 0.619 *p < .10; ** p < .05; *** p < .01; **** p < .001 El cuadro incluye dos variables de estatus socioeconómico y de estructura familiar porque ambas se trataron en el estudio como nominales con tres categorías, por lo que en cada caso se incluyeron en el análisis dos variables ficticias (dummy).		

FUENTE: LICHT, 1997: 35. TABLA 1.

Con los coeficientes de cada variable en unidades originales (no estandarizados), el que uno sea mucho mayor que otro no dice gran cosa: tanto el coeficiente del promedio de calificaciones como el de la prueba de rendimiento son significativos (el primero con $p < 0.001$ y el segundo $p < 0.01$), aunque el valor del primero es 0.367 y el del segundo solo 0.011. En su forma estandarizada la significatividad es la misma, pero el valor del segundo (0.300) es similar al del primero (0.363). Los coeficientes no estandarizados no se pueden comparar entre sí, y solo por el valor p sabemos que la variable dependiente es predicha algo por los tipos de prácticas de los padres; algo más por el resultado en la prueba de rendimiento y, sobre todo, el promedio de calificaciones; y prácticamente nada por las otras cuatro variables.

Con coeficientes estandarizados se confirma que los cambios en cuatro variables independientes afectan poco a la dependiente (en tres casos en sentido negativo); que un cambio en una desviación estándar de las tres variables de prácticas de los padres implica un cambio de ~.13, .15 y .14 σ en el promedio de calificaciones de los chicos; y que un cambio de una σ del promedio de calificaciones previo y del resultado en la prueba predicen cambios de .36 y .30 σ en el promedio.

Al tratarse de coeficientes parciales, en todos los casos se trata del impacto de una variable controlando el de todas las demás. Con ello tenemos la respuesta a la segunda pregunta antes planteada. Con respecto a la primera, el coeficiente de determinación

múltiple R^2 nos dice que, en conjunto, las variables incluidas en el modelo dan cuenta de casi el 62% de la varianza de la variable dependiente (un valor alto en ciencias sociales), y que hay más de 90% de probabilidades de que eso no se deba al azar ($r = 0.787$, $p < .10$). (Licht, 1997)

Lo dicho hasta ahora se refiere a situaciones en que se busca estudiar la relación de una o más variables independientes con una dependiente. En la terminología más aceptada actual en esos casos no se trata de análisis multivariado, ya que esta expresión se reserva a estudios que exploran simultáneamente la relación entre múltiples variables independientes y múltiples variables independientes.

De manera análoga al ANOVA, el modelo de regresión ha tenido varios desarrollos, que se presentan en el siguiente punto de este capítulo, sobre técnicas avanzadas de análisis.

Esas variantes del modelo de regresión se desarrollaron para enfrentar situaciones en que no se cumple uno o más de los requisitos del modelo básico, como si la variable dependiente no es continua sino categórica, si los factores que se quiere estudiar se agrupan formando subconjuntos relativamente homogéneos, al interior de cada uno de los cuales no se cumple el requisito de independencia, o si se quiere reducir un gran número de variables que no son realmente diferentes, sino más bien aspectos (indicadores) de constructos (dimensiones) subyacentes, y se busca además identificar las interacciones más o menos complejas entre esas dimensiones o factores.

También hay que tener en cuenta situaciones como las siguientes:

- Con variables independientes categóricas. Con variables en sí dicotómicas, (como sexo masculino o femenino, si no se consideran otras posibilidades), o artificialmente dicotomizadas (como personas casadas o no, en cuyo caso el no incluye a personas solteras, viudas, divorciadas, y cualquier otra posibilidad) se usan variables ficticias (dummy) codificadas de modo que una categoría tenga el valor 1 y la otra 0. Con variables con más de dos categorías se construye una variable ficticia para cada categoría.
- Con variables independientes altamente correlacionadas entre sí, lo que se conoce como *multicolinealidad*. El modelo de regresión supone que las variables independientes no están correlacionadas entre sí, y cuando lo están en un grado importante, los coeficientes de regresión no son suficientes para detectar correctamente la influencia particular de cada variable. Para detectar lo anterior los paquetes estadísticos permiten hacer diagnósticos con medidas de *tolerancia*

(la proporción de la varianza de una variable independiente que no es explicada por otras de las variables independientes consideradas), o del *factor de inflación de la varianza.*

- Cuando hay datos faltantes (*missing*) de alguna de las variables. Cuando tal situación se presenta en un número importante de casos, el resultado al que se llega aplicando el modelo puede distar bastante del que se tendría si se contara con información completa. Las alternativas ante tal situación incluyen eliminar los casos de información incompleta, o sustituir los valores faltantes por la media de los datos con que se cuenta.

Técnicas avanzadas

En este último apartado del Capítulo 4, como se ha anunciado ya varias veces, se presentan algunas de las técnicas más avanzadas que se usan actualmente para analizar resultados de investigaciones sociales y educativas.

En los dos apartados anteriores el contenido se presentó con bastante detalle, ya que con ellos se pretende que los lectores alcancen una buena comprensión de las nociones implicadas, que les permita luego hacer interpretaciones correctas de los resultados de sus estudios, evitando los errores que puede cometer quien aplica las técnicas en cuestión con apoyo de una computadora y un paquete de software, sin entender realmente lo que hace. La experiencia me ha mostrado que esto no es tan raro como se podría pensar, lamentablemente.

Señalo nuevamente que el contenido de este capítulo no puede sustituir a un buen curso de estadística, ni a un buen texto especializado, pero que espero contribuir a la comprensión conceptual que supone la utilización de una técnica.

En contraste con lo anterior, las técnicas de que trata este apartado se presentan en forma muy escueta, que obviamente no basta para que los lectores las dominen, y tampoco siquiera que alcancen una buena comprensión conceptual de ellas. Esto se debe a dos razones: una es la complejidad misma de las técnicas de que se trata, que exigiría explicaciones mucho más amplias, por alguien con excelente dominio de esos contenidos, que yo, desde luego, no tengo.

La otra razón es que dominar esas técnicas rebasaría con mucho lo que, a mi juicio, puede esperarse de una buena maestría e incluso un buen doctorado que busque formar investigadores educativos.

En mi opinión el contenido de esta obra, sin este último apartado, es suficientemente amplio para definir los contenidos metodológicos del currículo de un posgrado.

Por ello las breves descripciones que siguen de técnicas avanzadas de análisis de información cuantitativa solo buscan dar una idea de ellas, que permita al lector saber para qué sirven. Será el ejercicio profesional de los egresados de posgrados de investigación lo que haga necesario que amplíen su formación en la dirección y en la medida en que lo requieran los retos profesionales que deban enfrentar.

La noción de formación continua alude, precisamente, a la necesidad de que todo profesional se mantenga actualizado, complementando el conocimiento adquirido durante su formación de pregrado y posgrado con lo que no haya cubierto, y estando al tanto de los avances que surjan en su campo. Las referencias que se mencionan en relación con cada técnica que se presenta en seguida son orientaciones para iniciar esas búsquedas.

El contenido del apartado se organiza presentando primero técnicas avanzadas que tienen que ver con la medición y su calidad; en seguida se presentan variantes de las técnicas de regresión, incluyendo las técnicas estrictamente multivariadas, que hoy se entienden como aquellas en que, además de que hay más de una variable independiente, hay también más de una dependiente; un tercer grupo es el formado por técnicas apropiadas en particular para algunos diseños de investigación.

De medición

Teoría o Modelos de Respuesta al Ítem (TRI)

Con inicio hacia 1900, a mediados del siglo XX se consideró madura la Teoría Clásica de las Pruebas o Test (TCT). Luego se desarrollaron los modelos, o Teorías, de Respuesta al Ítem (TRI), para construir pruebas y otros instrumentos. La TCT y los modelos de TRI parten de estimaciones de la *dificultad* de los ítems y de su poder de *discriminación*. Las herramientas de la TCT para estimar esos elementos son:

- Índice de dificultad: % de sujetos que responden correcta o incorrectamente.
- Discriminación, la correlación (punto-biserial) entre las respuestas de un sujeto a cada ítem y las que da al conjunto de la prueba; o bien la diferencia entre la proporción de los sujetos de mayor rendimiento que responden correctamente un reactivo, y los de menor rendimiento que lo consiguen.

Se puede mejorar la calidad de la información eliminando ítems de discriminación baja o suprimiendo ítems demasiado fáciles o difíciles. Pero como las medidas no son métricas, no se pueden utilizar modelos más poderosos, como los de regresión.

Los modelos de TRI utilizan métodos para variables categóricas, transformables en métricas. Utilizan *momios* (*odds*) y *razones de momios* (*odds ratios, cfr.* infra), con los valores expresados en el logaritmo natural (ln), con lo que se transforman en lineales. Las herramientas de los modelos de TRI equivalente a las de TCT son:

- Índice de dificultad: logaritmo natural del momio del sujeto, su lógito.
- Discriminación: el lógito del momio del ítem de que se trate.

Con estas medidas transformadas en lineales pueden usarse modelos de regresión, en particular *log-lineal*. Se puede estimar con más precisión que con la TCT una variable fundamental: el nivel de habilidad de los sujetos, por ejemplo, el grado en que muestran dominar el contenido de una prueba de aprendizaje (con la técnica de *Estimación por Máxima Verosimilitud, Maximum Likelihood Estimation*).

Las estadísticas de los modelos de TRI parten de ecuaciones básicas como estas:

- Modelos de un parámetro (o de Rasch): Lógito de un ítem = puntaje de habilidad de los sujetos menos nivel de dificultad.
- Modelo de dos parámetros: Lógito de un ítem = discriminación del ítem multiplicada por el puntaje de habilidad menos el nivel de dificultad.

Los modelos TRI permiten elaborar instrumentos que dan información precisa sobre la habilidad de los sujetos con menos ítems de dificultad cercana a la habilidad de cada sujeto (pruebas adaptativas por computadora). Se pueden hacer variantes equivalentes, aunque no tengan los mismos ítems (equiparadas), fortaleciendo la solidez de comparaciones en el tiempo, entre otras posibilidades.

Para ampliar los conocimientos sobre el tema puede verse Vogt, 2007; Downing y Haladyna, 2006; Muñiz, 1997; Wilson, 2005.

Teoría de generalizabilidad (TG)

Con los modelos de respuesta al ítem, la TG es un avance importante respecto a las ideas sobre medición de la Teoría Clásica. Para esta el puntaje observado, el que resulta de la aplicación de un instrumento, es la suma de dos componentes, *puntaje verdadero* + *error*: el nivel real del sujeto en la habilidad o conocimiento que se mide, y el error que inevitablemente acompaña a su medición.

En la TCT, pues, el error se concibe como unidimensional. Desde la década de 1930 algunos estudiosos habían sugerido que el error podía tener varias dimensiones, pero fue hasta 1963 cuando Lee Cronbach y colaboradores propusieron la idea de la *generalizabilidad*, que plantearon formalmente en Cronbach *et al.*, 1972.

En vez del término único de error, la TG especifica un *universo de observaciones admisibles* que especifica varias fuentes de error o *facetas*, como las diversas formas de un instrumento, las personas que aplican un protocolo de observación, o las ocasiones o lugares en que se aplica un instrumento o un protocolo.

Cada una de esas formas, personas, ocasiones o lugares es una *condición* de una faceta. Una observación se define por una condición de cada faceta considerada. En lugar de un solo puntaje verdadero, la TG define un puntaje del universo de observaciones.

Aplicando técnicas del ANOVA, la TG permite hacer dos tipos de estudio, desde los más sencillos, con una sola faceta, hasta los más complejos, anidados y con múltiples facetas.

En lugar de los coeficientes de confiabilidad tradicionales que se desarrollaron en el marco de la TCT (alfa de Cronbach y otros), en los estudios de TG se usan otros coeficientes, uno de generalizabilidad ($E\rho^2$), análogo al de la TCT, y un *índice de confianza* Φ (*dependability*). Los dos tipos de estudio son:

- G, de generalizabilidad, para estudiar la confiabilidad de las observaciones con el coeficiente Eρ2 o el índice Φ.
- D, de decisión, para optimizar la calidad de las observaciones, al analizar qué tanto mejora la confiabilidad al reducir el error de cada faceta, lo que siempre implica costos adicionales: más formas, más observadores, más ocasiones, entre otras posibilidades.

Quien quiera saber más del tema puede consultar Brennan, 1983; Brennan, 2001; Cronbach *et al.*, 1972; Haertel, 2006; Hill *et al.*, 2012; Shavelson y Webb, 1991.

Análisis de patrones de respuesta

En países latinoamericanos como México, los resultados de PISA han mostrado una y otra vez una relación inversa entre los resultados en las pruebas y en las escalas de actitudes: los chicos con resultados más bajos en pruebas dicen tener actitudes más favorables, en relación con las mismas competencias, que los estudiantes de los países de resultados cognitivos altos. Esto ha llevado a hipotetizar que algunos rasgos de la cultura de unos países podrían llevar a responder positivamente las preguntas sobre

actitudes, en grado mayor que en otros medios culturales. Estudios de mercadotecnia y sobre opiniones políticas han identificado patrones consistentes que en ocasiones muestran las respuestas a preguntas de escalas tipo Likert.

> [...] hay evidencias que sugieren que diferencias en estudios internacionales o interculturales están contaminadas por características de los instrumentos de medición (artifacts of measurement) [...] parte de esos estudios se refiere a las diferencias interculturales asociadas al uso de escalas Likert, o de ítems categóricos individuales de ese tipo de escalas [...]. (Buckley, 2009: 4)

Se distinguen *estilos de respuesta, o diferencias en el uso de escalas de respuesta, que pueden llevar a sesgos en la investigación intercultural sobre actitudes:*

- Estilo de respuesta afirmativo (acquiescence response style, ARS), o sesgo positivo: estar de acuerdo con un ítem, aunque no sea la actitud real.
- Estilo de respuesta negativo (disacquiescence response style, DARS): tendencia a estar en desacuerdo con un ítem, aunque no sea la actitud real.
- Estilo de respuesta extremo (extreme response style ERS): tendencia a elegir opciones de respuesta en extremos de la escala de un ítem (*v. gr.* totalmente satisfecho o insatisfecho), de nuevo sea cual sea la actitud subyacente real.
- Estilo de respuesta azaroso (noncontingent responding, NCR): término que describe un estilo de respuesta aleatorio o descuidado. (Buckley, 2009: 4-5)

Buckley investigó la posible presencia en los resultados de la aplicación 2006 de PISA de esos tipos de sesgos debidos a estilos de respuesta distintos entre países.

- Calculó índices de sesgo afirmativo, negativo, respuesta extrema y azarosa.
- Estimó el valor de cada índice en los países que participaron en PISA 2006.
- Aplicó un *modelo simple de estimación del efecto del estilo de respuesta* calculando valores de las respuestas a las escalas de actitud de PISA 2006 de cada país, después de corregir el efecto del estilo de respuesta respectivo.
- Hizo una nueva corrección con un acercamiento jerárquico bayesiano.

Otros trabajos estudian patrones característicos de las respuestas de unos sujetos a ciertos tipos de preguntas, en especial escalas de tipo Likert. Identificar esos patrones permite corregir sesgos en la medición asociados con rasgos culturales.

Más sobre el tema en Buckley, 2009.

▪ Enfoque *basado en argumentos*

Verificar unidimensionalidad y confiabilidad de la información que da un instrumento es necesario, pero no suficiente. En la terminología tradicional hay que verificar la validez de constructo, que hoy se aborda estudiando la fuente de evidencias sobre los procesos de respuesta, para lo que se suelen usar entrevistas cognitivas que permiten captar lo que está pensando una persona al responder una pregunta o reaccionar a un estímulo, lo que sin tales entrevistas solo inferimos. Es posible que dos personas de contexto cultural o lingüístico diferente entiendan en forma también distinta una misma pregunta, lo que lleva ala llamada validez cultural.

Por otra parte, los resultados de la aplicación de una prueba se pueden usar para definir qué aspirantes entrarán a una carrera y quiénes quedarán fuera, quienes aprobarán un curso o lo reprobarán, o bien qué docentes o qué escuelas merecerán estímulos por conseguir que sus estudiantes tengan mejores resultados. En este caso se trata de la validez de consecuencias, y en cada caso será necesario distinto tipo de evidencias para defender que el uso de los resultados es correcto.

