CONSTRUCCIÓN DE UNA PRUEBA PARA EVALUAR APRENDIZAJES EN EDUCACIÓN SUPERIOR

EMILIA CRISTINA GONZÁLEZ MACHADO
ERIKA PAOLA REYES PIÑUELAS
MÓNICA LÓPEZ ORTEGA
UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

TEMÁTICA GENERAL: APRENDIZAJE Y DESARROLLO HUMANO

RESUMEN

En el presente trabajo se describe el diseño de una prueba alineada al currículo, construida con el propósito de evaluar los aprendizajes logrados en el curso introducción al pensamiento científico en el área de ciencias sociales. La unidad de análisis en el estudio corresponde a estudiantes de recién ingreso a un tronco común en educación superior. La prueba final está constituida por 33 reactivos de opción múltiple con tres alternativas de respuesta (K=33), la cual se aplicó a 173 examinados (n=173). Entre los resultados se muestran los análisis de discriminación y dificultad de los ítems, basados en la Teoría Clásica del Ítem, los cuales fueron aceptables. Entre las discusiones finales se plantean estrategias relacionadas con la muestra y mejorar el índice KR20.

Palabras clave: Evaluación del aprendizaje, Currículo, Educación superior.

Introducción

Este estudio reporta resultados y hallazgos relativos al diseño y propiedades psicométricas de confiabilidad y validez de un examen colegiado del aprendizaje de la asignatura de *Introducción al Pensamiento Científico* (IPC) de la etapa básica del área de las Ciencias Sociales en la Universidad Autónoma de Baja California (UABC). Se circunscribe dentro de la evaluación educativa, la cual ha sido promovida por los gobiernos a partir de las recomendaciones de organismos internacionales, destacando la planeación y el desarrollo institucional, perfil y desempeño del profesorado, programas de estímulos, productividad de los investigadores, calidad de los programas educativos, competencias de egreso, programas de posgrado, pruebas de gran escala. Con ello, las Instituciones de Educación

Superior (IES) han abierto procesos evaluativos y reorientado sus estructuras organizativas y programas de desarrollo institucional para cumplir con los indicadores de gremios evaluadores.

Al respecto de la evaluación educativa, Scriven (1991) la plantea como una disciplina que busca generar una serie de inferencias certeras y valiosas, en la que se determina el mérito o valor de las cosas, que conlleva un proceso analítico reflexivo de las disciplinas. En el cual se recolecta información o datos, y además se detectan los elementos relevantes, que, si bien en el ámbito educativo el resultado de una evaluación no implica una mejora en la calidad educativa, sí ofrece una dirección para su logro. Así, en virtud que los estudiantes son una de las razones de ser de las universidades, desarrollar instrumentos que permitan identificar los conocimientos adquiridos en las asignaturas que cursan, brinda elementos para su reflexión, análisis e implementación de estrategias para la mejora.

Este trabajo descansa en lo que se denomina una evaluación interna. Para Escudero (1997) este tipo de evaluación pone en relieve la capacidad de la institución para autotransformarse y resolver sus propias dificultades, y sugiere complementarse de la evaluación externa que facilita la reflexión y modula las interpretaciones realizadas desde el interior. De manera general se abona a la documentación de estudios que la utilizan, y particularmente en este caso al área de las Ciencias Sociales en educación superior. Con ello se pretende realizar una descripción del proceso y probar su solidez dentro de un contexto particular.

En este marco se apoya la presente investigación y se justifica por alcances diversos. Entre los que destaca, el diseño y desarrollo del instrumento a partir de modelo para desarrollar pruebas de gran escala de referencia criterial alineadas con el currículum propuesta por Contreras (2011), este modelo nace entre otras razones de la necesidad de contar con metodologías rigurosas para el diseño, validación y aplicación de pruebas. Con ello, este instrumento supera las críticas en relación al desarrollo de pruebas donde las técnicas y procedimientos normalmente utilizados en las investigaciones para reportar las propiedades psicométricas no son suficientes para abordar la variedad y complejidad de la evaluación (Borsboom, Mellenbergh & Heerden, 2004).

Hay que mencionar, además, en el desarrollo de los instrumentos una de las observaciones por expertos es la limitada participación de los sujetos que se encuentran directa o indirectamente inmersos en el proceso enseñanza-aprendizaje (Berk, 2006). En en este caso, se incorpora la participación de especialistas en el desarrollo de instrumentos y en el constructo a evaluar, así como de docentes en ejercicio.

