Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 11-31.
https://doi.org/10.15366/riee2017.10.1.001

La Objetividad en las Pruebas Estandarizadas

Objectivity in Standardized Tests

Agustín Tristán López*
Nancy Yahibé Pedraza Corpus

Instituto de Evaluación e Ingeniería Avanzada (IEIA)

La objetividad es un atributo necesario que debe detallarse claramente para satisfacer los propósitos científicos de todo proyecto de evaluación en ciencias de la salud, ciencias sociales y educación, así como en cada una de las etapas de producción y uso de las pruebas estandarizadas. El valor de la objetividad para el desarrollo de las pruebas se refuerza al emplearse como herramienta de vigilancia que garantiza la neutralidad en los estímulos presentados. Se detallan cinco propiedades principales distintivas: especificidad, neutralidad, independencia, imparcialidad e impersonalidad, fundamentales para interpretar los resultados, eliminar o reducir los sesgos inducidos por la influencia de estereotipos y preferencias en el diseño del instrumento o en la apreciación de jueces, entre otros factores que pueden afectar el uso ético de los resultados de las pruebas. Se muestra que la objetividad es el primer atributo que debe definirse en una prueba estandarizada, distinguiendo las cualidades que le son propias para evitar asociarlas incorrectamente con la validez o la confiabilidad.

Palabras Claves: Objetividad, Pruebas estandarizadas, Validez, Confiabilidad.

Objectivity is a needed attribute of standardized tests in different areas, such as health, social sciences and education, and in each one of the phases of the development of a test, from its initial definition to the interpretation of outcomes. Objectivity ensures fairness of the test from its design up to the appraisal of the judges or evaluators and on the treatment of results, grounded on five main properties: specificity, neutrality, independence, impartiality and impersonality. Objectivity is fundamental for the interpretation of the outcomes, eliminating or reducing the presence of stereotypes and preferences that produce several types of bias that may affect the ethical use of the results of the test. Objectivity should be the first attribute to consider in a standardized test, as it improves the definition of the traits to evaluate permitting the distinction of characteristics that are mistakenly associated with validity and reliability.

Keywords: Objectivity, Standardized tests, Validity, Reliability.

*Contacto: atristan@ieia.com.mx

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 23 de octubre de 2016
1ª Evaluación: 15 de enero de 2017
Aceptado: 25 de febrero de 2017

1. Presentación

Las pruebas estandarizadas en educación siempre están en el ojo inquisitivo de funcionarios y autoridades públicas, de asociaciones de padres de familia, de docentes y de los estudiantes que deben resolver la prueba. No es propósito de este trabajo hacer la diatriba de estas pruebas ni tampoco su defensa, sino apuntar de manera breve las cualidades de la objetividad que, junto con la validez y la confiabilidad, configuran los tres atributos fundamentales para el diseño, administración, interpretación y uso de las pruebas estandarizadas.

En general los proyectos de evaluación se centran en garantizar la confiabilidad del instrumento, frecuentemente dejando a la validez como atributo subordinado y, la mayoría de las veces, sin citar a la objetividad como atributo indispensable. Para ilustrar esta afirmación, pero sin el propósito de realizar un meta análisis, la tabla 1 muestra la cantidad de entradas que se obtuvieron en un popular buscador de Internet para validez, confiabilidad, objetividad (en inglés y en castellano) y sus combinaciones. Los datos no representan preferencias definitivas de investigadores y evaluadores, pero ilustran la incidencia de los tres atributos en la Web. La objetividad es el atributo con menos referencias o entradas y, notablemente, su incidencia es menor en combinación con los otros atributos.

Tabla 1. Frecuencia de entradas dentro del buscador Google para "validity, reliability, objectivity"

	Frecuencia (en miles)	
Atributo	**Inglés**	**Castellano**
Validez	89,000	19,400
Confiabilidad	169,000	8,860
Objetividad	11,000	4,920
Validez y confiabilidad	4,187	243
Validez y objetividad	107.8	60.5
Confiabilidad y objetividad	183.7	32.2
Los tres atributos	26.1	2.15

Fuente: Elaboración propia a partir del buscador Google, octubre de 2016.