Con lo anterior se relaciona el *enfoque a la validez basado en argumentos* propuesto por Kane, que parte de la idea de que la validez es un concepto unitario, y que para analizarlo es indispensable partir siempre del propósito que se quiere alcanzar con la información que se obtenga al aplicar un instrumento, ya que la validez se refiere siempre a lo que se hace con esa información, a las decisiones que sustenta.

Este enfoque implica sostener las interpretaciones de los resultados obtenidos con un instrumento, y los usos que se piensa hacer de ellos, mediante un conjunto de argumentos, interpretativos y de validez. Esto implica que ninguna medición es perfecta, y que varias en conjunto podrán ser mejores, pues cada una ofrece una imagen parcial de un aspecto del constructo global que se quiere estudiar. Por ello se requieren indicadores múltiples, y el análisis del patrón de las correlaciones que se observen entre dichos indicadores. La imagen siguiente ilustra la idea.

GRÁFICA 4.19. PATRÓN DE CORRELACIONES ENTRE INDICADORES

FUENTE: MARTÍNEZ, J. F., 2015. PARA SABER MÁS DEL TEMA PUEDE VERSE KANE, 2006.

Análisis de varianza, de regresión y factorial

Los análisis de varianza y regresión son similares, variantes del modelo matemático *lineal general*, lo que se aprecia comparando las fórmulas básicas de uno y otro. Con la abreviatura MS (*mean square*) para designar la varianza, las fórmulas son:

Análisis de varianza: MS total = MS entre grupos, between + MS intra-grupos, within
Análisis de regresión: MS total = MS de regresión + MS residual

En ambos la variación total es la suma de la variación predicha por, o asociada con, la variable independiente, y la variación de error (intra-grupos o residual).

El análisis de varianza implica variables independientes categóricas y dependientes métricas; en el de regresión se puede trabajar con variables independientes y dependientes de cualquier nivel de medición, y con un número considerable de variables, lo que no es fácil con el análisis de varianza.

El análisis de varianza más sencillo considera una variable independiente y suele designarse con la expresión "de una vía" (*one-way*), pero se puede trabajar con dos, tres o más variables independientes (*two-way, three-way*).

Como se usan variables independientes categóricas, las categorías de cada una se combinan *factorialmente* con las de las otras. Si una independiente tiene dos valores (*v.gr.* sexo masculino y femenino) y la otra tres (*v.gr.* tres tratamientos distintos en un experimento), se tratará de un diseño *two-way* de 2 x 3, y así sucesivamente, con lo que los cálculos se vuelven muy complicados.

Las versiones más simples de estos análisis son insuficientes para dar cuenta de la complejidad de muchas situaciones reales. En el campo educativo se puede dar por descontado que en todo fenómeno relevante incidirá una multiplicidad de factores. Por ello desde principios del siglo XX se entendía la necesidad de tener mediciones precisas de los numerosos aspectos de cualquier fenómeno educativo importante, y análisis robustos de las relaciones entre dichos aspectos. Esto se hizo posible con el avance de herramientas de medición como las pruebas de rendimiento, y el de técnicas de análisis, como las primeras versiones de los análisis de varianza y de regresión. Con la difusión de las computadoras, se volvieron accesibles técnicas más complejas, varias propuestas teóricamente décadas antes.

Murnane y Willet (2011: 3-10) comentan el impacto que significó para la difusión de técnicas avanzadas en las investigaciones educativas el famoso Informe Coleman, y la polémica que siguió a su difusión en 1966. Se multiplicaron los esfuerzos por superar las limitaciones del informe para hacer atribuciones causales con mejores técnicas, diseños longitudinales, modelos lineales jerárquicos, entre otros.

En seguida se presentan técnicas avanzadas, en una forma que no será suficiente para aplicarlas, y ni siquiera para entender los resultados de estudios en los que se utilicen, ya que dominar cada técnica implicaría un curso especial. Con las páginas siguientes se pretende solo que los lectores vislumbren la vastedad del campo en la actualidad. Llegado el momento deberán buscar la formación adicional necesaria para usar bien una u otra de estas técnicas.

ANCOVA y MANOVA

Las aportaciones de Fisher se dieron en el marco de sus estudios sobre rendimiento de semillas. La fundamental consistió en la noción misma del diseño experimental moderno, con asignación aleatoria de sujetos a los grupos experimentales y al de control, que es fácil utilizar en el campo de la investigación agronómica. Vino luego el ANOVA, para analizar datos obtenidos en un contexto experimental.

Una primera variante, el Análisis de Covarianza (ANCOVA), fue propuesto también por Fisher, siempre en el contexto de datos experimentales, para incrementar el poder de la prueba de la variable independiente.

No pocos investigadores han usado posteriormente ANCOVA en el marco de diseños no experimentales, que comparan grupos naturales sin asignación aleatoria, buscando *controlar* una tercera variable, una *covariante*, con una lógica similar a la del modelo de elaboración de Lazarsfeld. Como en este caso, sin embargo, introducir una o más

covariantes no sustenta una inferencia causal, que solo es aceptable, con salvedades importantes, en el marco de diseños experimentales estrictos.

La extensión del modelo ANOVA cuando se pretende valorar la significatividad estadística del efecto de una o más variables independientes sobre un conjunto de dos o más variables dependientes, se conoce como *Análisis multivariado de la varianza* (MANOVA). (Weinfurt, 1997)

A partir de técnicas de ANOVA para ver si la diferencia entre medias de la variable en los grupos considerados se debe al azar, con la t de Student o la prueba F, en análisis MANOVA se usan otras estadísticas, para tener en cuenta que en lugar de medias de valores individuales se trabaja con vectores, lo que implica complejas operaciones con matrices, y nuevas estadísticas, como el equivalente de la R^2 que es η^2 (eta cuadrada), o la lambda de Wilkins (λ o Λ).

También deben hacerse análisis del tamaño del efecto (*effect size)* y del poder estadístico (*power análisis*). Otras variantes son el análisis MANOVA de medidas repetidas, el análisis de reducción por la introducción progresiva de variables en el modelo (*step-down*), o el análisis discriminante.

Para ampliar los conocimientos sobre el tema puede verse Holland y Rubin, 1986; Miller y Chapman, 2001; Owen y Froman, 1998; Weinfurt, 1997; Weinfurt, 2000.

Regresión logística

Si la variable dependiente es categórica, sus valores no se distribuyen normalmente y no se pueden calcular medias ni varianzas. Si la variable independiente también es categórica, se pueden hacer tablas de contingencia y usar una prueba de X^2. El análisis implica *momios* (*odds, O*) y *razones de momios* (*odds ratios, OR o theta Θ*):

- *Momios*. Un momio (O) es la *relación* entre la probabilidad de que cierta condición se presente, frente a la de que no se presente. La suma de las dos probabilidades será siempre igual a 1, y cualquiera de ellas será igual a la unidad menos la otra (O = p / 1-p).
- *Razones de momios* (OR). Son relaciones de segundo nivel, relaciones de relaciones: precisamente, la relación entre dos momios. OR o $\Theta = O_1 / O_2$.

Ejemplo: 1,000 docentes, 300 hombres y 700 mujeres; 600 trabajan en educación básica y 400 no, sino en media superior. La Tabla 4.29 presenta cada combinación del sexo de los docentes y el nivel en que trabajan, en cifras absolutas y con la probabilidad de que una de sexo masculino o femenino trabaje en uno u otro nivel.

TABLA 4.29. DOCENTES SEGÚN SEXO Y GRADOS EN QUE ENSEÑAN

	Ed. básica	Media sup.	Totales
Hombres	95 (.3167)	205 (.6833)	300
Mujeres	505 (.7214)	195 (.2786)	700
Totales	600	400	1,000

FUENTE: VOGT, 2007: 198. TABLA 11.11.

- Momio de que hombres enseñen en educación básica $O_1 = .4634$ (95/205).
- Momio de que mujeres enseñen en educación básica $O_2 = 2.590$ (505/195).
- Razón de momios hombres/mujeres $OR = O_1/O_2 = .1789$ (.4634/2.590)
- Razón de momios mujeres/hombres $OR = O_2/O_1 = 5.589$ (2.590/.4634)

Modelos de regresión log-lineal. A partir de momios y razones de momios, estos modelos sirven para analizar datos de tablas de contingencia de fenómenos con variables categóricas, tanto independientes como dependientes, pero empleando pruebas con el poder estadístico del análisis de regresión, posible gracias a la transformación de datos en lineales, con el logaritmo natural de los momios (ln de base e = 2.718281...), de donde el nombre de modelos *log-lineales*. El ln de un momio (*log odds*) se llama *lógito* (*logit*). Al usar el logaritmo de valores nominales u ordinales, distribuciones originalmente irregulares y sesgadas, toman rasgos de distribuciones de variables métricas (de intervalo): regulares y simétricas.

Los métodos más poderosos para datos con variables independientes categóricas y una variable dependiente también categórica, son el análisis discriminante, la regresión *probit*, y la *regresión logística*, si la variable dependiente es dicotómica, y si tiene más de dos categorías, la *regresión logística multinomia*l.

Para ampliar las ideas sobre el tema puede verse Vogt, 2007; Wright, 1997.

Modelos de Regresión Lineal Jerárquica (HLM)

Un supuesto de la regresión es que haya independencia entre los sujetos de los grupos a estudiar. Esto no siempre se cumple en el campo educativo, que tiene una *estructura anidada*, de varios *niveles*: estudiantes y docentes trabajan en aulas, que se agrupan en escuelas, y estas en zonas de un estado o país. En cada nivel de la estructura hay variables que no son independientes entre sí, sino que tienden a variar en forma similar, ya que el nivel se distingue por tener elementos comunes:

- Las escuelas, docentes y estudiantes de un país pueden plan y programas de estudio comunes y los mismos libros de texto.
- Escuelas, docentes y alumnos de una entidad subnacional pueden compartir políticas, como la forma de tratar a ciertas minorías étnicas.
- En cada escuela docentes y alumnos comparten un espacio con cierto clima y recursos, un mismo director(a) y ciertas formas de gestión.
- En cada aula los alumnos tienen un mismo docente, que maneja ciertas prácticas de enseñanza, cierto tipo de disciplina y gestión del grupo.

Incluso con un nivel, en todo estudio de varias escuelas, por ejemplo, la situación de alumnos y docentes de cada una no cumple la condición de independencia, pues comparten director, clima escolar, recursos, entre otras cosas.

En estos casos no es adecuado un modelo de regresión de un solo nivel, y por ello se desarrollaron los *Modelos Lineales Jerárquicos, o Multinivel*, en los que hay que añadir una nueva ecuación para incluir variables de un nivel (o plantel) diferente.

La diferencia puede apreciarse con el caso del Informe Coleman (1966), que estudió la desigualdad de los resultados escolares de niños de diversos grupos étnicos. Un hallazgo del famoso informe, según una lectura todavía prevaleciente en algunos medios, sería que la proporción de la varianza de los resultados asociada a factores del entorno familiar y social de los niños sería muy grande, del orden de 90%, en tanto que la proporción asociada a factores de la escuela rondaría solo 5%. Los estudios actuales con un Modelo Lineal Jerárquico de dos niveles llegan a proporciones diferentes, alrededor de 70% de la varianza de los resultados asociada a factores del entorno y 30% a factores de la escuela. En estudios que distinguen tres niveles (niños, aulas y escuelas) la proporción de la varianza de los resultados asociada a factores del entorno se reduce aún más, por ejemplo, alrededor de 60%, mientras el resto se asocia con la escuela y el aula. Esto cuestiona la extendida idea de que la escuela es incapaz de contrarrestar, al menos parcialmente, la influencia de los factores del entorno.

Para profundizar en el tema se puede consultar Vogt, 2007; Bryk y Raudenbush, 1992; Gelman y Hill, 2007.

Análisis de componentes principales y factorial

La complejidad de los fenómenos sociales y educativos hace que, en general, el estudio de cualquiera de ellos deba considerar numerosas variables, advirtiendo que las variables

no siempre son realmente diferentes: pueden ser manifestaciones particulares de un constructo subyacente, indicadores de una misma dimensión, o variables manifiestas (observadas) de una variable latente. Por ello han surgido técnicas que buscan reducir muchas variables que no sean realmente diferentes a un número menor de factores o componentes: el *Análisis Factorial Exploratori*o (AFE), la técnica similar del *Análisis de Componentes Principales* (ACP), o el acercamiento más complejo que es el *Análisis Factorial Confirmatorio* (AFC).

Un uso frecuente del Análisis Factorial se da en la construcción de instrumentos de obtención de información, como las escalas de actitud, que consisten en cierto número de ítems, que posiblemente se refieren a un número menor de dimensiones; si estas pueden identificarse, la extensión del instrumento puede reducirse, sin sacrificar la calidad de la información que produce, e incluso mejorándola.

El punto inicial del AF es el conjunto de correlaciones entre todas las variables de que se trate (plasmado en una matriz de correlaciones), de donde se parte para identificar grupos de variables que tengan alta correlación entre sí, y muy poca con las demás, lo que se interpreta en el sentido de que las variables de un grupo con esas características forman un solo *factor*.

A partir de la matriz de correlaciones se calculan valores característicos de cada variable (eigenvalores), que son una medida de la cantidad de la varianza de todas las variables de la que da cuenta un posible factor. Dividiendo cada eigenvalor entre el número de variables se obtiene el porcentaje de la varianza asociado a ese factor. Los factores con eigenvalores muy pequeños se descartan, de diversas formas.

Una vez definido un número de posibles factores, se calcula la correlación de cada variable con cada factor, su *carga factorial*. En principio las variables con carga factorial mayor a cierto valor (*v.gr.* 0.4) forman un factor, pero para definir mejor el conjunto de factores a retener se utilizan diversas formas de *rotación*, dado que los factores pueden estar relacionados o no entre sí. Los vectores de los factores identificados pueden tener una relación ortogonal (en ángulo recto) cuando no hay relación entre ellos, o bien oblicua (en ángulo más o menos agudo) cuando sí hay relación. Gracias a la rotación de factores, las correlaciones entre las variables de un factor aumentan, al tiempo que disminuyen las que hay entre factores distintos.

El AFE implica cálculos de álgebra lineal muy laboriosos que solo es práctico hacer con apoyo computacional, pero la interpretación de resultados implica un elemento no estadístico: para precisar la variable latente (o constructo subyacente) que define

un factor hay que acudir a la teoría, a lo que se sabe sobre el fenómeno de que se trate, y al contenido de las variables que forman el factor.

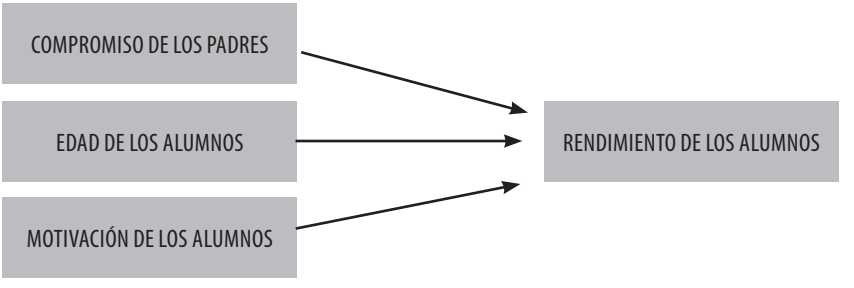
Más sobre el tema en Vogt, 2007; o en Bryant y Yarnold, 1997.

*Análisis de trayectorias (*Path Análisis*)*

Con una regresión se puede estudiar el grado en que una variable dependiente es predicha por un conjunto de variables independientes, sin analizar la manera en que las variables independientes interactúan entre sí: en muchos casos unas variables son independientes respecto a otras, que dependen de ella, pero a su vez pueden incidir en otras, y puede haber variables que influyen unas en otras. El *Análisis de trayectorias* es una extensión del modelo de regresión que propone una *estructura teórica* en la que se precisan las relaciones, simples o complejas, que se conjetura pueda haber entre las variables involucradas en un fenómeno.