Finalmente, el presente examen representa uno de los productos derivados del interés de la actual gestión rectoral 2015-2019 (UABC, 2015), que busca fortalecer la habilitación del personal docente con la capacitación en materia de evaluación colegiada del aprendizaje y elaboración de exámenes colegiados del aprendizaje y al mismo tiempo evaluar el grado de apropiación de las

competencias de la asignatura en los estudiantes y la identificación de áreas de oportunidad para la implementación de un plan de acción para una asignatura con altos niveles de reprobación.

Desarrollo

El objetivo del presente estudio es desarrollar un examen con criterios de calidad técnica, alineado al currículo del curso *IPC* que se imparte en el tronco común de Ciencias Sociales en la Facultad de Ciencias Humanas de la UABC.

Como todo instrumento de medición, este examen pretende evaluar el conocimiento a partir de un proceso estandarizado de puntuaciones, cuidando los procesos de calidad que aseguren se trate de una prueba válida y confiable. En este sentido, es importante recordar que los *Estándares para la evaluación educativa y psicológica* (American Educational Research Association, 2014) establecen que todo instrumento, independientemente de su estructura o propósito, debe cumplir con determinados criterios para ser considerado un instrumento apropiado para la medición. Ahora, los *Estándares* también hacen una distinción entre *tests* y *evaluación*, siendo éste último un término más amplio el cual normalmente integra la información derivada de los tests y la de otras fuentes, como por ejemplo resultados de otros tests, entrevistas o historial académico o psicológico.

La elaboración de tests es un proceso que permite medir un aspecto particular de un individuo a partir de una serie de preguntas o demostración de habilidades, según sean los propósitos del instrumento. Dicho proceso comienza con la consideración de los usos que se espera asignar a los resultados o puntajes. Son el contenido y el formato los elementos que proporcionan evidencia que da soporte a estas interpretaciones.

En el caso particular de las evaluaciones educativas, y en concreto de la presente, su fin es conocer la eficiencia de los estudiantes en relación al dominio de conocimientos en la asignatura de IPC. Por lo tanto, el elemento clave en un test educativo es la relación entre el currículo y el contenido del test.

En cuanto a la especificación de contenidos, que es el primer paso en el desarrollo de especificaciones, se deben mencionar de inicio los propósitos del instrumento, así como el constructo o definición del dominio de contenido, los cuales deben estar delimitados en un marco que indique cuáles aspectos en concreto deben ser cubiertos y, por lo tanto, pueden ser medidos por el test. Una vez que las decisiones de contenido han sido tomadas, es necesario establecer el formato del instrumento, es decir, establecer la naturaleza de los reactivos y los formatos de respuesta, los cuales se establecen considerando diferentes aspectos, el propósito del test, su dominio y la plataforma en que se desarrolla.

En cuanto a las especificaciones psicométricas, éstas indican las propiedades estadísticas que se esperan de los reactivos y del instrumento en su conjunto, comúnmente son la capacidad de discriminar entre contenidos similares, el nivel de dificultad y las correlaciones entre ítems, así como la dificultad del test.

Lo propuesto por los Estándares es congruente con las labores de evaluación que actualmente se realizan en México, donde el Instituto Nacional para la Evaluación Educativa (INEE) tiene la tarea de realizar evaluaciones en gran escala del aprendizaje mediante instrumentos diagnósticos, objetivos y estandarizados, los contenidos curriculares de los distintos niveles educativos, susceptibles de ser sometidos a este tipo de evaluación (INEE, 2015). De manera general, los criterios que se emplean para la elaboración de instrumentos de medición, son: (a) la alineación con los referentes de evaluación, (b) la calidad técnica y (c) los usos y consecuencias del este.

Ahora bien, es importante tener en cuenta que una sola prueba no es capaz de fundamentar decisiones de gran alcance pues si bien es cierto que dichas pruebas constituyen un elemento valioso al respecto, también lo es que resulta complejo el desarrollo de pruebas estandarizadas de calidad, válidas, confiables y justas. Para ello es fundamental que las pruebas estén alineadas a los contenidos que se deseen evaluar, comúnmente, se utiliza el currículo y se sienta la base en los estándares de contenido y desempeño.