¿Cómo explicar que de 169 millones de entradas para confiabilidad (las cantidades son notablemente inferiores en castellano), se tenga una reducción a casi 4.2 millones al combinarse con validez y se tengan solo 26 mil entradas combinadamente? No parece explicable que un atributo tan importante se haya escapado a los especialistas en evaluación o a los investigadores de la psicometría.

Una explicación es que, de los tres atributos fundamentales de la evaluación y de la medición, la objetividad es el más complejo de definir y de aprehender. De hecho, varios de los principales factores de la objetividad se transfieren, erróneamente, a la validez, reduciendo a la objetividad a pocos aspectos, importantes pero insuficientes, castigando a la validez al contener factores que le son ajenos y que la convierten en un atributo altamente complejo; que además de referirse al concepto primigenio de que "el instrumento de medida sirva para el propósito previsto", también se asocia con la interpretación de los resultados. Así se han acuñado nuevos términos como "validez de uso", "validez consecuencial" y "validez cultural", entre otros, olvidando que están asociados con la objetividad. También la confiabilidad ha tenido que absorber aspectos que atañen a la objetividad, como es el sesgo de diseño.

Los propósitos de este trabajo son varios: aclarar el concepto de objetividad, explicar su importancia como uno de los tres atributos fundamentales de la evaluación y, sobre todo, establecer su papel dentro de las pruebas estandarizadas. Para cumplir con estos propósitos debe recurrirse a diversas áreas del conocimiento, pero se ha optado por abordar tres facetas teóricas para identificar las cualidades ontológicas, epistemológicas y éticas de la objetividad, incluyendo ejemplos en el terreno de la evaluación estandarizada.

Reflexionar sobre la objetividad en las pruebas estandarizadas no es trivial ni ocioso. Una vez comprendida su importancia, permite liberar de complicaciones a la validez y a la confiabilidad de las pruebas en general y de las estandarizadas en particular. Este trabajo no pretende hacer una reseña histórica de la objetividad a través de la filosofía y otras áreas del conocimiento; el lector interesado puede referirse al trabajo de Gaukroger (2012).

2. La objetividad como desiderátum

La objetividad es la cualidad inherente de un objeto en sí mismo, ajeno a cualquier enfoque especulativo (Real Academia Española, 2016; Zamora, 2007). Se ha utilizado en el tratamiento metódico y controlado para definir y estudiar "entes" y como base de discusión científica y filosófica al cuestionar la existencia "real" de las cosas (García, 1955) para determinar la posibilidad de alcanzar un conocimiento "real" del mundo fuera de otras aproximaciones. Las discusiones en torno a la objetividad han tratado de esclarecer su relación con la verdad, la realidad, la existencia y el ser como tal de los objetos, orientando el trabajo filosófico y científico desde el positivismo del siglo XIX hasta el relativismo contemporáneo, provocando diversas aproximaciones, tendencias y conflictos ontológicos.

La tensión entre objetivismo y subjetivismo ha permitido encontrar elementos útiles dentro de ambos, contribuyendo al desarrollo del conocimiento científico, particularmente en áreas predominantemente especulativas en el tratamiento del objeto de estudio, como en ciencias sociales y de la salud. Una postura más moderada reconoce las limitaciones de la visión objetiva mecánica, que pretende despersonalizar al observador para evitar que sus juicios afecten las descripciones que hace del objeto que analiza y, en cambio, reconocer que el observador no puede ser ajeno del todo a lo que observa, pero consciente de esta implicación y conociendo sus prejuicios, debe poder apartarlos del objeto en estudio (Cupani, 2011; Morales de Barbenza, 2001).