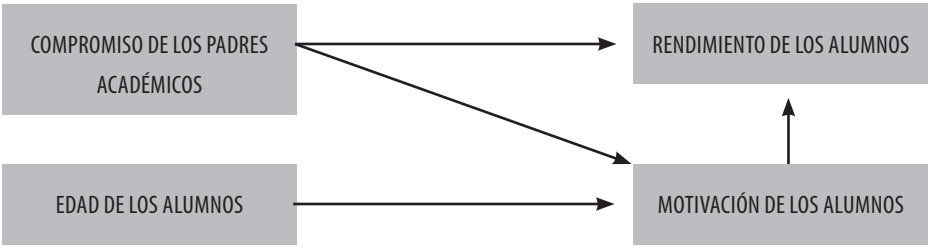
Ejemplo: dos estudios que busquen explicar la variable dependiente *rendimiento de unos alumnos*, con tres variables independientes: *compromiso de los padres, motivación de los alumnos, y edad de estos*. Un modelo de regresión plantea solo que las tres variables independientes *influyen de algún modo* en la dependiente, pero no propone una forma en que la influencia de cada variable independiente se combina con las de las otras para producir el efecto. El Análisis de trayectorias aporta *una estructura de relaciones* entre las variables independientes entre sí, y con la dependiente, que dé cuenta mejor del fenómeno a explicar.

GRÁFICA 4.20. MODELO DE REGRESIÓN



FUENTE: VOGT, 2007: 248. FIGURA 14.2.

GRÁFICA 4.21. MODELO DE ANÁLISIS DE TRAYECTORIAS



FUENTE: VOGT, 2007: 248. FIGURA 14.1.

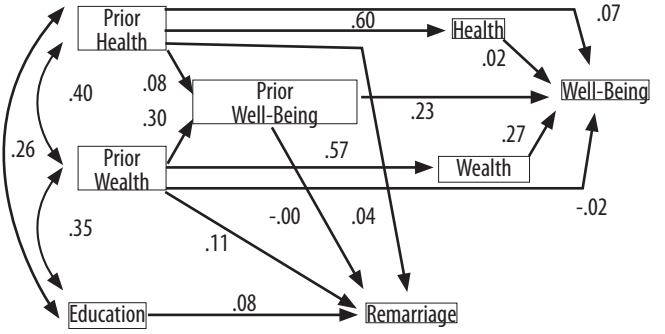
En el primer caso tenemos solo tres relaciones simples; en el segundo hay una estructura con las mismas variables, pero que propone una posible forma en que las variables independientes se combinen para incidir en la dependiente.

Los estudios con Análisis de Trayectorias suelen incluir más variables que, según su posición en la estructura propuesta hipotéticamente, pueden ser *exógenas* (si no dependen de ninguna otra), o *endógenas* (las demás, que dependen al menos de otra variable, y en muchos casos de dos o más). La estructura se puede representar con un conjunto de ecuaciones de regresión, pero es más fácil visualizarla con un *diagrama de trayectorias* que representa el flujo causal hipotetizado y muestra:

- A la izquierda las variables exógenas, unidas por curvas con puntas en los dos extremos, indicando la correlación que hay entre cada par de ellas.
- En la parte central variables endógenas, de izquierda a derecha, con líneas rectas con punta en el extremo que indica el flujo causal hipotetizado, y un coeficiente de trayectoria (*path coefficient*), el coeficiente de regresión que indica la importancia del efecto directo de una variable sobre la otra.
- A la derecha la(s) última(s) variable(s) endógena(s), que se quiere explicar con el modelo y recibe(n) la influencia directa e indirecta de las otras incluidas en él, pero ya no influye(n) en ninguna más.
- Flechas con punta dirigidas a variables endógenas, que no salen de ninguna otra e indican el efecto de variables no identificadas; sus coeficientes indican el componente error de la ecuación de regresión (*residual path coefficients*).

La Gráfica 4.22 es un diagrama de trayectorias, sin flechas ni coeficientes residuales:

GRÁFICA 4.22. DIAGRAMA DE TRAYECTORIAS



FUENTE: KLEM, 1997: 71. FIGURA 2.

A partir de un diagrama de este tipo puede definirse el conjunto de ecuaciones que representan matemáticamente el modelo, y calcular los efectos directos e indirectos de unas variables sobre otras, el efecto total de las variables consideradas sobre la(s) que se quiere explicar, así como la varianza no explicada, de error o residual de cada variable endógena y del conjunto.

Para saber más sobre el tema véase Klem, 1997.

Modelos de Ecuaciones Estructurales (SEM)

A partir de los principios del Análisis de Trayectorias, los Modelos de Ecuaciones Estructurales (*Structural Equation Modelling,* SEM) plantean estructuras hipotéticas de relaciones causales entre variables, incluyendo variables latentes, lo que hace necesario un modelo de medición que permita integrar datos de las variables observadas con las que se construye la información de cada variable latente.

En el *Path Analysis,* desarrollado hace casi 100 años, hay *una sola medición* de cada variable incluida, que se supone puede ser observada o medida directamente, por lo que no hace falta utilizar un modelo de medición.

En un modelo SEM, en cambio, y de acuerdo con la visión actual de la medición, se entiende que muchas variables no se pueden medir directamente, sino que para hacerlo se necesitan varios indicadores (por ejemplo, los ítems de una escala), lo cual hace necesario un modelo de medición, junto al componente estructural. Debe añadirse que, como este componente define una estructura teórica, hipotética, en la que se integran las variables latentes consideradas, el análisis factorial que se debe utilizar no es exploratorio, sino confirmatorio.

Las gráficas que se usan en los estudios con Modelos de Ecuaciones Estructurales son similares a las de estudios con Análisis de Trayectorias, pero más complejas, porque cada variable del modelo de trayectorias es acompañada por el modelo de medición correspondiente.

La estructura de las relaciones entre las variables de un modelo SEM, como en un Análisis de Trayectorias, permite ir más allá de la descripción y aproximarse a la explicación. Si la relación entre una variable independiente y una dependiente está mediada por una o más variables intermedias, hay cierta base para sostener que hay una relación causal, que esas variables intermedias especifican. Una relación entre dos variables sin mediaciones, en cambio, permanece inexplicada.

Para tener una idea más completa de estos modelos puede verse Vogt, 2007; Klem, 2000; o Kaplan, 2000.

Análisis de datos en diseños particulares

Las herramientas analíticas descritas en apartados anteriores pueden aplicarse a los datos de cualquier diseño de investigación, pero toman formas particulares en el caso de diseños que buscan dar sustento sólido a la atribución de causalidad.

La noción de causa no es fácil de definir. Puede sostenerse incluso que no se puede definir de la manera convencional, que implica identificar una categoría *superior* en la que estaría incluido el concepto en cuestión, o sea mediante una definición "por género y especie". Esto se debe a que la noción de causa es una de las categorías más básicas de nuestra forma de conocer la realidad. Así lo planteó hace más de dos siglos Immanuel Kant, que reelaboró las categorías aristotélicas y propuso que, además de las de espacio y tiempo, habría un pequeño conjunto de tales nociones que no pueden reducirse a otras más generales, entre las que se incluye la de causa-efecto o, en términos abstractos, de causalidad y dependencia.

Las condiciones que se deben cumplir para establecer la presencia de una relación causal entre dos fenómenos son: *concomitancia* (los fenómenos van siempre juntos: si se presenta uno, el otro también lo hace siempre); *precedencia* (el orden importa: un fenómeno antecede siempre al otro); y *¡causalidad*! (lo anterior no se debe a sincronización o a otro factor, sino que un fenómeno es *producido* por el otro). La circularidad de la tercera condición es evidente pero no hay manera de evitarla, ya que la noción de causa es una categoría básica del modo de conocer humano.

No dice otra cosa el que *correlación no necesariamente implica causalidad*; y, traduciendo a jerga metodológica los planteamientos hechos en clave filosófica, se dan tres

criterios para establecer casualidad: hay correlación entre dos variables; una precede a la otra; y no hay explicación alternativa. (Vogt, 2007: 51-53)

Observar empíricamente correlación entre dos variables quiere decir establecer concomitancia; la precedencia se puede observar añadiendo un sentido o dirección a la correlación, ya que el efecto no puede preceder a la causa. Pero además hay que verificar que se cumpla el tercer criterio, o sea descartar que el fenómeno a explicar se deba a otra u otras variables. Esto parece imposible, ya que descartar cualquier explicación alternativa implica *controlar todas las otras variables que podrían ser la verdadera causa* del fenómeno a explicar. Se pueden enumerar decenas de variables potencialmente relacionadas con el fenómeno a explicar, y se pueden descartar otras por considerar que no es verosímil que influyan, pero, además de que el sustento de la decisión de descartar ciertas variables puede ser endeble, sigue presente que puede haber otras variables en las que ni siquiera se ha pensado, y que por lo mismo no se han observado, y menos controlado.

Los experimentos con asignación aleatoria de sujetos resuelven este problema, que parece tan difícil, de una manera en principio sorprendentemente sencilla. La idea clave es que, si se asigna aleatoriamente un número suficiente de sujetos a dos grupos, y en uno de ellos se interviene modificando una variable, la diferencia que se pueda encontrar luego entre ambos grupos en cuanto a esa variable solo se podrán atribuir a la que se manipuló, y a ninguna otra, ya que los dos grupos serán equivalentes en todos los aspectos, conocidos o no, por la asignación aleatoria. En los incisos siguientes se presentan estrategias analíticas sobre estos temas.

Estudios experimentales y cuasiexperimentales

Por las razones mencionadas, el experimento estricto, con asignación aleatoria, fue considerado el diseño de investigación ideal, el único que permitía llegar en forma sólida a conclusiones causales, a explicar realmente un fenómeno. Después se advirtió que también este tipo de diseños tiene límites que, en la práctica, impiden llegar a conclusiones totalmente seguras. Campbell y Stanley (1963) propusieron una primera sistematización de las amenazas a las dimensiones de la validez de un experimento. Una versión más reciente de tales amenazas es la siguiente.

TABLA 4.30. AMENAZAS A CUATRO TIPOS DE VALIDEZ

Tipo de validez	Amenazas
De conclusiones estadísticas	Bajo poder estadístico; Incumplimiento de supuestos de las pruebas estadísticas; Búsqueda abusiva de relaciones significativas (*fishing*); Baja confiabilidad de medidas; Restricción de rango; No fidelidad de la implementación; Varianza extraña en los arreglos experimentales; Heterogeneidad de unidades; Mala estimación de tamaño de efecto.
Interna	Precedencia temporal ambigua; Selección; Historia; Maduración; Regresión; Deserción de sujetos; Efecto prueba; Instrumentación; Adición o interacción de los efectos de las amenazas anteriores.
De constructo	Explicación inadecuada o confusión de constructos; Sesgo por una sola operacionalización o método; Confusión constructos-niveles; Estructura factorial sensible a tratamiento; Autorreportes alterados por asignación a un grupo o por tratamiento; Sesgo por expectativas del experimentador; Novedad; Igualación-rivalidad compensatoria; Desaliento de participantes; Difusión del tratamiento.
Externa	Interacciones de la relación causal con unidades, con variantes del tratamiento, con los resultados, o con los arreglos experimentales; Dependencia del contexto de una mediación.

FUENTE: SHADISH, COOK Y CAMPBELL, 2002: 45, 55, 73 Y 87. TABLAS 2.2, 2.4, 3.1 Y 3.2.

Algunas técnicas de análisis estadístico enfrentan estas amenazas, pero además es fundamental tener ciertos cuidados en lo que se refiere a la forma de implementar los tratamientos y, en general, de realizar el experimento.

En medicina, cuando se experimenta con un nuevo fármaco para ver su efecto en los sujetos (ratones o humanos), es fácil asegurar que la sustancia y la dosis que se proporciona a cada sujeto es idéntica, pero tratándose de humanos es posible que su reacción se vea afectada por el hecho mismo de saber que se les ha dado cierto fármaco (sugestión), lo que se busca evitar con el uso de sustancias inocuas (*placebos*) que se proporcionan a los sujetos del grupo control, con lo cual tanto ellos como los del grupo experimental creerán que se les aplicó el tratamiento, y el efecto real de este se podrá distinguir del efecto de la sugestión. Este es un estudio *ciego*, en que los sujetos ignoran si están en el grupo experimental o el de control, lo que no evita otra amenaza a partir de las expectativas del experimentador, que pueden llevarlo a que sesgue la medición de resultados, privilegiando lo que busca, aun inconscientemente. Por ello hay experimentos *doblemente ciegos*, en los que también el experimentador ignora en qué grupo está cada sujeto, lo que solo saben quienes conducen el estudio, que no intervienen en la recolección de datos.

Un trabajo reciente muestra la dificultad de controlar todas las variables que pueden afectar los resultados de un experimento. Un grupo internacional de investigadores (Serge *et al.*, 2014) propusieron una solución para un añejo problema de estudios sobre estrés en ratones, con resultados inconsistentes por más cuidado que se pusiera en replicar exactamente todos los detalles. El grupo logró identificar la variable no controlada que afectaba los resultados: el sexo de los experimentadores: si eran varones, los ratones presentaban fuerte reacción de estrés, que los hacía menos sensibles al dolor; con investigadoras dicho efecto no se presentaba. Para confirmar que el efecto se debía al olor de varones o mujeres se pusieron en las jaulas de los ratones camisetas usadas por experimentadores de uno y otro sexo, y se observó un efecto idéntico. El sudor de los varones contiene más feromonas que el de las mujeres, lo que los ratones huelen y perciben como la presencia de machos y les produce estrés. (*Cfr*. González de Alba, 2014)

En educación los tratamientos son complejos, *v.gr.* un nuevo método de enseñanza de matemáticas o lectoescritura. Aún con buen entrenamiento, es difícil asegurar que todos los sujetos del grupo experimental apliquen de manera idéntica el nuevo método, y tampoco es fácil asegurar que ningún miembro del grupo control lo use al menos en parte, porque lo conoce de otra manera. Por eso importa cuidar la *fidelidad de la implementación* en los experimentos educativos.

Por razones prácticas o éticas, en educación no siempre hay condiciones para que los sujetos se asignen aleatoriamente, lo que lleva a hacer cuasiexperimentos, en los que se usan otras estrategias para asegurar la equivalencia de los dos grupos.

Aleatorización implícita, en diseños de *regresión con discontinuidad* (*regression discontinuity*). Si los sujetos cuyo resultado en una medición previa fue superior a un umbral definido son tratados de manera diferente a los de resultado inferior a ese *valor de corte*, por el margen de error de la medición se puede considerar que aquellos cuyo puntaje cae dentro del margen de error se asignan aleatoriamente.

Grupos similares por emparejamiento (*Propensity Score Matching*). Si se cuida que los grupos sean tan parecidos como se pueda en variables que son explicaciones alternas, por incidir en los resultados de cierta manera (tienen propensión). Para probar la hipótesis de que un método hace a los alumnos aprender más (*produce* aprendizaje) la variable que definirá los grupos a comparar es el método, y para descartar que la influencia causal se deba a variables como el sexo de los alumnos, o su nivel socioeconómico, se asegura que en los grupos cuasi-experimental y control haya la misma proporción de niños y niñas, de NSE alto y bajo.

Experimentos naturales. Los grupos a comparar se forman con sujetos que tienen o no naturalmente la característica que se piensa podría explicar el efecto.

Analizar los resultados implica usar técnicas como ANOVA o regresión, diferencias de medias, pruebas de significatividad, o medidas del Tamaño del Efecto.

Sobre este tema es fundamental la obra de Shadish, Cook y Campbell, 2002. Puede verse además Guo y Fraser, 2010; Morgan y Winship, 2007; Murnane y Willet, 2011; Murnane y Nelson, 2007; Nelson *et al.*, 2012; Rosenbaum, 2002; Rubin, 2006.

Estudios longitudinales

Las técnicas a las que se ha hecho referencia en incisos anteriores se aplican también al análisis de datos longitudinales, y en este caso son de especial interés los Modelos Lineales Jerárquicos, por tratarse de muestras no independientes.