La recomendación principal es encontrar el equilibrio entre la rendición de cuentas y el mejoramiento de los procesos de enseñanza. Esto es apoyado por autores como Martínez (2003), quien resalta la importancia de no perder de vista que la calidad educativa no solamente consiste en aumentar el nivel promedio del aprendizaje de los estudiantes, se deben considerar además dimensiones como la cobertura educativa, eficiencia terminal, eficiencia en costos y equidad.

Las puntuaciones generadas a partir de los instrumentos de medición deben reunir esencialmente dos características: confiabilidad y validez. La primera hace referencia al grado en el cual las medidas de una prueba o de un procedimiento de medición están libres de error, es decir, el grado de consistencia que las mediciones presentan cuando la prueba es repetida en una población de individuos o grupos (American Educational Research Association [AERA]; American Psychological Association [APA]; National Council on Measurement in Education [NCME] (2014). Por su parte, el concepto de validez se refiere al grado en el cual la evidencia y la teoría apoyan las interpretaciones de las medidas de una prueba de acuerdo con los usos previstos (AERA, APA, NCME, 1999). El proceso de acumulación de evidencias de validez tiene como propósito determinar qué tan apropiadas, significativas y útiles resultan las inferencias derivadas de la aplicación de la prueba en función del uso específico para el cual se diseñó. Aun cuando se identifican diversas fuentes, se plantea como un concepto unitario (Messick, 1995).

Metodología

A través del instrumento se pretendió determinar el nivel de logro de aprendizaje que obtuvieron los estudiantes con respecto a un criterio definido. Para este caso, se delimitó a partir de los contenidos establecidos en el programa de la unidad de aprendizaje (PUA) denominado: IPC; lo que permitió conocer lo que un estudiante aprendió o no, por lo anterior, se consideró relevante definir con claridad lo que se pretendía medir.

La propuesta evaluativa se basó en el modelo de desarrollo de pruebas criteriales alineadas al currículo, la metodología implicó procedimientos específicos para cada una de las dos etapas descritas en la figura 1.

Figura 1. Etapas del método de investigación

Participantes. Para la etapa del diseño de la prueba se contó con el apoyo de dos equipos de trabajo: 1) el equipo externo, cuatro expertos en la metodología, uno en investigación educativa, un especialista psicómetra en evaluación del aprendizaje y; dos más, con experiencia en análisis de datos con sustentos en la teoría clásica de los test (TCT) y la teoría de respuesta al ítem (TRI); y 2) el equipo interno, compuesto por especialistas en el currículo, integrado por once miembros, cinco profesores que imparten el curso, dos que colaboraron en el diseño del curso, un experto en diseño curricular, tres con dominio en instrumentos de evaluación y, una conocedora de la metodología.

En la etapa de identificación de la calidad técnica del instrumento, se procedió con la aplicación de la prueba a 173 estudiantes universitarios, el primer criterio para seleccionar a los sujetos muestrales fue que estuvieran inscritos en el curso. El segundo criterio en la selección de la muestra determinística, fue considerar las dos modalidades de estudio ofertadas del programa en la Facultad de Ciencias Humanas (n=173). La distribución de los estratos (ver tabla 1) se conformó de cinco grupos, a cargo de tres profesores.

Resultados

Etapa 1. Elaboración de la prueba.

Análisis curricular. El primer paso de la metodología consistió en revisar cada contenido conceptual y procedimental plasmado en el PUA. En este proceso, un grupo de cuatro especialistas discutió y estableció un mínimo de acuerdos sobre los contenidos que deben ser facilitados y evaluados. Con el fin de identificar el currículo importante a evaluar se requirió del PUA, los textos y demás material bibliográfico de donde se desprenden los contenidos revisados en el curso. El producto que se diseñó es una retícula, el cual se define como un instrumento de análisis de currículo cuyo propósito es estructurar los conocimientos y habilidades que este prevé, y como herramienta analítica documenta las relaciones entre los componentes requeridos del currículo y los aprendizajes pretendidos (Contreras, 2011).

Al revisar y definir los contenidos, se identificó la función de cada uno de ellos y se mencionó de qué forma se relaciona. Una vez representado el currículo se orientaron los ítems de la prueba. Se identificaron 20 contenidos conceptuales y cinco contenidos procedimentales, distribuidos por unidades temáticas en cuatro momentos: (1) naturaleza del conocimiento científico; (2) nociones básicas de la ciencia; (3) epistemología y nociones de la ciencia; y (4) paradigmas científicos.