La objetividad es un desiderátum, es decir, inalcanzable plenamente por varias razones. Por una parte la ciencia, sus productos y motivaciones, son resultado de la actividad cognitiva que hace cada individuo sobre un objeto en particular; por otra parte, las representaciones o definiciones de un objeto están sujetas a la aprehensión consciente del investigador, lo cual queda obligatoriamente vinculado a su subjetividad al observar, medir, valorar, controlar o asignarle categorías lógicas dentro de un sistema teórico (Cupani, 2011). El estar consciente de esta limitante brinda la oportunidad de plantear aproximaciones al objeto por conocer, merced a un alto grado de "indiferencia en el juicio, que puede estar en conflicto con nuestras necesidades y deseos" (Gaukroger, 2012). En consecuencia el juicio se despersonaliza como si fuera hecho desde el exterior del propio sujeto.

Si se parte del argumento filosófico de que ningún objeto es aprehensible directamente, porque su existencia implica una declaración metafísica alrededor de cualidades inherentes u ontológicas, entonces, se acepta que todo objeto por conocer (real o ideal) posee un ser inteligible e identificable como correlato del objeto respecto de un conjunto de características específicas (García, 1955) que, al encontrar una expresión material (concreta o abstracta) adquieren realidad objetiva. El citado correlato puede ser resultado de una medición sobre el objeto, producto de un razonamiento deductivo formal o designarse por consenso de una comunidad científica o profesional, lo cual se denomina realidad subjetiva.

La objetividad es el resultado de un proceso dual que se basa, por una parte, en la contrastación de conocimientos e ideas en un mundo empírico y, por otra, en la intersubjetividad donde un grupo acepta la construcción de esa idea como válida por un acuerdo convencional sobre un mismo objeto partiendo de apreciaciones subjetivas. El proceso de objetivación reconoce las manifestaciones materiales de los objetos, independientes del observador, aceptando que, a pesar de que cada observador es único o singular, está en posibilidad de establecer criterios sobre sus afirmaciones, de modo que el acuerdo convencional elimina toda discrepancia entre observadores.

Los investigadores pueden definir formalmente, de manera conceptual o empírica, el objeto en estudio, en función de un contenido, una representación y una estructura con fundamento en cada área del conocimiento. El acuerdo convencional se fundamenta en definiciones dinámicas, aprovechando que la ciencia es autocrítica y se auto corrige, con estructuras que se construyen y reconstruyen al incorporar nuevo conocimiento proveniente de evidencia empírica o de postular nuevas categorías formales, en un ejercicio de honestidad intelectual (Gaukroger, 2012).

Una implicación evidente de no buscar sistemáticamente la objetividad como criterio científico en el sentido que le da Popper (citado por Larroyo, 1968), es que el conocimiento derivado de su estudio, puede no corresponder a las características o atributos del objeto o, peor aún, que los elementos estudiados sean plenamente dependientes del observador, haciendo que los atributos adjudicados al objeto estén más bien vinculados a otros procesos que no definen ni explican en absoluto lo que pasa con dicho objeto. En consecuencia, los atributos descritos por un observador serán discrepantes de lo que puede establecer otro observador, induciendo a que el tratamiento del objeto no sea sistemático y dependa de la interpretación de cada persona, de lo que se desprende la necesidad de la intersubjetividad citada previamente. Queda claro que la objetividad es un criterio fundamental en el desarrollo de la investigación científica, porque permite generar conocimientos válidos sobre los objetos investigados.

La objetividad depende de dos aspectos fundamentales: la especificidad y la interpretación.