Ya en 1962, en una reunión de especialistas para analizar los problemas que planteaba la medición del cambio de los fenómenos sociales se hacían propuestas sobre la confiabilidad de las mediciones, el uso de técnicas de análisis univariado o multivariado de varianza, el análisis factorial de matrices tridimensionales, e incluso una propuesta de Donald Campbell para interpretar las tendencias como cuasi-experimentos. (Harris, 1963)

Los estudios de esta familia incluyen las series de tiempo, de las que son de especial interés para la atribución de causalidad las series interrumpidas (*Interrupted Time-Series*), en las que se detecta un cambio en el posible efecto cuando ha habido un cambio en la serie, una *interrupción*.

Según Shadish, Cook y Campbell (2007: 175-179), mientras más larga sea la serie de observaciones *antes* de la interrupción, más razonable será descartar otras explicaciones del cambio observado *después* de la interrupción, como regresión a la media, deserción de participantes o maduración. Por otra parte, el análisis podrá revelar cambios bruscos en el intercepto o en la pendiente de la serie, o bien cambios pequeños y diferidos.

Los análisis del cambio con el concepto de *cohorte* distinguen el efecto de cohorte propiamente dicho, el de la edad de los sujetos, y el del tiempo en que se hace la medición (Ryder, 1965; Keeves, 1997). Los datos de estudios longitudinales se pueden analizar también con modelos causales y con modelos para la medición de variables latentes (Keeves, 1997: 148). El uso de Modelos Lineales Jerárquicos es apropiado. (Willet, 1997)

Además de las referencias citadas en los párrafos anteriores, dos obras sobre avances recientes en las técnicas para el estudio del cambio son la de Collins y Horn, 1991, y la de Singer y Willet, 2003.

Síntesis de investigaciones (meta-análisis)

Una consecuencia de la consolidación de los grupos de investigación educativa que ha tenido lugar en muchos países en las últimas décadas es la multiplicación de los estudios sobe un mismo tema, con lo que se han desarrollado también técnicas que tienen como propósito la integración de los resultados de estudios similares.

Si los hallazgos de estudios de distintos estudiosos, hechos en contextos diferentes, pero con mediciones y técnicas de análisis comparables, resultan coincidentes, el sustento de las conclusiones correspondientes será, desde luego, más sólido. Al integrar cuantitativamente los hallazgos de varios estudios, un *meta-análisis* reduce el margen de error de las mediciones en que se basa cada trabajo, al trabajar con un número de casos mucho mayor que el de cada uno de los estudios considerados.

El interés de este tipo de trabajos se puede apreciar considerando la atención que han suscitado varias obras de John Hattie y colaboradores, en las que se presentan resultados de cientos de meta-análisis, que sintetizan los hallazgos de miles de estudios, que cubrieron a millones de sujetos, explorando los factores que inciden en el desempeño de los estudiantes. (Hattie, 2009; Hattie y Anderman, 2013)

Las técnicas que se usan en este tipo de estudios incluyen pruebas estadísticas combinadas (de Fisher, Winer, Stouffer), con las que se busca descartar que los resultados obtenidos se deban al azar. Para estimar la fuerza de la relación entre las variables en juego se usan medidas del tamaño del efecto como la diferencia estandarizada de medias de Cohen (d), la razón de momios, o el coeficiente r de Pearson (Wolf, 1986).

Es posible analizar tamaños del efecto basados en regresiones (Hedges, 1994; Raudenbush, 1994), pero esto no es frecuente, ya que para ello es necesario que los estudios que se revisen, además de medir de igual forma la(s) variable(s) dependiente(s), para predecirlas o explicarlas consideren —al menos en parte— las mismas variables independientes, y que también las mediciones de estas sean comparables (Vogt, 2007: 312).

Para profundizar sobre este tema una obra importante, en Cooper y Hedges, 1994; también pueden verse Rosenthal, 1984; y Wolf, 1986.

Conclusión

Las observaciones o mediciones hechas en cualquier estudio nunca son perfectas; siempre tienen cierto grado de error que el investigador debe minimizar. De manera similar, los análisis de la información obtenida nunca serán totalmente precisos.

Las capacidades cognitivas de los humanos son notables, pero inevitablemente limitadas. Nunca podemos pretender que *hemos captado la realidad tal cual es, de manera*

perfectamente objetiva, sino que siempre hay que interpretar de alguna manera lo que se capta, en esos saltos cognitivos más o menos grandes que se conocen como inferencias, presentes en todo momento en la investigación, sea que se trate de describir uno o varios aspectos de la realidad, de encontrar asociaciones entre esos aspectos, de llegar a establecer relaciones de causalidad entre ellos, de generalizar lo encontrado en cierto número de casos a un universo más amplio, de identificar tendencias en el tiempo, etc.

Hay inferencias de distinto tipo, pero no se puede prescindir de ella, en el conocimiento ordinario o en el que genera la mejor investigación. Todo conocimiento que pretenda ser científico debe identificar las amenazas que pueden afectar las inferencias que se hagan y la forma de reducir su impacto. *La metodología de investigación se puede entender como la sistematización de las formas de cuidar la calidad de las inferencias.*

Lo anterior se aplica, desde luego, tanto a los estudios llamados cuantitativo como a los que se designa como cualitativos. Retomando la cita de King, Keohane y Verba:

> [...] una buena investigación puede ser cuantitativa o cualitativa en cuanto a estilo, pero en cuanto a diseño todas tienen estas características: su objetivo es la inferencia [...] Los procedimientos son públicos [...] Las conclusiones son inciertas [...] El contenido es el método [...] (1994: 9)

Las técnicas usadas en trabajos extensivos (cuanti) o intensivos (cuali) pueden ser distintas, pero las inferencias son similares, para describir, encontrar asociaciones, establecer causalidad, generalizar o identificar tendencias. Y los cuidados a tener para cuidar la solidez de las inferencias son también similares; no usar estadísticas no exime, por ejemplo, de que para interpretar una asociación en términos causales hay que descartar explicaciones alternativas, o que hay que desconfiar de la objetividad de cualquier observador, incluso (y, tal vez, sobre todo) de un observador participante.

Ningún diseño, y ninguna técnica de obtención o análisis de información, es superior en sí misma a otras, y la elección de cualquiera debe estar subordinada al propósito de la indagación y a las preguntas que la guíen. El autor que más he seguido en este capítulo escribe:

> Mi consejo es no comenzar revisando la gama de pruebas estadísticas que se podrían utilizar en una investigación. Hay docenas de ellas. No es eficiente aprender muchas técnicas previamente, por si acaso se llega a ofrecer utilizar alguna. Recomiendo, más bien, planteando una pregunta de investigación importante y diseñar un plan para recabar buenas

> evidencias relevantes para responderla. Una vez que usted conozca su diseño, y el tipo de evidencia que producirá, entonces usted deberá revisar los tipos de inferencias estadísticas relevantes para su trabajo. (Vogt, 2007: 143)

En otro pasaje este autor señala:

> [...] el punto crucial es que no hay manera estadística de que un investigador decida cuál es el mejor modelo para un estudio. Lo que se puede hacer con la estadística, después de haber construido el modelo y recogida la evidencia, es determinar que modelo se ajusta mejor a la evidencia, pero antes hay que tener el modelo. Sin él no se puede tener idea de qué datos reunir y cómo hacerlo. No hay sustituto estadístico, y ciertamente no hay sustituto computacional, por útiles que sean estas herramientas, que reemplace la reflexión seria sobre los problemas de investigación y sobre cómo recabar evidencia para resolverlos. La investigación cuantitativa de buena calidad no se trata principalmente de matemáticas. Tiene mucho más que ver con la lógica, con el conocimiento que tenga el investigador del tema y con el diseño escogido para recabar evidencia. Una onza de diseño vale una libra de análisis. (Vogt, 2007: 49)

En una línea similar, es pertinente citar una anécdota de uno de los autores más reconocidos del campo *cuantitativo*, que también hemos encontrado, Lee Cronbach, que también subraya la primacía que deben tener en todo proyecto las preguntas de investigación, por encima de las técnicas.

Otro investigador de primer nivel, Lee Shulman, comenta que, en 1986, al iniciar el desarrollo de un sistema para evaluar maestros que fuera más allá de los limitados enfoques entonces prevalecientes, buscó al asesor más calificado para cuidar la calidad de los instrumentos a desarrollar, y se dirigió a Cronbach, que era entonces considerado el *Zeus de la psicometría*, y explica que Cronbach aceptó dejar su retiro y atender la invitación, con una advertencia:

> Trabajaré en este proyecto con la condición de que ustedes se pregunten primero qué tipo de evaluaciones serán más fieles a la forma en que entienden la enseñanza y el aprendizaje, y es más probable que sean útiles para el campo. Entonces mi tarea será pensar cómo hacer que esas evaluaciones sean viables en términos psicométricos. El día en que yo vea que ustedes corrompen lo que estén haciendo para atender algún principio o práctica psicométrico, en ese momento regresaré a mi retiro. No podemos permitir que el rabo metodológico sea el que dirija al perro de la buena práctica pedagógica (Shulman, 2009: 240).

Referencia

Introducción. De la obtención de datos al análisis: la codificación

Andrews, F. M. ***et al.*** (1998). *Selecting Statistical Techniques for Social Science Data: A Guide for SAS Users*. Cary, NC: SAS Institute.

Best, J. (2009). *Uso y abuso de las estadísticas. La distorsión en la percepción pública de los problemas sociales y políticos*. Santiago de Chile: Editorial Cuatro Vientos. Edición original (2004). *More Damned Lies and Statistics. How Numbers Confuse Public Issues.* University of California Press.

Lewis-Beck, M. S. (1995). Data Analysis: *An Introduction*. Series Quantitative Applications in the Social Sciences, N° 103. Thousand Oaks: Sage.

Salkind, N. J. (2007). *Statistics for People Who* (*Think They) Hate Statistics. The Excel Edition*. Thousand Oaks: Sage Publications.

Vogt, W. P. (2007). *Quantitative Research Methods for Professionals.* Boston: Pearson Education.

Vogt, W. P., Vogt, E. R., Gardner, D. C. y Haeffele, L. M. (2014). *Selecting the Right Analyses for Your Data. Quantitative, Qualitative and Mixed Methods*. New York-London: The Guilford Press.

Fundamentos

Análisis descriptivo y exploratorio

Escobar, M. (1999). *Análisis gráfico/exploratorio.* Madrid: La Muralla.

Hartwig, F. y Dearing, B. E. (1979). *Exploratory Data Analysis*. Series Quantitative Applications in the Social Sciences, N° 16. Newbury Park: Sage.

Jacoby, W. G. (1998). *Statistical Graphics for Univariate and Bivariate Data*. Series Quantitative Applications in the Social Sciences, N° 117. Thousand Oaks: Sage.

Palmer Pol, A. L. (1999). *Análisis de datos. Etapa exploratoria.* Madrid: Ediciones Pirámide.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison Wesley.

Tukey, J. W. (1962). *The Future of Data Analysis. The Annals of Mathematical Statistics,* Vol. 33 (1): 1-67.

Weisberg, H. T. (1992). *Central Tendency and Variability*. Series Quantitative Applications in the Social Sciences, N° 83. Newbury Park: Sage.

Inferencia estadística

Coe, R. (2002). It's the Effect Size, Stupid. What effect size is and why it is important. En congreso anual de la *British Educational Research Association*, Exeter, 12-14 de septiembre.

Cohen, J. (1994). The Earth Is Round (p < .05). *American Psychologist*. Vol. 49 (12): 997-1003.

Flaherty, C. (2016). American Statistical Association seeks to usher in new era of statistical significance. *Inside Higher Ed*. Recuperado en agosto 14, 2019: https://www.insidehighered.com/news/2016/03/15/

Gigerenzer, G., Krauss, S. y Vitouch, O. (2004). The Null Ritual. What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. En Kaplan, D. (Ed.). *The SAGE Handbook of Quantitative Methodology for the Social Sciences* [21, pp. 391-408]. Thouand Oaks: Sage

Norton, B. y Strube, M. (2001). Understanding Statistical Power. *Journal of Orthopaedic & Sports Physical Therapy*, Vol. 31(6): 307-315.

Nuzzo, R. (2014). Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, 506: 150-152.

Prentice, D. A. y Miller, D. T. (1992). When Small Effects Are Impressive. *Psychological Bulletin,* Vol. 112 (1): 160-164.

Wasserstein, R. L. y Lazar, N. A. (2016). *The ASA's Statement on p-Values: Context*, Process, and Purpose. *The American Statistician*, 70:2, 129-133

Muestreo

Henry, G. T. (1998). Practical Sampling. En Bickman, L. y Debra J. R., *Handbook of Applied Social Research Methods* (pp. 101-126). Thousand Oaks: Sage.

Izcara Palacios, P. (2007). *Introducción al muestreo.* México: Miguel Ángel Porrúa.

Jaeger, R. M. (1984). *Sampling in education and the social sciences*. New York: Longman.

Análisis de la calidad de la información

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington: Authors.

Clark, L. A. y Watson, D. (1995). Constructing Validity: Basic Issues in Objective Scale Development. *Psychological Assessment*, Vol. 7 (3): 309-319.

Gardner, P. L. (1995). Measuring Attitudes to Science: Unidimensionality and Internal Consistency *Revisited. Research in Science Education*, Vol. 25 (3): 283-289.

Messick, S. (1989). Validity. En R. L. Linn (Ed.). *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education & Macmillan.

Shulman, L. S. (2009). Assessment of Teaching or Assessment for Teaching? Reflections on the Invitational Conference. En Gitomer, D. (Ed.). *Measurement Issues and Assessment for Teaching Quality*. Thousand Oakes: Sage, pp. 234-244

Técnicas básicas

Diferencias y asociación

Boudon, R. (1969). *Les méthodes en sociologie*. Collection Que sais'je? N° 1334. París: Presses Universitaires de France.

Dickinson Gibbons, J. (1993). *Nonparametric Measures of Association*. Series Quantitative Applications in the Social Sciences, N° 91. Newbury Park: Sage.

Liebetran, A. M. (1983). *Measures of Association*. Series Quantitative Applications in the Social Sciences, N° 32. Beverly Hills: Sage.

Correlación espuria y control de variables

Boudon, R. (1970). A propos d'un livre imaginaire. En Lazarsfeld, P. *Philosophie des sciences sociales* (Introduction, pp. 7-72). París: Gallimard.

Cortés, F. y Rubalcava, R. Ma. (1987). *Métodos estadísticos aplicados en investigación en ciencias sociales*. Análisis de asociación. México: COLMEX.

Lazarsfeld, P. (1955). Interpretation of statistical relations as a research operation. En Lazarsfeld, P. F. y Rosenberg, M. (Eds.). *The Language of Social Research* (pp. 115-125). New York-London: The Free Press-Collier Macmillan Ltd.

Análisis de varianza

Salkind, N. J. (2007). *Statistics for People Who (Think They) Hate Statistics. The Excel Edition.* Thousand Oaks: Sage Publications.

Análisis de regresión

Licht, M. H. (1997). Multiple Regression and Correlation. En Grimm y Yarnold, *Reading and Understanding Multivariate Statistics* (pp. 19-64). Washington: American Psychological Association.

Vogt, W. P. (2007). *Quantitative Research Methods for Professionals.* Boston: Pearson Education.

Técnicas avanzadas

De medición

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City: American College Testing Program. (Edición revisada en 1992).

Brennan, R. L. (2001). An Essay on the History and Future of Reliability from the Perspective of Replications. *Journal of Educational Measurement*, Vol 38 (4): 295-317.

Buckley, J. (2009). Cross-national Response Styles in International Educational Assessments: Evidence from PISA 2006. Original no publicado.

Cronbach, L. J., Gleser, G. C., Nanda, H., y Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Downing, S. M. y Haldyna, T. M. (Eds.). (2006). *Handbook of Test Development.* Mahwah, N. J.: Lawrence Erlbaum Assoc. Publ.

Haertel, E. H. (2006). Reliability. En Brennan, R. L. (Ed.). *Educational Measurement*, 4th Ed. [3, pp. 65-110]. Westport: American Council on Education-Praeger Publ.

Hill, Heather C. Charalambos Y. Ch. y Kraft, Matthew A. (2012). When Rater Reliability is not enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, 41 (2): 56-64.