Contenido importante a evaluar. Una vez analizado el currículo se identificó y determinó la importancia relativa de los contenidos del PUA, es decir, lo que justifica el contenido importante a evaluar (ver tabla 1). Dicho proceso se llevó a cabo por medio de la validación entre jueces de cada contenido, a partir de los cuales se construyó un índice de relevancia curricular (IRC) basado en los siguientes criterios: contribución al logro de la competencia de la unidad (20%), dosificación (10%), carga horaria (10%), relevancia disciplinaria (20%), proporción de servicios que recibe (20%) y, proporción de servicios que brinda (20%).

Se determinó el IRC cuando un par especialistas valoró a través de una calificación el total de contenidos según los criterios mencionados, posteriormente se procedió con un análisis de varianza en donde se estimó confiabilidad entre jueces mediante el coeficiente de correlación intraclase (CCI), donde se consideraron resultados con valores de r= 0.51 a 0.60 como moderada correlación y, r=0.61 a 0.83 como sustanciales, según el criterio propuesto de Landis y Koch. Los análisis se procesaron en el software libre llamado *vassarstats.*

Tabla 1. *Importancia de los contenidos desde una validación interjueces*.*

Contenido	Logro de la competencia de la unidad (20%)	Dosificación (10%)	Carga horaria (10%)	Relevancia Disciplinaria (20%)	Servicios que recibe 20%	Servicios que proporciona 20%	Índice de Relevancia Curricular
Conocimiento ordinario y científico	0.167	0.083	0.083	0.200	0.022	0.060	**0.615**
Características del conocimiento científico	0.200	0.100	0.100	0.133	0.022	0.120	**0.675**
Ciencia	0.200	0.083	0.083	0.067	0.067	0.200	**0.700**
Evolución de las Ciencias Sociales	0.133	0.100	0.100	0.133	0.067	0.140	**0.673**
El papel de la teoría y del método en la construcción del conocimiento	0.200	0.100	0.100	0.200	0.111	0.160	**0.871**
El proceso de investigación científica	0.200	0.100	0.100	0.067	0.111	0.140	**0.717**
Las revoluciones científicas: cambios en el concepto del mundo	0.100	0.050	0.050	0.200	0.133	0.080	**0.613**

Paradigma Empírico-Positivista	0.200	0.100	0.100	0.200	0.156	0.020	**0.775**
Paradigma Dialéctico-Materialista	0.200	0.100	0.100	0.200	0.156	0.020	**0.775**
Paradigma Comprensivo-Relativista	0.200	0.100	0.100	0.200	0.156	0.020	**0.775**
Nuevos paradigmas	0.100	0.067	0.050	0.200	0.200	0.020	**0.636**

Fuente: Elaboración propia. *Se mencionan solo aquellos con valores superiores a .600 en el IRC

Tabla de especificaciones. Con sustento en los productos diseñados anteriormente se procedió con la redacción de 20 especificaciones para construir el banco de reactivos. En el producto que corresponde a la tabla de especificaciones (ver tabla 2 a manera de ejemplo), se precisó que todo el instrumento contendría ítems de opción múltiple, se estableció con base en su IRC, el número de ítems, el foco evaluativo y el nivel cognitivo taxonómico para cada contenido (Anderson y Kratwohl, 2001).

Tabla 2. *Tabla de especificaciones. Unidad I. Naturaleza del conocimiento científico*.

Contenido	IRC	Nº especificaciones	Nº ítems	Foco del ítem	Nivel taxonómico
1.1 Espíritu y curiosidad científicos	0.29	1	2	El ítem pondrá a prueba la comprensión de la función que desempeña el Espíritu científico en el	Comprender

Contenido	CIR	Nº especificaciones	Nº ítems	Foco del ítem	Nivel taxonómico
				desarrollo de la ciencia Y por otro que comprende la curiosidad científica respecto al desarrollo de la ciencia.	
1.2 Conocimiento ordinario y científico	0.61	1	2	El ítem podrá a prueba la comprensión del concepto de conocimiento ordinario.	Comprender
				Otro ítem podrá prueba la comprensión del concepto de conocimiento científico.	Comprender
1.3 Características del conocimiento científico	0.67	1	1	El ítem podrá a prueba la de las características del	Recordar

Contenido	IR C	Nº especificaciones	Nº ítems	Foco del ítem	Nivel taxonómico
				conocimiento científico.	
1.4 Objetivos y alcances de la ciencia	0.57	1	2	El ítem pondrá a prueba la identificación de los objetivos de la ciencia.	Recordar
				Otro ítem que ponga a prueba la identificación de los alcances de la ciencia.	Recordar
1.5 Pseudociencia	0.39	1	1	El ítem podrá a prueba la comprensión de lo que es la pseudociencia.	Comprender

Nota: Competencia: Distinguir las características del conocimiento científico y no científico, mediante aportaciones teóricas de la ciencia, para su aplicación en el estudio de la realidad social, potenciando su integridad intelectual hacia la formación de su espíritu científico.