1) La especificidad es la representación de la realidad, contenida en una definición completa, pertinente, precisa del objeto y que lo distingue de otros. Para esta definición se justifica el uso de un arquetipo como referencia para los juicios de valor que se pueden hacer de los objetos de su mismo género o especie. En consecuencia de su definición, la objetividad no es un constructo universal que todas las personas perciben de la misma forma, sino que requiere de la aceptación convencional a partir de las cualidades intrínsecas incluidas en la definición. La especificidad implica que la definición del objeto debe distinguir claramente entre

cualidades inherentes y otros elementos que pueden catalogarse como requisitos, criterios de inclusión o de exclusión, condiciones reglamentarias o administrativas para tener derecho a participar en un proceso de evaluación. Por ejemplo, el que la prueba PISA se administre a jóvenes de 15 años es un criterio de inclusión para el proyecto que restringe a otras personas a ser parte de la población focal, pero esta condición no se considera como un sesgo o una valoración subjetiva respecto de dicha población. Los requisitos no producen medidas respecto del objeto, por lo tanto no deben aportar calificaciones o puntajes a las personas de la población focal, ni tampoco generalizaciones que comprometan el uso ético de la información

2) La interpretación se asocia con las justificaciones de los usos y juicios de valor que pueden postularse a nivel contextual, cultural, grupal, o de otra índole, a partir de datos obtenidos de la realidad. Las interpretaciones y justificaciones responden a la necesidad de identificar, prevenir, medir y, de preferencia, eliminar o reducir al mínimo la presencia de sesgos en las apreciaciones de las personas que van a emitir juicios de valor sobre los objetos en estudio. Este tópico es complejo porque hay fuentes de sesgo imputables al proyecto, al evaluador y a la población, lo cual incide en problemas de diseño, al construir los ítems, al administrar la prueba, al emitir juicios de valor, de tal modo que las interpretaciones se ven afectadas por todas estas condiciones.

Al aceptar que la objetivación es factible de ser alcanzada y definida, se avanza contra el escepticismo que niega dicha factibilidad, en particular en las ciencias sociales, psicología, educación y áreas de la salud; aunque, en el extremo, la objetividad matemática tampoco sería alcanzada por tratar con entes abstractos cuya manifestación real es siempre imperfecta. Por ejemplo, el concepto de triángulo como figura plana cerrada de tres lados, cuyos ángulos internos suman 180º es geométricamente perfecto, pero solo puede dibujarse en un papel de manera aproximada por un dibujante experto. Sin embargo, el arquetipo del triángulo es una formalización que perfecciona la percepción que se hace de un objeto real que solo puede existir de manera imperfecta; este perfeccionamiento lo hace objetivo y permite que se interpreten sus propiedades de la misma manera por el dibujante, el ingeniero que dirige la edificación, el albañil que ejecuta la construcción o la persona que va a contemplar la obra terminada. De igual modo, en el campo de la evaluación, se tiene que mirar la objetividad de los modelos estadísticos como paradigmas esperados del comportamiento de un ítem o de un test estandarizado, lo cual invalida las objeciones que rechazan la construcción de modelos teóricos al centrarse en lo que denominan "evidencia empírica", dejando abierta la relatividad de la recolección de los datos, el juicio del observador y cualquier otra fuente de subjetividad en la definición del constructo, el diseño del instrumento y la interpretación de los resultados. Como apuntan Myers y Hansen (2002), puede afirmarse que la objetividad no niega la utilidad de recabar datos de la realidad, pero advierte que no son suficientes para garantizar que se obtienen conclusiones correctas.

Algunas cualidades que posee la objetividad se pueden asimilar a propiedades asociadas con ella (Gaukroger, 2012), en particular (1) la ausencia de sesgo en la interpretación y la toma de decisiones, (2) la eliminación de prejuicios personales, por lo tanto, libres de supuestos y valores individuales, (3) la facultad de distinguir entre dos ideas o teorías contrastantes o hasta en conflicto respecto del objeto, sustentándose en (4) la definición exacta del objeto. Dicho autor advierte en no asociar la objetividad exclusivamente con la cuantificación y la medida o con la acumulación e interpretación de datos. La medida

es parte importante de la objetivación, pero no es el único elemento que la constituye, porque reduciría el concepto de estandarización a la obtención de resultados que pueden medirse, analizarse en forma matemática o estadística para su comparación.