Kane, M. T. (2006). Validation. En Brennan, R. L. (Ed.). *Educational Measurement*, 4th Ed. [2, pp. 17-64]. Westport: American Council on Education-Praeger Publ.

Martínez, J. F. (2015). *Buena enseñanza, error y validez: aspectos conceptuales y metodológicos.* 2do Congreso Latinoamericano de Medición y Evaluación Educacional COLMEE, México DF, Marzo 13.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.

Shavelson, R. J. y Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park: Sage.

Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach*. Mahwah, N. J.: Lawrence Erlbaum Assoc. Publ.

ANCOVA y MANOVA

Holland, P. W., y Rubin, D. B. (1986). Research Designs and Causal Inference: On Lord's Paradox. En Pearson, R. W., y Boruch, R. F. (Eds.). Survey Research Designs: *Towards a Better Understanding of Their Costs & Benefits* [Lecture Notes in Statistics, 38: 7-37]. New York: Springer.

Miller, G. A. y Chapman, J. P. (2001). Misunderstanding Analysis of Covariance. *Journal of Abnormal* Psychology, Vol. 110 (1): 40-48.

Owen, S. V. y Froman, R. D. (1998). Uses and Abuses of the Analysis of Covariance. *Research in Nursing & Health*, 21: 557-562.

Weinfurt, K. P. (1997). Multivariate Analysis of Variance. En Grimm, L. G. y Yarnold, P. R. (Eds.). *Reading and Understanding Multivariate Statistics* [pp. 245-276]. Washington: American Psychological Association.

Weinfurt, K. P. (2000). Repeated Measures Analyses: ANOVA, MANOVA and HLM. En Grimm, L. G. y Yarnold, P. R. (Eds.). *Reading and Understanding More Multivariate Statistics* [pp. 317-361]. Washington: American Psychological Association.

Análisis de regresión

Bryk, A. S., y Raudenbusch, S. W. (1992). *Hierarchical Linear Models.* Advanced Quantitative Techniques in the Social Science Series 1. Newbury Park: Sage Publications.

Gelman, A. y J. Hill. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge-New York: Cambridge Univ. Press.

Kaplan, D. (2000). *Structural Equation Modeling. Foundations and Extensions*. Thousand Oaks: Sage.

Klem, L. (1997). Path Analysis. En Grimm, L. G. y Yarnold, P. R. (Eds.). *Reading and Understanding Multivariate Statistics* [pp. 65-97]. Washington: American Psychological Association.

Klem, L. (2000). Structural Equation Modeling. En Grimm, L. G. y Yarnold, P. R. (Eds.). *Reading and Understanding More Multivariate Statistics* [pp. 227-260]. Washington: American Psychological Association.

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks: Sage.

Wright, R. E. (1997). Logistic Regression. En Grimm, L. G. y Yarnold, P. R. (Eds.). *Reading and Understanding Multivariate Statistics* [pp. 217-244]. Washington: American Psychological Association.

Análisis factorial

Bryant, F B., y Yarnold, P. R. (1997). Principal-Components Analysis and Exploratory and Confirmatory Factor Analysis. En Grimm, L. G. y Yarnold, P. R. (Eds.). *Reading and Understanding Multivariate Statistics* [pp. 99-136]. Washington: American Psychological Association.

Análisis de datos en diseños particulares

Estudios experimentales y cuasi-experimentales

Guo, S. y Fraser, M. W. (2010). *Propensity Score Analysis. Statistical Methods and Applications.* Advanced Quantitative Techniques in the Social Science Series 11. Thousand Oaks: Sage Publications.

Morgan, S. L., y Winship, Ch. (2007). *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. Oxford-New York: Oxford University Press.

Murnane, R. J., y Willet, J. B. (2011). Methods Matter. Improving Causal Inference in Educational and Social Science Research. Oxford-New York: Oxford University Press.

Murnane, R. J. y Nelson, R. R. (2007). Improving the Performance of the Education Sector: The Valuable, *Challenging, and Limited Role of Random Assignment Evaluations. Economics of Innovation and New Technology,* Vol. 16 (5): 307-322.

Nelson, M. C., Cordray, D. S., Hulleman, Ch. S., Darrow, C. L., y Sommer, E. C. (2012). A Procedure for Assessing Intervention Fidelity in Experiments Testing Educational and Behavioral Interventions. *The Journal of Behavioral Health Services & Research*, Vol. 39 (4): 374-396.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects.* Oxford-New York: Oxford University Press.

Shadish, W. R., Cook, Th. D., y Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston-New York: Houghton Mifflin Co.

Estudios longitudinales

Collins, L. M., y Horn, J. L. (1991). *Best Methods for the Analysis of Change. Recent Advances, Unanswered Questions, Future Directions*. Wasington: American Psychological Association.

Harris, Ch. W. (Ed.). (1963). *Problems in Measuring Change*. Madison: The University of Wisconsin Press.

Keeves, J. P. (1997). Longitudinal Research Methods. En Keeves, J. P. (Ed.). *Educational Research, Methodology and Measurement. An International Handbook* [pp. 138-149]. Oxford-New York: Pergamon.

Ryder, N. B. (1965). The Cohort as a Concept in the Analysis of Social Change. *American Sociological Review,* Vol. 30 (6): 843-861.

Singer, J. D., y Willet, J. B. (2003). *Applied Longitudinal Analysis. Modeling Change and Event Occurrence*. Oxford-New York: Oxford Univ. Press.

Willet, J. B. (1997). Change, Measurement of. En Keeves, J. P. (Ed.). *Educational Research, Methodology and Measurement. An International Handbook* [pp. 327-334]. Oxford-New York: Pergamon.

Síntesis de investigaciones

Cooper, H. M. y Hedges, L. V. (Eds.). (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation. En especial: Hedges, Larry V. Fixed Effects Models, pp. 285-300. y Raudenbush, S. E. (1994). Random Effects Models, pp. 301-322.

Hattie, J. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement.* London-New York: Routledge.

Hattie, J. y Anderman, E. M. (Eds.) (2013). *International Guide to Student Achievement.* London-New York: Routledge.

Rosenthal, R. (1984). *Meta-analitic Procedures for Social Research.* Beverly Hills: Sage.

Wolf, F. M. (1986). *Meta-analysis. Quantitative Methods for Research Synthesis.* Beverly Hills: Sage.

Apéndice. Tablas del Modelo de Elaboración de Lazarsfeld

Preferencia por programas de tipo religioso

TABLA 4A.1. PREFERENCIA POR PROGRAMAS RELIGIOSOS, POR EDAD

$r_t = - 0.2$

Audiencia	Edad		Totales
	Jóvenes	Viejos	
Escuchan	170	338	508
No escuchan	830	962	1792
Totales	1000	1300	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.2. PREFERENCIA POR PROGRAMAS RELIGIOSOS POR EDAD, CONTROLANDO NI

NI Alto r_t -.04

NI Bajo r_t -.05

Audiencia	Nivel de instrucción				Totales
	Alto		Bajo		
	Jóvenes	Viejos	Jóvenes	Viejos	
Escuchan	55	45	115	285	500
No escuchan	545	355	285	615	1800
Totales	600	400	400	900	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.3. PREFERENCIA POR PROGRAMAS RELIGIOSOS, POR NIVEL DE INSTRUCCIÓN

$r_t = - 0.5$

Audiencia	Nivel de Instrucción		Totales
	Alto	Bajo	
Escuchan	100	400	500
No escuchan	900	900	1800
Totales	1000	1300	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.4. PREFERENCIA POR PROGRAMAS RELIGIOSOS POR NI, CONTROLANDO EDAD

Jóvenes r_t -.49

Viejos r_t -.47

Audiencia	Edad				Totales
	Jóvenes		Viejos		
	NI Alto	NI Bajo	NI Alto	NI Bajo	
Escuchan	55	115	45	285	500
No escuchan	545	285	3555	615	1800
Totales	600	400	400	900	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

Preferencia por programas políticos

TABLA 4A.5. PREFERENCIA POR PROGRAMAS POLÍTICOS, POR EDAD

$r_t = -0.18$

Audiencia	Edad		Totales
	Jóvenes	Viejos	
Escuchan	340	585	925
No escuchan	660	715	1375
Totales	1000	1300	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.6. PREFERENCIA POR PROGRAMAS POLÍTICOS POR EDAD, CONTROLANDO NI

NI Alto r_t -.23

NI Bajo r_t -.26

Audiencia	Nivel de Instrucción				Totales
	Alto		Bajo		
	Jóvenes	Viejos	Jóvenes	Viejos	
Escuchan	240	220	100	360	920
No escuchan	360	180	300	540	1380
Totales	600	400	400	900	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.7. PREFERENCIA POR PROGRAMAS POLÍTICOS, POR NIVEL DE INSTRUCCIÓN

$r_t = 0.17$

Audiencia	Nivel Instrucción		Totales
	Alto	Bajo	
Escuchan	460	460	920
No escuchan	540	840	1380
Totales	1000	1300	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.8. PREFERENCIA POR PROGRAMAS POLÍTICOS POR NI, CONTROLANDO EDAD

Jóvenes r_t .17

Viejos r_t .27

Audiencia	Edad				Totales
	Jóvenes		Viejos		
	NI Alto	NI Bajo	NI Alto	NI Bajo	
Escuchan	240	100	220	360	920
No escuchan	360	300	180	540	1380
Totales	600	400	400	900	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

Preferencia por programas musicales

TABLA 4A.9. PREFERENCIA POR PROGRAMAS MUSICALES, POR EDAD

$r_t = 0.02$

Audiencia	Edad		Totales
	Jóvenes	Viejos	
Escuchan	300	377	677
No escuchan	700	923	1623
Totales	1000	1300	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.10. PREFERENCIA POR PROGRAMAS MUSICALES POR EDAD, CONTROLANDO NI

NI Alto r_t -.32

NI Bajo r_t .20

Audiencia	Nivel de Instrucción				Totales
	Alto		Bajo		
	Jóvenes	Viejos	Jóvenes	Viejos	
Escuchan	192	208	112	171	683
No escuchan	408	192	288	729	1617
Totales	600	400	400	900	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.11. PREFERENCIA POR PROGRAMAS MUSICALES POR NIVEL DE INSTRUCCIÓN

$r_t = 0.33$

Audiencia	Nivel Instrucción		Totales
	Alto	Bajo	
Escuchan	400	283	683
No escuchan	600	1017	1617
Totales	1000	1300	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

TABLA 4A.12. PREFERENCIA POR PROGRAMAS MUSICALES POR NI, CONTROLANDO EDAD

Jóvenes r_t .08

Viejos r_t .54

Audiencia	Edad				Totales
	Jóvenes		Viejos		
	NI Alto	NI Bajo	NI Alto	NI Bajo	
Escuchan	192	112	208	171	683
No escuchan	408	288	192	729	1617
Totales	600	400	400	900	2300

FUENTE: ELABORACIÓN PROPIA, CON DATOS DE LAZARSFELD, 1955.

CONTENIDO

Introducción

A mediados de la década de 1980 me parecía que muchos posgrados que buscaban formar investigadores en educación o ciencias sociales no daban a sus egresados una buena preparación, sobre todo en lo metodológico, gracias a la cual estuvieran al menos en condiciones de *detectar el carácter espurio de las relaciones que parezca haber entre las variables clave.* A mi juicio tal hecho constituía un fraude, que daría derecho a esos exalumnos a pedir que se les devolviera lo que hubieran pagado por una formación tan limitada, según el dicho de W. Spady. A pesar de los avances alcanzados, hoy pienso lo mismo de muchas maestrías y no pocos doctorados. Esta opinión, basada en el conocimiento directo o indirecto de muchos programas, está en el origen de este *Nuevo Oficio del Investigador Educativo,* que pretende ayudar a que la formación de los futuros investigadores sea más sólida, en particular en lo metodológico.

¿Por qué no se ha avanzado en este terreno? No es algo exclusivo de México; algo similar se encuentra en otros países, comenzando por Estados Unidos. En 1989 un especialista en temas de metodología observaba que los estudiantes de posgrado que atendía en esas fechas tenían una preparación en matemáticas peor, y llevaban menos cursos de formación metodológica, que los alumnos de las décadas de 1950 y 1960. Añadía que los aspirantes a entrar a posgrados de ciencias sociales suelen tener resultados bajos en las pruebas de ingreso, "y no sólo en razonamiento cuantitativo, sino también en razonamiento verbal...". (Blalock, 1989: 448-458)

Hacia 1960 y 1970 se dio en México un fuerte rechazo de la metodología llamada cuantitativa, en parte debido a fallas reales: no pocos trabajos ocultaban sus límites tras la cortina de humo de cuadros estadísticos de apariencia impresionante, aunque

de contenido poco claro. Una visión pobre de la metodología, reducida a un conjunto limitado de técnicas *cuantitativas*, provocó el rechazo de estudiosos *cualitativos*, pero muchas veces sin distinguir el grano de la paja, incluyendo el rechazo de elementos sólidos de la metodología convencional.

En momentos álgidos de la reacción contra lo *cuantitativo* se llegaron a suprimir los cursos de estadística en carreras como sociología y economía, por considerar que se trataba de una disciplina burguesa. Hasta mediados del siglo XX, la investigación social y educativa se basaba sobre todo en corrientes teóricas como el conductismo en psicología y el funcionalismo en sociología; en lo metodológico predominaban diseños experimentales y encuestas, del enfoque llamado *cuantitativo*. El rechazo de los años 1960 fue acompañado por el surgimiento de trabajos de la orientación alternativa, (*cualitativa*), con estudios antropológicos, etnográficos, interpretativos, basados en la llamada teoría fundamentada (*grounded theory*), entre otros.

Hoy los momentos más duros del enfrentamiento entre esas formas de entender la investigación (*guerras paradigmáticas*) parecen quedar atrás, pero en los posgrados que forman investigadores educativos sigue presente cierto rechazo de las técnicas de investigación extensiva y estructurada, y preferencia por enfoques intensivos e interpretativos, que no siempre se manejan con rigor y con frecuencia muestran deficiencias serias, como las que advertía Eduardo Weiss (2017) en el artículo citado al hablar de los estudios de casos en el Capítulo 2.

Hay razones para temer que el rechazo de la estadística y todo lo que suene a cuantitativo no se deba a decisiones basadas en el tipo de pregunta de investigación que se quiere responder, sino simplemente en la débil formación matemática de muchos estudiantes. Hay que reconocer que también en México las políticas de admisión a posgrado de muchas instituciones permiten que lleguen a programas orientados a investigación, que deberían ser los más exigentes, aspirantes sin la capacidad necesaria para cursar estudios superiores.

Cómo mejorar la formación de los futuros investigadores

La necesidad de fortalecer la preparación de los futuros investigadores es aún más clara por los cambios que enfrenta el oficio. Sigue siendo necesario asegurar la capacidad de plantear preguntas investigables; utilizar diseños adecuados para darles respuesta, y herramientas apropiadas para obtener información de buena calidad; analizar la información con técnicas adecuadas; llegar a inferencias sólidas descriptivas, asociativas, causales o generalizadoras, y a conclusiones sustentadas en la evidencia.

Pero, a diferencia de hace 15 o 20 años, quien hoy quiere dedicarse de manera profesional a esta actividad debe tener doctorado y pertenecer al SNI; navegar en grandes bases bibliográficas en línea; utilizar redes digitales para obtener información; manejar con soltura paquetes de software; publicar en revistas indexadas; formar parte de redes internacionales de estudiosos del tema, etc. Por ello, la formación debe actualizarse, para desarrollar esas competencias.

Una competencia implica integrar conocimientos, habilidades y actitudes. Las que debe dominar quien quiera dedicarse a investigación educativa tienen, en efecto, un componente cognitivo, uno actitudinal y uno afectivo, que en otro lugar propuse designar con tres términos griegos clásicos: l*ogos, ethos* y *pathos.*

Los componentes actitudinal y afectivo (*ethos-pathos*) no se desarrollan con materias o cursos especiales, sino ante todo por la vivencia de unos valores compartidos, y de éxitos y fracasos compartidos. El componente cognitivo, el *logos* sí debe ser objeto de una formación sistemática, y comprende la formación teórica y la metodológica.