Especificaciones de ítem. Se incluyeron datos sobre la justificación del contenido a evaluar, la estrategia evaluativa, las indicaciones para responder al ítem, información, gráfica, tabular o textual a utilizar en el ítem; las dimensiones del conocimiento y del proceso cognitivo; la base del ítem, la especificación de la respuesta correcta y de dos distractores; finalmente se redactaron ítems muestra.

Elaboración de ítem. Con base en las especificaciones, se redactaron versiones de ítems para cada contenido, lo constituyó un banco de reactivos para elaborar distintas pruebas de IPC.

Etapa II. Calidad técnica de la prueba. La primera versión del examen contiene 33 ítems con tres opciones de respuesta, la cual evalúa los contenidos teóricos del PUA. La escala podría variar entre 0 y 33 puntos. La puntuación máxima de la prueba fue de 28 con un porcentaje de 88.4% de aciertos, mientras que la puntuación mínima fue de 8 con un porcentaje de 24.2% de respuestas correctas. La media en la puntuación fue de 17.838, lo que representa un 54.1% de respuesta acertadas. Los análisis se basan en la Teoría Clásica de los Test por medio de la aplicación Microsoft Excel 2010, además se utilizó el Test Analysis Program versión 6.65 (TAP).

Confiabilidad. La consistencia interna de la prueba se realizó a través de la prueba KR20, en donde se obtiene un alpha α=0.57, de acuerdo con la convención el nivel aceptable es de α=0.85, lo que indica que no se alcanza a cumplir este criterio.

Análisis de Dificultad. La dificultad promedio (p) refleja un p= 0.541, lo que muestra que en promedio se contesta el 54% de la prueba de forma correcta. La mayor proporción de ítems representan mediana dificultad para responderlos, se puede observar a partir de los valores de *p* que aquellos que fueron muy fáciles de responder fueron el 11, 20, 21 y 22, mientras los ítems 2 y 5 son los que representaron mayor dificultad como se puede apreciar en la tabla 4.

Análisis de Discriminación. En relación con la capacidad para discriminar entre los que saben y no saben, se observan los valores D y rpbis. El índice de discriminación (D) promedio que se obtuvo fue D= 0.276, y la media del punto biserial (rpbis) =0.252 por lo que se cumple con el criterio el cual debe ser igual o mayor a 0.20, lo que indica que hay poder de discriminar entre los que saben de aquellos que no dominan el currículo.

Los ítems con mayor capacidad de discriminar son el 28, 8 y 10, por su parte aquellos que no cumplen con el criterio son los ítems 4, 12, 17,23 y el 31.

Tabla 4. *Análisis de dificultad y discriminación de los ítems.*

Ítem	p	D	rpbis
1	0.64	0.21	0.19
2	**0.13**	0.18	0.22
3	0.54	0.36	0.31
4	0.59	**0.04**	**0.05**
5	**0.16**	0.17	0.24
6	0.38	0.26	0.20
7	0.42	0.27	0.20
8	0.57	**0.47**	0.37
9	0.51	0.32	0.29
10	0.66	**0.40**	0.40
11	**0.84**	0.24	0.25
12	0.43	**0.17**	**0.16**
13	0.36	0.30	0.25
14	0.52	0.38	0.33
15	0.59	0.28	0.21
16	0.40	0.23	0.20
17	0.37	**-0.04**	**0.04**
18	0.55	0.39	0.24
19	0.40	0.35	0.28
20	**0.84**	0.15	0.21
21	**0.80**	0.24	0.28
22	**0.90**	0.21	0.31
23	0.46	**0.07**	**0.11**
24	0.72	0.35	0.31
25	0.39	0.38	0.35
26	0.67	0.39	0.35
27	0.28	0.21	0.24
28	0.63	**0.55**	0.46
29	0.71	0.37	0.28
30	0.51	0.51	0.41
31	0.76	**0.08**	**0.09**
32	0.55	0.24	0.21

33	0.55	0.33	0.27

Nota: las ponderaciones del valor de p<0.20 y p>0.79 aparecen en negritas.