Entrando en el terreno de la evaluación y como consecuencia de lo expresado, la definición del objeto es el punto de partida para el proceso de evaluación porque es el que permite hacer apreciaciones cualitativas o cuantitativas respecto de los atributos inherentes del objeto, propiciando profundizar en sus características y funciones. La definición del objeto a medir debería ser una de las responsabilidades y preocupaciones de los diseñadores de pruebas y cuestionarios, porque sin una definición objetiva, es altamente probable que se obtengan resultados y conclusiones poco fieles del objeto que, en este caso, se trata de cualidades de las personas evaluadas; donde el resultado de esa medición tendrá un impacto muy importante en su vida; como en el caso de ser admitido o rechazado en la universidad; del diseño de políticas y programas gubernamentales; del establecimiento de una campaña de salud pública. De no hacer una adecuada definición puede verse el enorme riesgo de generar un instrumento que proporcione información deficiente en términos de objetividad.

La evaluación depende de la objetivación para definir el objeto de medida. Por ejemplo, puede tenerse interés de disponer de un instrumento para medir la temperatura corporal, lo cual se resuelve fácilmente con adquirir un termómetro en la farmacia más cercana. Este aparato debe ser válido y confiable para obtener medidas certeras de temperatura al utilizarlo con cualquier persona. Todo parece simple, siempre y cuando el concepto de temperatura tenga una definición objetiva, sin confundir "temperatura" con "calor", "fiebre", "dilatación" u otro concepto con el cual esté posiblemente asociada la temperatura, pero que se manifiesta, mide e interpreta de otra forma. El funcionamiento del termómetro será válido para el propósito de medir el objeto deseado (temperatura), sin cuya definición sería inapropiado el uso del instrumento e inútiles las medidas resultantes. La objetividad, como desiderátum precede, por lo tanto, a la validez y a la confiabilidad.

3. La objetividad como herramienta epistemológica

Una forma de aproximarse a un objeto concreto es identificar algunas propiedades físicas observables: dimensiones geométricas, características de materiales y forma, cantidad de un atributo como peso o temperatura (Nunnally y Bernstein, 1995). Los investigadores al medir buscan "asignar símbolos a objetos de manera que (1) representen cantidades o atributos de forma numérica (escala de medición) o (2) definan si los objetos caen en las mismas categorías o en otras diferentes con respecto a un atributo determinado (clasificación)" (Nunnally y Bernstein, 1995). La forma de identificar las propiedades puede ser tanto teórica como experimental, por lo que se han ideado instrumentos y estrategias para medir uno o varios atributos del objeto en escalas apropiadas.

La medición de un objeto abstracto (como la inteligencia, la percepción de un síntoma hepático, la depresión, el aprendizaje, entre otros), presenta importantes limitaciones prácticas, ya que la relación existente entre el objeto y su realidad objetiva no es directa, como en los objetos concretos (Kerlinger y Howard, 2008b). Para el análisis y la medición los investigadores precisan definir un conjunto de características del objeto, concentrándose en una o en algunas manifestaciones de ellas, denominadas "rasgo" (Nunnally y Bernstein, 1995). En el caso particular de los rasgos observables de forma

indirecta a través de comportamientos y expresiones diversas que realice una persona, se habla de rasgos latentes y se presume que de forma indirecta pueden ser medidos al observar tales manifestaciones que contienen la característica prevista del objeto. La forma de vincular la realidad objetiva y el rasgo latente es altamente compleja, con gran probabilidad de confusión e imprecisión, máxime que distintos enfoques científicos, áreas del conocimiento o sistemas teóricos pueden estudiar simultáneamente el mismo rasgo latente atribuyéndole propiedades, funciones y manifestaciones distintas.