- Formación teórica: se refiere al núcleo básico de conocimientos necesario para delimitar objetos de estudio con precisión suficiente que haga posible un trabajo común. Eso incluye unos conceptos clave, ciertos principios con carácter casi axiomático, unos tipos de explicación y determinados autores de referencia. Se trata de aprovechar lo que ya se sabe sobre el objeto de estudio, y se consigue mediante la lectura de literatura pertinente, pero precisando dos ideas:
  - Distinguir teorías macro y micro. En cada disciplina hay grandes teorías de autores como Skinner, Piaget, Vygotsky o Bruner en psicología, o Durkheim, Weber, Parsons o Bourdieu en sociología. Quien quiera trabajar un campo en que esas disciplinas son relevantes no puede ignorar a tales autores, sabiendo que aportan poco sobre un tema particular. Hay además trabajos menos ambiciosos (teorías micro) de autores cuya aportación no alcanza la trascendencia de los más grandes, pero más útiles para explorar un tema.
  - Distinguir que, además de campos en los que se cuentas con teorías micro que dan cuenta de lo que ocurre en ellos, hay campos en que no hay tal cosa, pero que en ellos trabajan profesionales que tienen conocimientos ricos de dichos fenómenos, aunque en muchos casos sean implícitos. La educación es uno de estos campos, y por ello lo que se sabe sobre muchos temas (como los factores de la deserción, el impacto de ciertas estrategias didácticas, o las escuelas eficaces) puede provenir de la literatura, pero también del conocimiento de maestros o directores experimentados.

- Formación metodológica: ha sido el foco de atención de esta obra, incluyendo el dominio de técnicas de obtención y análisis de la información, y de diseños de investigación, y no es necesario abundar al respecto. En varios lugares he aludido a cuestionamientos de quienes critican los enfoques *cuantitativos*, algunos de los cuales no se refieren a detalles técnicos, sino que incluyen puntos relativos a la naturaleza de la realidad y el conocimiento que se puede tener de ella, o sea que llegan a un nivel filosófico, y en particular de teoría de la ciencia y el conocimiento, de epistemología. En el segundo apartado de esta conclusión se abordará este punto.

Por otra parte, la formación necesaria para llegar a ser buen investigador ha sido abordado de dos maneras opuestas y, en mi opinión, igualmente insuficientes.

- En un extremo, se pretende formar investigadores mediante un entrenamiento de contenido preciso y enfoque técnico. A partir de una visión escolar y simplista de las estrategias de la ciencia (*el método científico*), y de la absolutización de un diseño como la encuesta, se enseñan técnicas de muestreo, elaboración de cuestionarios y procesamiento de datos, a nivel elemental, y se espera que quien sea capaz de aplicar aceptablemente tales técnicas será un buen investigador.
- En el polo opuesto, se afirma que es imposible sistematizar la enseñanza de la investigación, sobre la base de una afirmación que, tomada literalmente y sin matices, es una obviedad: el carácter único de cada objeto y su conexión con la totalidad. Con una preferencia por estudios de caso y enfoques etnográficos, la noción clave, escolar y simplista también, es la de creatividad: de nada sirven los cursos; lo que se necesita es lanzarse al terreno y aprender sobre la marcha. La única forma de aprender a investigar es investigando, se dice, como la única de aprender a nadar es lanzándose al agua.

Considero inaceptables los dos extremos. Reconozco que la investigación no se reduce a ninguna técnica, ni combinación de técnicas, por lo que la capacidad analítica y sintética del investigador y, si se quiere, su creatividad, son ingredientes indispensables de un trabajo de calidad, pero creo que la capacidad de investigación es compleja, y que sus componentes son susceptibles de desarrollarse en diferente medida y distinta forma. Siguiendo la comparación con la natación: es posible aprender a nadar lanzándose

sencillamente al agua, pero así nadie llegará a ser un excelente nadador. Quien aspire a serlo deberá someterse a la disciplina de un largo entrenamiento, durante el cual, en algunos momentos deberá practicar solo una habilidad particular —el movimiento de las piernas, la forma de respirar— y en otros integrará diversas habilidades particulares en la habilidad mayor de *nadar*.

Sin duda el aprendizaje de la investigación no puede asimilarse al de destrezas psicomotrices como la de nadar, por lo que la comparación no debe exagerarse, pero la idea fundamental es importante: debemos identificar analíticamente los componentes de la habilidad general de hacer investigación, y entonces podremos preguntarnos sobre la forma de propiciar el desarrollo de cada uno.

Los componentes de la capacidad de investigación

Los elementos que deben conjuntarse para que se forme un buen investigador se pueden resumir en seis puntos:

a. Capacidad intelectual. La tarea de un investigador no supone alguna inteligencia particular, verbal o numérica, espacial, artística o emocional, pero sí la forma general de inteligencia que se manifiesta en la capacidad de análisis y síntesis, o de operaciones abstractas.
b. Capacidad de lectura y expresión oral y escrita. Para investigar sobre cualquier tema hay que saber lo que han encontrado otros; es necesario también producir textos bien estructurados y redactados, de acuerdo al género literario de un artículo especializado, una obra monográfica o un texto de divulgación, para poner al alcance de otros el resultado del trabajo propio, en beneficio de los demás y del mismo autor, que será el primero en enriquecerse con las críticas.
c. Buen conocimiento del campo de que se trate. Sin desconocer que sus avances no son lineales, un rasgo de la ciencia es su carácter acumulativo, como fruto del trabajo de personas y grupos que dedican su atención a ciertos temas. Con la dificultad que implica la publicación de millares de artículos al año, en centenares de revistas especializadas, es impensable en la actualidad que un buen investigador ignore los trabajos importantes relacionados con su área.
d. Dominio de técnicas de obtención y análisis de información. Aunque el dominio de técnicas no hace por sí sólo al científico, el buen investigador no puede ignorarlas. Para recabar información podrá tener auxiliares, pero los deberá capacitar y supervisar. Podrá tener el apoyo de un especialista para el análisis, y

la última versión de un software, pero si no puede seleccionar el tipo de análisis más adecuado, e interpretar los resultados, no será un buen investigador. Los ejemplos no implican que la única forma de hacer buena investigación es con técnicas estadísticas complejas, pero la idea se aplica igual en el caso de otros enfoques: si quiere ser un buen investigador, el responsable de un proyecto con un acercamiento de tipo etnográfico no puede ser lego en los procedimientos respectivos, aunque tenga recursos para pagar un auxiliar especializado.

e. Actitudes y disposiciones adecuadas. Como curiosidad, rigor, laboriosidad, exigencia, crítica y autocrítica, hábitos de trabajo intenso y regular, o disposición favorable para el trabajo en equipo. La naturaleza colectiva del trabajo académico destaca la importancia de estos rasgos, sin los cuales la capacidad intelectual y la formación teórico-metodológica pueden resultar poco productivas.
f. Capacidad de conjuntar los ingredientes. Un excelente investigador necesita los elementos anteriores, pero además debe combinarlos de manera armoniosa en el contexto de cada trabajo particular. El todo es más que la suma de las partes.

El desarrollo de los componentes: la formación

Identificados los ingredientes de la capacidad de investigación, preguntemos de nuevo: ¿es posible desarrollarlos de manera intencional y sistemática? ¿Se puede sistematizar la formación de un investigador? Veamos.

a. Capacidad intelectual. Según las ciencias cognitivas es posible desarrollar la capacidad de pensamiento lógico y razonamiento abstracto, pero no es fácil y mientras más tarde, peor. Aplica el viejo dicho de que lo que la naturaleza no dio no lo presta Salamanca. No es razonable que los programas para formar investigadores incluyan entre sus objetivos tareas remediales tan complejas. Por ello maestrías y doctorados que pretendan formar para la investigación deben seleccionar rigurosamente a sus alumnos, cuidando que en el proceso de admisión se asegure un nivel adecuado de capacidad intelectual.
b. Capacidad de lectura, expresión oral y escrita. Estas habilidades sí pueden desarrollarse sistemáticamente y no implican enfoques sofisticados, sino constancia en el esfuerzo y un proceso de correcciones y retroalimentación constantes, para que la interiorización de esas prácticas las vuelva *hábitos*. Los programas que pretendan formar investigadores de alto nivel deberán hacer que sus alumnos lean y escriban mucho, con los mecanismos de retroalimentación necesarios

para que la calidad de la lectura y la redacción alcance niveles que correspondan a una formación de posgrado.

c. Conocimiento de un(os) campo(s) del conocimiento. Suponiendo la presencia de las dos habilidades anteriores, también esta se puede desarrollar en forma sistemática. La lectura de autores clave; la comprensión cabal de sus ideas; la contrastación de esas ideas con las de otros autores; la crítica que detecte los puntos débiles y fuertes; y, por fin, la construcción de síntesis propias, son tareas esenciales de la formación de un investigador, que pueden y deben hacerse en forma continua y sistemática, conjuntando el esfuerzo del estudiante, la orientación del maestro y la discusión y el diálogo en el grupo de personas de diversos niveles de experiencia que constituye un buen seminario.

d. Dominio de técnicas. También es un componente que se puede enseñar y aprender de manera sistemática. Se tiende a menospreciar su importancia, dando la impresión de que puede prescindirse de la tarea laboriosa de dominar diversas técnicas. La postura correcta no es difícil de establecer: un buen investigador deberá dominar los principios de las principales técnicas de su campo, y haber adquirido un buen dominio de algunas.

Esta obra es una propuesta sobre lo que debería comprender una sólida formación metodológica, incluyendo lo relativo a técnicas de obtención y análisis de la información, y también los diseños de investigación, con énfasis en los acercamientos estructurados. Cada uno de sus capítulos podría dar lugar a un curso semestral, sin olvidar los consejos con que cierra el Capítulo 3, de que no tiene sentido tratar de aprender el mayor número de técnicas que se pueda, por si llega a presentarse la necesidad de utilizar alguna, sino que basta tener una buena visión panorámica, reconociendo que las preguntas de investigación deben tener la primacía.

e. Actitudes y disposiciones adecuadas. Es más difícil sistematizar este ingrediente que los anteriores, de suerte que no resulta adecuado un enfoque directo, en un hipotético curso o taller de actitudes o algo similar. Pero tampoco estamos ante algo azaroso: el desarrollo de actitudes y disposiciones favorables para la investigación se da en la interacción cotidiana del aprendiz con quienes han desarrollado previamente tales elementos y los ponen en práctica en su quehacer diario. Por ello la formación para la investigación se da, deseablemente, en el seno de grupos establecidos en los que prevalezca ese tipo de *ethos*.

f. Capacidad de conjuntar los elementos. Este componente tampoco debe ser objeto de cursos especiales, pero sí de un apoyo tutorial al aprendiz por un investigador con más experiencia que, en diálogo, le ayude a clarificar ideas y alcanzar la síntesis personal que es la culminación del trabajo.

En el texto en que propuse la terna *logos-ethos-pathos*, al referirme al tercero de estos elementos, decía que, como sugiere la palabra, se refiere a aspectos emotivos o afectivos que también deben aglutinar a un grupo de investigación: esta es una tarea ardua, que puede ser considerada tediosa por quien no la comprenda. Además, nunca podrá competir con otras en cuanto a los ingresos que produce. El ejercicio de las profesiones liberales, el comercio o las finanzas, por no hablar de los espectáculos, o el deporte profesional, proporcionan ingresos mucho mayores. Lo que atrae al buen investigador es algo más profundo: un gran aprecio por el valor intrínseco de la ciencia, por la satisfacción incomparable de ir conquistando palmo a palmo el conocimiento de una parcela de la realidad; es la emoción inmensa del *eureka* de Arquímedes corriendo semidesnudo por las calles de Siracusa. Es ese *pathos* que a veces parece llegar al *eros*: se dice de la ciencia que es una amante celosa que no tolera rivales.

Conformar un grupo de investigación es el largo y complicado proceso de identificar un objeto de estudio común y construir una perspectiva compartida para abordarlo, es la interiorización y la vivencia cotidiana de un conjunto de valores y de normas de comportamiento, y es la sintonía emotiva que hace que un grupo se apasione por los mismos propósitos, sufra por los mismos tropiezos, y comparta su gozo por los mismos triunfos. Supone mucho más que la contratación de varios especialistas. Un mínimo de recursos es condición necesaria, pero no suficiente. Si se logra integrar un grupo alrededor de una tradición, la búsqueda de los recursos será una parte más del reto cotidiano. En cambio, si no hay unidad de miras, de estilo de trabajo y de entusiasmos, los recursos materiales podrán producir resultados de corto plazo que engañen a quien los vea superficialmente, pero no lograrán resultados de calidad a largo plazo.

La receta que propongo puede resumirse diciendo que un programa logrará formar buenos investigadores si selecciona cuidadosamente a sus alumnos; si los hace leer y escribir mucho y los retroalimenta; si los hace dialogar con buenos autores de su campo y llegar a síntesis propias; si los hace adquirir un dominio de una gama adecuada de técnicas; si, gracias a la vivencia diaria en el grupo de trabajo, propicia en ellos el desarrollo de un *ethos* de investigación; y si los investigadores de mayor experiencia

del grupo consideran como su mayor logro el que sus alumnos lleguen a producir obras propias bien acabadas, por medio de las cuales los superen.

La formación en aspectos epistemológicos

En ocasiones el rechazo de lo *cuantitativo* y la opción por lo *cualitativo* pretende sustentarse en la epistemología, pero con una perspectiva superficial e inexacta del tema, que se manifiesta en dicotomías simplistas, en especial la que contrapone al positivismo con posturas interpretativas, críticas, hermenéuticas, fenomenológicas y similares. Esas dicotomías se extienden al terreno metodológico, donde se oponen los acercamientos *cualitativos* a los *cuantitativos*, asociando los primeros con las posturas interpretativas y críticas, y los segundos con el positivismo. Se suele decir, además, que el positivismo y los enfoques cuantitativos serían adecuados para las ciencias de la naturaleza, pero no para las ciencias sociales y humanas, ya que en estas sólo serían aceptables enfoques cualitativos, con epistemologías alternativas.

La especificidad de los temas epistemológicos hace impensable que los dominen todos los investigadores educativos, lo que corresponde a los especialistas de ese campo; pero como la epistemología se usa como sustento de las posturas simplistas mencionadas es conveniente que quienes desean dedicarse de manera profesional a la investigación tengan ideas claras al respecto, pues las posturas radicales que se encuentran en la literatura de investigación educativa no tienen sustento real en lo que hoy sostiene en general la comunidad de profesionales de la epistemología.

Con este espíritu abordaré estos temas a partir de una reflexión que dura décadas y recogen varias publicaciones (Martínez Rizo, 1986; 1991; 1993; 1997; 1999; 2000; 2002). Para un tratamiento más amplio ver Phillips, 2014; Phillips y Burbules, 2000; Blaug, 1980; Onwuegbuzie y Wisdom, 2014. En particular, sostendré:

- Que el positivismo decimonónico y el neopositivismo dejaron hace tiempo de ser defendidos como tales, y han dejado el lugar a varios postpositivismos.
- Que la versión decimonónica del historicismo y la hermenéutica también ha sido sustituida por ideas más actuales.
- Que las versiones actuales de ambas posturas tienen amplias coincidencias, en especial el reconocer que no hay un conocimiento perfectamente objetivo de la realidad, y que siempre hay algún tipo de inferencia o interpretación.
- Que ambas posturas coinciden también en rechazar las versiones extremas del irracionalismo posmoderno, pues no hay conocimientos perfectos, pero sí los

hay de muy distinta solidez, y la tarea del investigador es, justamente, producir conocimientos lo más sólidos que pueda.