Para analizar el funcionamiento de los distractores, se observó que tanto los grupos con puntuaciones altas y bajas seleccionaron las dos opciones, sin embargo, aquellos con puntuaciones bajas seleccionarían en mayor medida los distractores. El grupo alto está compuesto por una n=47 con una puntuación mínima de 21 respuesta correctas. El grupo bajo está compuesto por una n=56 con una puntuación máxima de 15 respuestas correctas. A partir del análisis de distractores se logró constatar que todos los distractores fueron elegidos.

Conclusiones

En suma, los resultados indican una consistencia interna por medio de la prueba KR20 con un valor no aceptable (0.57); en el caso del análisis de dificultad, en promedio se obtiene mediana dificultad (p= 0.541); mientras que, en el índice de discriminación todos de los ítems presentan valores aceptables, excepto cinco de estos, que no cumplen con el criterio de corte; en el funcionamiento de los distractores, se destaca que, el grupo con puntuaciones bajas seleccionaron con mayor medida los distractores que el grupo de puntuaciones altas. A partir de lo expuesto, se plantea que los siguientes estudios consideren: un tamaño de muestra mayor y, analizar y en su caso modificar los ítems que presentaron un índice de discriminación no aceptable y los distractores que el grupo con altas puntuaciones seleccionó.

Cabe destacar que, los resultados de este estudio se consideran preliminares, y en proceso se encuentra la aplicación a la población de estudiantes que cursan la asignatura en el primer periodo del 2017, la cual incluye las recomendaciones del presente estudio.

Finalmente, el desarrollo de instrumentos de calidad está ligado a la necesidad de contar con insumos válidos y confiables para determinar el grado de conocimiento de los estudiantes respecto a un tema en particular. Por lo anterior, está la idea de mejorar las prácticas educativas en pro del desarrollo de la comunidad educativa. Este es un objetivo de la UABC, el cual busca a partir de implementar evaluaciones como la presente, que parten de un instrumento construido con gran rigor metodológico, donde fueron seguidas las recomendaciones plasmadas en los Estándares y los criterios planteados por el INEE, lo cual garantiza de cierta forma el poder contar con un instrumento que ayude a identificar los contenidos del currículo cubiertos.

Referencias

American Educational Research Association (2014). Standards for Educational and Psychological Testing. American Educational Research Association, American, Psychological Association & National Council on Measurement in Education: EE. UU.

Anderson, L. W. & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing. New York, EE.UU.: Longman.

Berk, R. (2006). Thirteen Strategies to Measure College Teaching: a consumer´s guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians. Virginia, EE.UU.: Stylus Publishing.

Borsboom, D., Mellenbergh, G. & Heerden, J. (2004). The Concept of Validity. Psychological Review, 111(4), 1061-1071.

Contreras, L. (2011). Modelo para desarrollar pruebas de gran escala de referencia criterial alineadas con el currículum. En E. Luna (Coord.), Aportaciones de la investigación a la evaluación de estudiantes y docentes. México: UABC, Porrúa.

Escudero, T. (1997). Enfoques modélicos y estrategias en la evaluación de centros educativos. RELIEVE, 3(1). Recuperado de http://www.uv.es /RELIEVE/v3n1/RELIEVEv3n1_1.htm

Instituto Nacional para la Evaluación de la Educación (2015). Las pruebas ENLACE y EXCALE. Un estudio de validación. Felipe Martínez Rizo, coord. Cuaderno de investigación, número 40.

Martínez, F. (2003). Pruebas y rendición de cuentas. Instituto Nacional para la evaluación de la Educación Cuaderno de investigación, número 12.

Messick, S. (1995). Validity of Psychological Assessment, Validation of Inferences from Person´s Responses and Performances as Scientific Inquiry Into Score Meaning. American Psychologist, 50 (9), 741-749.

Scriven, M. (1991). Evaluation Thesaurus. (4ª ed.). California, EE.UU.: Sage publications.

Universidad Autónoma de Baja California (2015). Plan de Desarrollo Institucional 2015-2019. UABC: México.