Al definir el objeto de medida en un proyecto de evaluación se pueden identificar los límites, alcances, interpretación y uso de los resultados (Jornet y Suárez, 1996). De nuevo, la objetividad precede a los atributos de validez y confiabilidad, dirigiéndolos en el rumbo previsto por el proyecto evaluativo, de tal modo que las consecuencias de una evaluación no son inherentes a estos dos atributos, sino a la objetividad que las antecede. De este modo la objetividad no se limita a la definición del objeto de medida y a su interpretación, sino también contiene un conjunto mucho más amplio de propósitos, entre ellos:

a) Aprehender las cualidades inherentes del objeto, al definirlo, caracterizarlo, categorizarlo, compararlo, ponderarlo, valorarlo o medirlo, entre otras formas.

b) Emitir juicios de valor sobre uno o varios rasgos o características inherentes del objeto, en función del objeto mismo, de una población dada, o respecto de criterios externos de referencia o de comparación.

c) Plasmar (en forma conceptual, simbólica, matemática o de otra índole) las cualidades, características o rasgos de un objeto para su análisis y aprehensión por diversas personas, incluyendo el evaluador y el evaluado, o un público independiente.

d) Reducir el sesgo de diseño del instrumento para propiciar la apreciación formal del objeto con especificaciones definidas en forma concreta u operacional, en forma independiente de las poblaciones en las que se utilice.

e) Acotar la interpretación subjetiva del evaluador respecto del rasgo evaluado, en un momento dado, o a lo largo del tiempo por cambios de criterios que experimenta el evaluador.

f) Reducir la diferencia de apreciación de diversos evaluadores, en función de criterios, consideraciones o prejuicios personales.

g) Evitar la diferencia de apreciación entre el evaluador y el evaluado, haciendo que este último perciba su dictamen como aceptable.

h) Eliminar el efecto de fatiga o influencia cualitativa por el número de juicios emitidos en un tiempo dado ante una población numerosa.

i) Anular el efecto de halo, de prejuicios discriminatorios o por influencia de estereotipos en el evaluador.

j) Eliminar la discrepancia de opinión respecto de la respuesta correcta o más aceptable, facilitando la calificación por personal no experto e, inclusive, por medio de un programa informático con base en una clave de respuestas.

k) Comparar las cualidades métricas de varios instrumentos, incluyendo el error de medida y la consistencia de resultados que se obtienen con una población focal dada.

l) Obtener medidas de los ítems independientemente de la población o personas particulares que intervienen en la aplicación del instrumento y, en contraparte, obtener medidas de las personas de la población focal independientes del conjunto de ítems utilizados en el instrumento.

Las pruebas estandarizadas son los instrumentos de medición más utilizados en psicología, educación, ciencias de la salud y ciencias sociales, que cuentan con un amplio desarrollo técnico y metodológico con formas perfeccionadas para medir los rasgos observables o latentes, en la población focal específica y con un grado de precisión previamente establecido y controlado por procedimientos logísticos y administrativos igualmente objetivos. Los atributos de validez y confiabilidad de las pruebas estandarizadas han sido objeto de muchos debates y de críticas que no se repetirán en este trabajo, sin embargo, parte de ellas se deben al limitado, por no decir nulo papel que se le ha dado a la objetividad como atributo de las pruebas estandarizadas (Borsboom, Mellenbergh y Heerden, 2004; Embretson, 2007; Kane, 2008; Kerlinger y Howard, 2008a-b; Mislevy, 2007; Newton y Baird, 2016; Padilla, Gómez, Hidalgo y Muñiz, 2006; Sijtsma, 2009).