Sobre la relación del positivismo con sus herederos

En un agudo análisis de las inexactitudes que se comenten al hablar del positivismo, Bouveresse decía que tratar el tema en 1980 no era fácil, por tres razones: porque el valor de la ciencia estaba en tela de juicio y las corrientes, como el positivismo, que daban un lugar primordial al conocimiento científico, estaban desacreditadas; porque el positivismo era la corriente filosófica que más mala fama tenía en Francia; y porque la palabra tenía un sentido polémico sin contenido descriptivo real. Según Bouveresse, *en la epistemología de Althusser, positivista no significa gran cosa más que no histórico, no marxista o incluso simplemente no francés*. (1980: 51)

En medios especializados, en cambio, se cree que el positivismo y, en particular el neopositivismo —de moda en la primera mitad del siglo XX, con el Círculo de Viena— es una corriente que ya nadie sostiene como tal, y ha dejado el lugar a posturas en las que influyó, pero que difieren en puntos centrales. El más fundamental se refiere a la forma de concebir el conocimiento, en general y el propio de las ciencias, que para el positivismo y el neopositivismo depende necesariamente de la experiencia sensorial. De manera más precisa, siguiendo a Ladriere:

> [...] habría que formular este principio en términos lógicos como lo hacen los neopositivistas: únicamente tienen sentido las proposiciones analíticas (que, siendo tautológicas, son independientes de la experiencia), y las proposiciones sintéticas a posteriori. (1971:118)

Según las dicotomías simplistas, las corrientes opuestas al positivismo diferirían de este por postular que un conocimiento directamente basado en la experiencia no puede existir, ya que siempre está presente algún tipo o grado de interpretación. La laxitud con que se usa el término positivismo, además, hace que se le quiera aplicar igual a los post-positivistas, comenzando por Popper, aunque desde 1934 este autor se veía a sí mismo como el principal crítico del neopositivismo. En su autobiografía intelectual, Popper dice que se ha llegado a dar como un hecho que el positivismo está muerto, pero que nadie se pregunta quien lo mató, y añade: "creo deber asumir tal responsabilidad" (*I fear that I must admit responsibility*). (1976: 88)

Otra cita no deja lugar a dudas en cuanto a que Popper dice expresamente que en todo conocimiento humano la interpretación está presente de manera fundamental y por ello es absurdo ubicarlo en los rangos del positivismo o el neopositivismo:

> Locke, Berkeley, e incluso el escéptico Hume y sus numerosos sucesores, en especial Russell y Moore, compartían con Descartes la idea de que las experiencias subjetivas son particularmente seguras y son por ello un punto de partida sólido [...] Frente a esto yo sugiero que no hay nada de directo o de inmediato en nuestra experiencia [...] Todo es decodificación o interpretación. Nosotros aprendemos tan bien a decodificar que todo nos parece muy directo o inmediato [...] De todas maneras, algunas veces hacemos errores al decodificar [...] y no hay certeza absoluta, aunque hay certeza suficiente para la mayor parte de los fines prácticos. Hay que abandonar la búsqueda de certeza, de un fundamento seguro del conocimiento. (Popper, 1978: 46-47)

Lo que distingue mejor la postura de Popper, y el post-positivismo, del positivismo y el neopositivismo, se refiere al fundamento del conocimiento. En una obra sobre epistemología e investigación educativa, Phillips y Burbules plantean que la postura de Descartes y otros pensadores racionalistas, al igual que la de Locke, Hume y otros empiristas, tenían en común la búsqueda de ese fundamento seguro, por lo que denominan fundacionalistas a unos y otros.

Las corrientes epistemológicas surgidas en la segunda mitad del siglo XX tienen en común el abandonar esa pretensión de identificar un fundamento indiscutible para nuestro conocimiento, por lo que las denominan *no fundacionalistas*. (Phillips y Burbules, 2000: 5)

La continuación del párrafo de Popper citado antes muestra que su postura debe situarse dentro de este grupo de autores *no fundacionalistas*:

> Hay que abandonar la búsqueda de certeza, de un fundamento seguro del conocimiento. Yo veo el problema del conocimiento de una manera diferente de la de mis predecesores. La seguridad y la justificación de las pretensiones al conocimiento no me interesan. Por el contrario, mi problema es el crecimiento del conocimiento. ¿En qué sentido podemos hablar de un crecimiento o de un progreso del conocimiento y cómo podemos realizarlo? (Popper, 1978: 47)

Los post-positivismos y las posturas interpretativas actuales

El Apéndice de esta Conclusión presenta una visión histórica de las ideas sobre el conocimiento, que permite apreciar que, desde la época griega hasta la baja Edad Media, prevalecieron distintas versiones del idealismo o racionalismo de Platón, en tanto que las posturas empiristas tenían una presencia muy débil. A partir de los siglos XVI y XVII, en particular con el desarrollo de la ciencia moderna, esto cambió con una presencia cada vez mayor de las ideas empiristas.

Se podrá ver que Kant propuso una tercera postura, que ha marcado las visiones posteriores, en la que el conocimiento es una síntesis de pensamiento y experiencia, ya que las percepciones del mundo que nos llegan a través de los sentidos se captan inevitablemente con ciertas categorías de nuestro aparato cognitivo.

Se verá que las posturas actuales, como se ha apuntado ya, no pretenden encontrar un sustento totalmente seguro de los conocimientos, como hacían racionalismos y empirismos anteriores, que se consideran *fundacionistas*, mientras las posturas actuales tienen en común ser *no fundacionistas*.

Una postura más, el escepticismo, que también se remonta a la época griega, está representada hoy por las variadas corrientes que se engloban bajo la etiqueta de posmodernas, como los constructivismos radicales o el anarquismo metodológico.

Sin llegar a estos extremos, las posturas del rubro de interpretativas, críticas, hermenéuticas y similares, destacan la omnipresencia de la interpretación, lo que lleva necesariamente a subrayar también la imposibilidad de que cualquier punto de vista particular agote la realidad y pretenda ser poseedor de la verdad absoluta.

Es importante añadir, sin embargo, que estas posturas no deben identificarse con un rechazo de elementos asociados con el positivismo y posturas que se relacionan con él, en particular el planteamiento de que las ciencias sociales y humanas (las ciencias del espíritu, las *geisteswissenschaften* de la tradición historicista de Dilthey y otros pensadores alemanes de fines del siglo XIX) implicarían una metodología totalmente distinta de la que sería adecuada para las ciencias de la naturaleza.

En este sentido puede citarse al iniciador de la neo-hermenéutica, Hans G. Gadamer, que la ve no como sustituto de la *metodología científica*, sino como complemento indispensable para que las ciencias no rebasen sus propios límites, pretendiendo ser el elemento clave de la vida social, sustituyendo decisiones políticas y valorales:

> Fue desde luego un tosco malentendido el que se acusara al lema "verdad y método" de ignorar el rigor metodológico de la ciencia moderna. Lo que da vigencia a la hermenéutica es algo muy distinto y que no plantea la menor tensión con el ethos más estricto de la ciencia [...] la misma ciencia no podrá ejercer adecuadamente su función social más que si no se oculta a si misma sus propios límites [...] Por suerte, puede existir un acuerdo objetivo tanto en el hecho de que sólo existe una única "lógica de investigación", como también en que ésta no lo es todo [...] con todas las diferencias que puedan existir entre las ciencias naturales y las del espíritu, en realidad la vigencia inmanente de la metodología crítica de las ciencias no es discutible en ningún sentido. Pero ni el racionalista crítico más extremo podrá negar que a la aplicación de la metodología científica le preceden una serie de factores determinantes que tienen que ver con la relevancia de su selección de temas y sus planteamientos [...] Como se ve, no es sólo el papel de la hermenéutica en las ciencias lo que está aquí en cuestión, sino toda la auto-comprensión del hombre en la moderna era de la ciencia. (Gadamer, 1977: 641-647)

Ladriere reconoce que es imposible tener un criterio simple e indiscutible de verdad, y subraya lo que puede reconciliar a la hermenéutica con el post-positivismo. Señala que Popper dice que el concepto de verdad se puede entender en relación con la realidad, y que sería deseable poder comparar directamente con ella conceptos y juicios para decidir sobre su valor de verdad, pero que esto no es posible:

> Como Popper mostró en 1934, esta manera de comprender el empirismo encierra una concepción demasiado ingenua y, en definitiva, inaceptable del lenguaje [...] no hay forma de comparar de manera directa el lenguaje con la experiencia [...] todo lenguaje, incluso el lenguaje descriptivo más sencillo, es una interpretación de la experiencia, no una representación adecuada de su estructura. Las proposiciones que formulamos deben ser sometidas a verificación, y es en definitiva hacia la experiencia donde habrá que dirigirse para saber si una proposición es aceptable o no. Es aquí donde reside la parte de verdad del empirismo. Pero no existe un lenguaje observacional puro [...] todo lenguaje es interpretativo [...] Nosotros siempre estamos ya en la teoría. Hay niveles más o menos elevados de teorización, pero la lectura del mundo no reposa sobre una base de evidencias privilegiadas; es una creación imaginativa, es siempre trascendente con respecto al dato, a lo dado, y sólo puede apuntar a lo que se siente en la experiencia por un inmenso rodeo, a través de construcciones cada vez más abstractas, por un esfuerzo ininterrumpido del logos cuya concordancia con la physis será siempre fragmentaria,

incierta, de carácter conjetural y perfectamente vacilante. Todo lenguaje es una hermenéutica. Tratar de decir el mundo es hacerlo venir a la temblorosa claridad de la palabra. Y toda palabra es productora. El hombre habita este mundo como poeta, incluso y sobre todo cuando es un matemático. Pero [...] ¿cuál es pues el sentido que se anuncia así, en medio del lenguaje? ¿Cómo puede llegar un significado a los conceptos? ¿Qué decimos cuando hablamos del mundo, de nosotros mismos o de Dios? ¿Qué es pues la interpretación? (Ladriere, 1971: 131)

Para terminar

He propuesto entender la metodología como un esfuerzo permanente por cuidar la solidez de las inferencias presentes en todos los pasos de una investigación. Las disciplinas cognitivas muestran que todo conocimiento (y no sólo el de la interioridad humana) implica interpretación, reconociendo que hay varios niveles:

TABLA C.1. NIVELES DE CONOCIMIENTO Y DE INTERPRETACIÓN

Niveles de conocimiento	Niveles de interpretación
Descriptivo estático: parte de percepción, pero implica paso a ideas abstractas, con procesos de conceptualizar, clasificar, comparar y medir, construyendo tipos y categorías, tipologías y taxonomías. Descriptivo dinámico: implican identificar y construir regularidades, tendencias y patrones, a partir de ciertas percepciones, organizadas de alguna forma en el tiempo, haciendo predicciones y retrospecciones.	De sensaciones en caso del conocimiento descriptivo estático, y de tendencias en el dinámico. El hombre, *animal buscador de patrones (pattern seeking animal),* da coherencia a multiplicidad de estímulos que llegan por los sentidos, ordenándolos en un tipo de interpretación para el que parecemos genéticamente programados, para la supervivencia de individuos y especie.
Explicativo de causalidad física: más allá de la descripción, indaga el por qué de las cosas; busca y construyen esquemas causales simples-complejos, estos últimos con una noción de causalidad múltiple, interactiva y no lineal. Distinción entre correlación y causalidad; cuidado por eliminar aparentes causas que no lo son, las causas espurias.	De causalidad física, en conocimiento explicativo, simple o complejo: noción de causa en Kant es categoría básica, noción primera; de allí la dificultad de definirla, pero también su presencia inevitable. El hombre parece también programado genéticamente para inferir las causas de lo que ocurre. También en esto somos *pattern seeking animals.*
Explicativo de causalidad intencional: se refiere a fines, objetivos, propósitos, metas de los actores, e implica la identificación o atribución de intenciones.	De fines y propósitos; propia de seres con interioridad. Lo que se interpreta son, a su vez, interpretaciones; se trata de terreno *metainterpretativo*.

FUENTE: ELABORACIÓN PROPIA.

En todos los niveles, el conocimiento humano implica un componente perceptual y otro estructurante; en todos hay una parte que no está en el objeto conocido (sea la naturaleza o la sociedad) sino, de alguna manera, en el sujeto que conoce. Esto es lo que denota la expresión *omnipresencia de la interpretación:*

Nuestro aparato cognitivo no es una cámara fotográfica, al estilo de un empirismo primitivo. Es algo más complejo, que la biología moderna explica como resultado de millones de años de evolución; así surgieron exquisitos mecanismos que pueden captar unos cuantos fotones o tenues vibraciones de la frecuencia apropiada; también se desarrollaron sistemas, más complejos aún, de neuronas y estructuras cerebrales que procesan las sensaciones para formar conceptos abstractos, detectar y construir patrones y relaciones causales, inferir o atribuir sentido.

No es esperable que todo investigador tenga formación avanzada en epistemología; de hecho, la gran mayoría no la tiene. Lo importante es evitar confusiones que obscurecen innecesariamente la tarea de investigar, en sí misma tan compleja.

Pero, aunque no tenga un conocimiento detallado de corrientes epistemológicas, si un investigador concibe su tarea como búsqueda diaria de conocer cada vez mejor la parcela de la realidad que le interesa, aceptando que nunca lo conseguirá del todo, estará del lado de las posturas que hoy se consideran más sólidas, frente a los extremos del empirismo positivista y el irracionalismo postmoderno.

MISLEVY Y EL SEÑOR JOURDAN

Un gran especialista en medición y evaluación educativa entiende su trabajo en forma similar a la que propongo. No encuentro mejor colofón que ideas de un texto de Robert Mislevy (2014), que identifica tres posturas epistemológicas en forma que coincide con las aquí expuestas, con una terminología diferente:

Modernismo. El mundo externo está allí, independiente de nosotros; nuestro conocimiento puede captar objetivamente su estructura y propiedades.
Postmodernismo. Nuestro conocimiento y lenguaje no captan el mundo; lo construyen de diversas formas, y no hay manera de sostener la superioridad de una u otra.
Postmodernismo neo-pragmático. No podemos conocer plenamente la verdad sobre el mundo, pero solo tenemos nuestro limitado conocimiento, por lo que debemos actuar con base en él, tratando de mejorarlo todo lo que podamos.
Comparándose irónicamente con el personaje de Moliere, el señor Jourdan, que descubrió con sorpresa que toda la vida había hablado en prosa, Mislevy dice:

- Durante años he estado simplemente haciendo mi trabajo: tratando de mejorar las evaluaciones educativas aplicando ideas de la estadística y la psicología. Y ahora descubro que he estado defendiendo una teoría de las pruebas postmoderna neo-pragmática, sin haberlo intentado nunca.

Captar perfectamente el mundo está fuera de nuestro alcance, pero los conocimientos no son iguales. Una distancia abismal separa las visiones mágicas y míticas que prevalecieron por milenios, o las tonterías de tanto charlatán de nuestros días, de las conclusiones de indagaciones serias. El trabajo de los que hemos querido dedicarnos a la investigación es, simplemente, procurar que nuestros esfuerzos sean cada vez más rigurosos.
Siguiendo una vieja costumbre, termino *yéndome hasta la Biblia,* en este caso hasta los inevitables griegos con el filósofo presocrático Xenófanes, de quien es una cita que gustaba a Popper, quien la tradujo así (Magee, 1974: 37):

- Los dioses no nos revelaron todas las cosas desde el principio, pero en el transcurso del tiempo, a través de la búsqueda, podemos aprender y conocer las cosas mejor. Pero por lo que se refiere a la verdad segura, ningún hombre la ha conocido ni la conocerá, ni sobre los dioses ni sobre todas las cosas de las que yo hablo. Pues ni aunque, por casualidad, llegara a pronunciar la verdad final, se enteraría él mismo: pues todo es una red de conjeturas entretejidas.

Apéndice. Visión histórica de corrientes epistemológicas

La reflexión filosófica sobre el conocimiento es tan antigua como la filosofía, pero no era tan importante como hoy, sino más bien parte de la ontología. Hacia el siglo XVI, y sobre todo a partir del XVII, con la ciencia moderna, se incrementó el interés por el conocimiento en general, y sobre todo por ese tipo particular de conocimiento que es la ciencia. Las obras filosóficas en español distinguen *gnoseología* y *epistemología*, de manera que la primera se refiere a filosofía del conocimiento en general, y la segunda a filosofía de la ciencia. En inglés epistemology designa a las dos. En español se usan las expresiones filosofía o teoría del conocimiento y filosofía o teoría de la ciencia, pero *filosofía de la ciencia* incluye cuestiones sobre el papel de la ciencia en la sociedad, aspectos éticos y otros, en tanto que *epistemología* se refiere solo a la ciencia en cuanto forma de conocimiento.