La idea de base de las pruebas estandarizadas como instrumentos de medidas de objetos abstractos o rasgos latentes cuenta con una profunda influencia del positivismo del siglo XIX, que buscaba establecer con el mayor rigor metodológico posible una definición del objeto de estudio, por ejemplo, la inteligencia o el rendimiento escolar (Binet, 1910), asumiendo que las manifestaciones observables de los rasgos latentes son objetivaciones que se quieren medir en el objeto y cuyos resultados se reportan en una escala es un eje cartesiano igualmente objetivo, que corre de menos a más respecto del atributo. Para garantizar la precisión de los resultados se cuenta con medidas objetivas del error que requieren de un control también objetivo de las situaciones en las que realiza la medición, todo lo cual anula o, por lo menos, reduce la influencia de variables que afecten la medición, por ser dependientes de varios agentes: el evaluador en el diseño del instrumento objetivo, el aplicador al administrar la prueba de forma objetiva y de la persona a evaluar al responder ítems objetivos por medio de un desempeño objetivo (Binet, 1910). Todos los elementos fueron adjetivados con la palabra "objetividad", recordando la necesidad de que la prueba sea objetiva, pero no debe pensarse que todas las pruebas objetivas están estandarizadas y, desde luego, no puede garantizarse que todas las pruebas estandarizadas disponibles en el mercado sean objetivas.

La objetividad sirve al propósito de la medición, ayuda a definir y delimitar el objeto a evaluar, así como a proporcionar elementos de control de dicha medición, para limitar que variables externas afecten el resultado, y que el medio ambiente, incluyendo al administrador de la prueba, no interfiera con el resultado.

Los valores y los propósitos de la objetividad contribuyen a reducir o constreñir la intervención subjetiva de quien administra o evalúa una medición (Cupani, 2011), reduce la influencia del evaluador cuando corrige una prueba, al valorar el resultado final, el nivel de desempeño de un estudiante en una asignatura escolar o el grado de enfermedad de un paciente ante un síntoma (Céspedes, 2009). De nada sirve contar con una prueba estandarizada de buena calidad, si la persona que va a aplicarla e interpretar los

resultados no está capacitada o si el dictamen final depende de criterios no ligados al objeto. Por lo tanto, la objetividad incide en varias fases del proyecto de evaluación: la planeación y diseño del instrumento, su administración, el control del proceso y la logística, la calificación y la interpretación, por ello no solo precede a los atributos de validez y confiabilidad como se indicó en la sección anterior, sino que funciona como un control de calidad de cada etapa de desarrollo de ambos atributos.

Puede verse, por lo tanto, que la definición primigenia de validez como el grado en que una prueba mide el propósito que se pretende medir es muy apropiada, porque asume que el objeto de medida fue definido claramente (Kelley, 1927). Además, cada etapa que permite obtener evidencias de validez se concentra en la sensibilidad del instrumento para captar el objeto y los atributos definidos en su objetivación. Este reconocimiento hace evidente que toda medida es imperfecta y como tal, tiene un margen de error que se vincula y calcula a través de procedimientos estadísticos, que objetivan a la confiabilidad. La definición de estos atributos, entre muchos otros, ha sido emprendida por diversas agencias o instituciones (APA, 1954-2010; AERA, APA, NCME, 2014), las cuales han sido sometidas a análisis, críticas y escrutinios dentro de la comunidad académica (Campbell, 1960; Chan, 2014; Guilford, 1987; Jeffrey, 2003; Kimberlin y Winterstein, 2008; Lane, 1999; Moss, 2007; Newton y Baird, 2016; Sireci, 2007; Sireci y Padilla, 2014).

La definición primigenia de validez es objetiva en los conceptos de validez de contenido, de constructo, de criterio (predictiva, concurrente, discriminante…) y de escala, pero se modificó el modelo al plantearse que la validez no es un atributo inherente del instrumento sino que depende del uso e interpretación que se haga de los resultados, lo cual involucra implicaciones éticas (Borsboom, Mellenbergh y Heerden, 2004; Chan, 2014; Jeffrey, 2003; Lissittz y Samuelsen, 2007; Messick, 1995; Zumbo, 2009). De esta forma, el uso y la interpretación caen en el terreno de la objetividad, no siendo pertinente adjudicarlos a la validez, porque esto complica y enturbia su significado dentro de la evaluación y despoja a la objetividad de algunos de sus propósitos.