De la época griega a la baja edad media (s. IV aC — s. XIV dC)

Dos grandes formas de entender el conocimiento en general (gnoseología):

- *Racionalismo*: el conocimiento es producto del pensamiento; en el límite no hay un mundo externo (*idealismo*); el pensamiento crea el mundo.
- *Empirismo*: el conocimiento es producto de la experiencia, de lo que captan los sentidos del mundo objetivo que existe fuera del sujeto cognoscente.

Postura racionalista. Predominó durante todo este lapso a partir de Platón (427-347 aC), que distinguía el mundo inteligible y el sensible (*kosmos noetós-tópos oratós*). Como ilustra su alegoría de la caverna, el conocimiento que dan los sentidos es *engañoso*; el sólido consiste en recordar (memoria, *anámnesis*) las ideas del mundo inteligible. Las sensaciones despiertan el recuerdo de las ideas.

Aristóteles (384-323 aC) daba un papel mayor como fuente de conocimiento a las sensaciones (*lo primero que conocemos es por medio de la experiencia, epagogé*), pero en lo esencial seguía a Platón, su maestro: debe intervenir el entendimiento agente (*nous poietikós*) para sacar la esencia, la especie inteligible (*eidos epistetón*) de las especies sensibles (*eidos aistetón*), las sensaciones (*fantásmata*).

Estas ideas siguieron con Plotino (204-269 dC) y San Agustín (354-430: *no busques fuera*; *vuelve a ti mismo; en lo interior del hombre habita la verdad*). Así fue hasta la escolástica, con Anselmo de Canterbury (1033-1109). Tomás de Aquino (1224-1274) retomó de Aristóteles el papel de la experiencia, pero descartó empirismo: la experiencia

debe ser trascendida al intervenir el *entendimiento agente* para llegar a los universales a partir de los fantasmas de la experiencia sensible.

Postura empirista. Estuvo presente en la época griega con el sensismo de *cínicos estóicos y epicúreos*, pero fue secundaria. En la escolástica se encuentra el primer planteamiento del papel de experiencia como verdadera causa de conocimiento, con el nominalismo de Guillermo de Ockam (ca 1300-1349)

Postura escéptica. Considera imposible llegar a verdades universales objetivas. Es el caso de los sofistas, como Protágoras (ca 481-411 aC, *como cada cosa me aparece, así es para mí, y como aparece a ti, es para ti*), y otros pensadores de la llamada Academia Media y de la Academia Nueva.

Gnoseología y epistemología en los siglos XVI a XVIII

El rasgo fundamental de la época fue el surgimiento de la ciencia moderna, primero con la astronomía y la física (Galileo, 1564-1642; Kepler, 1571-1630; Newton, 1643-1729), después con la química (Boyle, Lavoisier) y más tarde las demás.

Las dos grandes corrientes gnoseológicas anteriores siguieron presentes, con expresiones distintas, desde luego, pero la importancia del empirismo aumentó, en tanto que disminuía la del idealismo.

Racionalismo

El representante más destacado fue René Descartes (1596-1650; en 1637 su *Discurso del Método*). Al buscar un fundamento seguro para el conocimiento identifica una primera idea de la que no puede dudar: que existe, puesto que duda, piensa (*cogito ergo sum*). Continúa tradición platónico-agustiniana, incluyendo postular ideas innatas claras-distintas y por ello verdaderas (substancia infinita -Dios- y finitas —*res cogitans, res extensa*— *vs* ideas adventicias y que nosotros mismos formamos, siempre oscuras y dudosas).

Otros autores de esta corriente fueron Baruch Spinoza (1632-1677, que distinguía ideas falsas, a partir de sensaciones, y verdaderas, fruto del entendimiento, que no es raciocinio sino intuición *sub specie aeterni*) y Wilhelm Leibniz (1646-1716, que en sus *Nuevos ensayos sobre el entendimiento humano* de 1704, polemizando con Locke, afirma que el alma no es *tabula rasa*; por el contrario, tiene ideas innatas en las que se basa todo saber auténtico. Evocando a Platón, Leibniz dice que los sentidos son necesarios, pero *proporcionan más error que verdad; el espíritu se libera de la materia en el puro conocimiento de las verdades eternas y alcanza por ello su perfección. Hay en*

nuestro espíritu ideas innatas que representan las esencias de las cosas. Nuestro conocer, por lo tanto, es un recordar.

Empirismo

Francis Bacon (1561-1626; obras *Instauratio Magna* y *Novum Organum* en 1620). El conocimiento se ve afectado si, en lugar de atenerse a la experiencia, se hace caso a la tradición (*idola theatri*); a lo que dice la mayoría de la gente (*idola fori*); a las ideas personales favoritas (*idola specus*); o los prejuicios genéricos (*idola tribus*).

Thomas Hobbes (1588-1679). En *De corpore* (1655) plantea que la experiencia es la fuente principal del conocimiento. Contra Descartes adopta postura materialista, eliminando la *res cogitans* y reduciendo la realidad a únicamente la *res extensa.*

John Locke (1632-1704, autor clave del empirismo inglés, con su *Essay concerning human understanding* de 1690). No hay ideas innatas; la mente es *tabula rasa* y todo lo que contiene llega por la experiencia externa (sensación) o la interna (reflexión).

George Berkeley (1685-1753). *Treatise concerning principles of human knowledge* de 1709 lleva a extremo postura de Locke identificando al ser con el ser percibido.

David Hume (1711-1776; *An enquiry concerning human understanding*, 1748). Con Locke dice que no hay ideas innatas, y que todo contenido de la conciencia viene de la experiencia sensible. Añade que para explicar conceptos complejos hay que acudir a la asociación.

La tercera postura: Kant

En la *Crítica de la Razón Pura* (1781), Immanuel Kant (1724-1804) propuso una tercera postura sobre la naturaleza del conocimiento, como síntesis de percepción y categorías mentales. Integrando racionalismo y empirismo, Kant planteó la postura que está en la base de las ideas actuales: el *idealismo o racionalismo crítico.*

Kant distinguía la materia del conocimiento (la experiencia sensible) y la forma (las categorías universales-necesarias). Planteaba también la distinción entre *noúmeno y fenómeno.* El idealismo crítico es diferente de idealismo en su sentido más fuerte: hay mundo externo, pero no lo podemos conocer como es, pues las sensaciones que llegan por los sentidos no lo reflejan tal cual, sino modificadas, captadas según dos *formas a priori* (espacio y tiempo) y 12 *categorías* (vs 10 de Aristóteles):

De cantidad: Unidad-Pluralidad-Totalidad.
De cualidad: Realidad-Negación-Limitación.

De relación: Substancia-accidente; Causa-dependencia; Acción-pasión.

De modalidad: Posible-imposible; Existencia-inexistencia; Necesario-contingente

Después de Kant: posturas fundacionistas s. XIX y XX

Durante el s. XIX y la primera mitad del XX, varios pensadores sostuvieron nuevas versiones de las dos corrientes clásicas. Las que se incluyen en este inciso tienen un rasgo que comparten con sus antecesores de los siglos XVI a XVIII: la búsqueda de un fundamento completamente sólido del conocimiento. Por esto actualmente se les denomina posturas *fundacionalistas o fundacionistas.*

Racionalismo. Pueden identificarse variantes:

Idealismo post-kantiano Johann GottliebFichte (1762-1814); Friedrich Schelling (1775-1854); G. Wilhelm F. Hegel (1770-1831); en una variante de la dialéctica de Hegel debe mencionarse a Karl Marx (1818-1883) y el materialismo dialéctico.

Hermenéutica original literaria, jurídica y bíblica. *Subtilitas intelligendi* (sentido textual, semántica); *subtilitas explicandi* (sentido intertextual, sintáctica); *subtilitas applicandi* (sentido contextual, pragmática).

Historicismo. Wilhelm Dilthey (1833-1911). *Introducción a las ciencias del espíritu* (1883); *El nacimiento de la hermenéutica* (1900). Causa mecánica vs causa final. Explicar (*Erklären*) vs comprender (*Verstehen*).

Fenomenología. Edmund Husserl (1859-1938). La intuición de la esencia. La *epojé*, puesta entre paréntesis de los datos de la experiencia.

Escuela de Frankfurt. Teoría Crítica. Combina elementos del marxismo con el psicoanálisis Theodor Adorno (1903-1969). Dialéctica como perspectiva de la totalidad vs lógica formal. M. Horkheimer (1895-1973); H. Marcuse (1898-1979).

Empirismo

Positivismo de Auguste Comte (1798-1857) distingue estados mitológico-teológico, metafísico y positivo. Rechazo teórico de la metafísica. En la práctica concebía la sociología (ciencia reina) en forma idealista.

Otros positivismos franceses. Convencionalismo de Henri Poincaré (1854-1912); Pierre Duhem (1861-1916); E. Le Roy (1870-1954).

Empiriocriticismo alemán. Ernst Mach (1838-1916). Richard Avenarius (1843-1896).

Experimentalismo francés de Claude Bernard (1813-1878, medicina científica), así como empirismo inglés de John Stuart Mill (1806-1873, *A system of logic* de 1843), con su

propuesta de diseño experimental; métodos de coincidencias, diferencias, combinación de ambos, residuos y mutaciones paralelas.

Filosofía analítica-lógica matemática. Gottlob Frege (1848-1925); Bertrand Russell (1872-1970). Ludwig Witgenstein (1889-1951).

Neopositivismo lógico. Círculo de Viena. Moritz Schlick (1882-1936), Rudolph Carnap (1891-1970); Karl Hempel (1905-1997).

Pragmatismo. Ch. Pierce (1839-1914); W. James (1842-1910); John Dewey (1859-1952); Operacionalismo. Percy Williams Bridgam (1882-1961).

Posturas epistemológicas en la actualidad

Desde la primera mitad del XX, pero especialmente en la segunda, pensadores de las dos corrientes epistemológicas coincidieron en que el conocimiento perfecto de la realidad no está al alcance de los humanos. Reconocen que todo conocimiento implica interpretación, inferencia, y siempre es susceptible de error, lo que no quiere decir que todos los conocimientos sean iguales, ya que los hay de distinta solidez. Al abandonar la búsqueda de un fundamento perfecto para el conocimiento, estas posturas comparten la designación de no *fundacionalistas* o *no fundacionistas*.

Racionalismo no fundacionista:

Neo-hermenéutica. Hans G. Gadamer (1900-2002); Emilio Betti (1890-1968); Karl Otto Apel (1922-); Maurice Merleau Ponty (1908-1961); Paul Ricoeur (1913-2005); Jürgen Habermas (1929-).

Empirismo no fundacionista:

Post-positivismo. Karl Popper (1902-1994, *Lógica de la Investigación Científica* de 1934, primera crítica radical del neopositivismo). Thomas Kunh (1922-1996, *Estructura de las revoluciones científicas*, 1962). Imre Lakatos (1923-1974).

El irracionalismo posmoderno

La postura escéptica, que considera imposible llegar a conocimiento sólido alguno, se manifiesta en autores a los que se aplica el calificativo de postmodernos, como Jean Lyotard (1924-1998); Jean Baudrillard (1929-2007) Michel Foucault (1926-1984); Jacques Derrida (1930-2004, deconstruccionismo). Paul Feyerabend (1924-994, anarquismo metodológico); o Paul Watzlawick (1921-2007) y Ernst von Glazersfeld (1917-2010), con el constructivismo radical.

Referencias

Blalock Hubert, M. Jr. (1989). The real and unrealized contributions of quantitative Sociology. *American Sociological Review*, vol. 54 (junio), pp. 447-460.

Blaug, M. (1980). Lo que Ud. siempre quiso saber y nunca se atrevió a preguntar sobre la filosofía de la ciencia [pp. 17-72]. En *La metodología de la economía o cómo explican los economistas*. Madrid: Alianza. (Ed. original en inglés, 1980).

Bouveresse, J. (1980) Les positivistes. *Encyclopaedia Universalis*, Vol. 17, pp. 50-63. Paris, Encyclopaedia Universalis.

Gadamer, H. G. (1977). *Verdad y método,* Salamanca: Sígueme, pp. 641-647. (1a edición en alemán 1960).

Ladriere, J. (1971). Langage scientifique et langage spéculatif. *Revue Philosophique de Louvain*, Tome 69, N° 1, pp. 92-132 y N° 2, pp. 250-282.

Magee, B. (1974). *Popper*. Barcelona: Grijalbo. (Ed. original 1973).

Martínez Rizo, F. (1986). Teoría de la ciencia y teoría del método. Estado de la discusión en ciencias del hombre. Voz *Universitaria*. Año VIII (28): 5-24.

Martínez Rizo, F. (1991). The Controversy about Quantification in Social Research. *Educational Researcher,* Vol. 20 (9): 9-12.

Martínez Rizo, F. (1993). La polémica sobre la cuantificación en el campo de las ciencias del hombre. *Papers*, Nº 42, pp. 13-34.

Martínez Rizo, F. (1997). La metodología de la investigación y los límites del conocimiento humano. En *Caleidoscopio* (Aguascalientes, UAA) Vol. 1, Nº 1, 1997, pp. 95-111.

Martínez Rizo, F. (1999). ¿Es posible una formación sistemática para la investigación educativa? Algunas reflexiones. *Revista Electrónica de Investigación Educativa.* Vol. 1, No. 1, pp. 1-6.

Martínez Rizo, F. (2000). Exigencias de calidad de instrumentos de evaluación. Comparación de las tradiciones llamadas cuantitativa y cualitativa. *Caleidoscopio*. N° 7 (enero-junio) pp. 49-57.

Martínez Rizo, F. (2002). Las disputas entre paradigmas en la investigación educativa. *Revista Española de Pedagogía*. Año LX, N° 221, pp. 27-49.

Mislevy, R. J. (2014). Postmodern Test Theory. *Teachers College Record,* Vol. 116 (11): 1-24.

Onwuegbuzie, A. J. y Wisdom, J. P. (2014). Qualitative versus Quantitative Methods and Beyond. En Phillips, D. C. (Ed.). *Encyclopedia of Educational Theory and Philosophy.* [Vol. Two, pp. 677-681]. Los Angeles: Sage Reference.

Phillips, D. C. (2014). Postpositivism. En Phillips, D. C. (Ed.). *Encyclopedia of Educational Theory and Philosophy*. [Vol. Two, pp. 645-649]. Los Angeles: Sage Reference.

Phillips, D. C. y Burbules, N. C. (2000). *Postpositivism & Educational Research*. Lanham: Rowman & Littlefield Publ.

Popper K. R. (1976). *Unended Quest. An intellectual autobioqraphy*, London: Fontana.

Popper K. R. (1978). *La connaissance objective*. Paris: Complexe.

Weiss, E. (2017). Hermenéutica y descripción densa vs teoría fundamentada. *Revista Mexicana de Investigación Educativa*, Vol. 22 (73): 637-654.

La versión digital del libro *El nuevo oficio del investigador educativo*, de la Serie Investigación Educativa, se terminó de editar en diciembre de 2020.

Para su composición se usaron las tipografías Garamond Premier Pro, ITC Franklin Gothic y Myriad, en sus versiones Pro.

El libro pretende contribuir a subsanar una debilidad de muchos investigadores: su formación metodológica. A partir de una visión que integra diseño, observación y análisis, a partir de preguntas investigables, se aborda el inicio de un proyecto con la construcción del objeto de estudio basada en la revisión de la literatura, y que culmina en la formulación de preguntas precisas o hipótesis. Sigue luego una gama de diseños de investigación, después un conjunto de técnicas para obtención de información empírica y, finalmente, buen número de técnicas de análisis, presentando en detalle las más básicas, con menor amplitud las de complejidad media, y muy brevemente las avanzadas. Se refiere una extensa bibliografía para profundizar en cualquier tema. Con base en su experiencia, en la conclusión el autor hace recomendaciones a quienes se dedican a formar investigadores, subrayando la importancia de buscar siempre el rigor, sin perderse en vanas disputas epistemológicas, con la esperanza de superar la dicotomía empobrecedora que opone el enfoque cualitativo al cuantitativo.

ISBN: 978-607-7923-29-9
9 786077 923299

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

COMIE
Consejo Mexicano de Investigación Educativa A.C.