Respecto de la posible confusión entre objetividad y validez, es importante citar que, de acuerdo con Borsboom et al. (2004), una prueba es válida cuando el atributo existe y sus variaciones producen causalmente variaciones en la medición. Esta definición de validez, parece un sano retorno al concepto inicial pero con base en un sustrato distinto, al surgir de una reflexión ontológica (André y Loye, 2015; Jeffrey, 2003) sobre la objetivación de "aquello" que se quiere medir, distinguiendo los rasgos inherentes al objeto de los que no lo son. Si un objeto cambia, entonces se debe reflejar un cambio en su medida, lo que requiere de un proceso constante de objetivación y mantener esa vigilancia durante el proceso de medición. En caso contrario, es indispensable objetivar nuevamente el objeto y su medida, lo cual puede repetirse las veces que sean necesarias para garantizar que las medidas y las unidades que se utilizan miden lo que deben medir. Aceptando que la objetividad es el sustrato de la validez, en ausencia de ella, la validez queda seriamente comprometida.

Una prueba estandarizada debe tener claramente objetivado el rasgo con elementos de la realidad objetiva y de la realidad subjetiva. Para operacionalizarlo es posible utilizar enunciados, categorías y variables susceptibles de ser exploradas de forma cualitativa o cuantitativa. Todas las pruebas, en particular las estandarizadas, deberían usar diversas técnicas para comprobar que la operacionalización corresponde a los rasgos que se pretende medir. Esta comprobación puede hacerse a través del consenso del juicio de

expertos (evaluación de realidad subjetiva por terceros), con pruebas de correlación entre ítems, ítem contra prueba, entre pruebas distintas, con la misma prueba a lo largo del tiempo o con poblaciones de contraste, entre muchas otras formas.

En los propósitos de la evaluación objetiva se asocia la operacionalización con la independencia entre el evaluador y el evaluado, entre la medida del ítem y la del sujeto. La independencia es una cualidad de la objetividad que sistematizó Rasch (1980) con el concepto de independencia local y que garantiza que la probabilidad de respuesta de un sujeto ante un estímulo dado es una función que depende de la medida del sujeto y de la dificultad del ítem, independientes entre sí. Este modelo se ha extendido al análisis de facetas múltiples que permite incluir la opinión de los evaluadores y de variables de contexto (Linacre, 1994).

En general la confiabilidad ha tenido menos conflictos de interpretación que la validez, especialmente si se toma en el sentido de expresar valores relacionados con el grado de precisión de las medidas (Nunnally y Bernstein, 1995), pudiendo provenir de modelos que estiman la consistencia de los datos, la homogeneidad de los ítems y de la población, o la repetitividad de los resultados cuando la prueba es administrada a lo sujetos en condiciones controladas (Argibay, 2006; Carvajal-Carrascal, 2012; Kerlinger y Howard, 2008a; Sánchez-Meca, López-Pina y López, 2009; Zúñiga y Montero, 2007), siendo el Alfa de Cronbach, la teoría G y la separación logística, los modelos más utilizados en la práctica, dentro de un abanico enorme de modelos que persiguen calcular el error de medida de cada ítem, de la prueba en su conjunto, de los puntos de corte, entre otros elementos que tratan de brindar medidas objetivas de la precisión de la medida, aunque no de la calidad del instrumento. Tradicionalmente, los valores aceptables del Alfa de Cronbach se dejan a juicio del evaluador, es decir, quedan supeditados a criterios subjetivos (Blanco-Villaseñor, 1991; Nunnally y Bernstein, 1995) por lo que no se ve problema en aceptar un valor de Alfa de 0.8 en una prueba estandarizada y se rechaza que una de las partes de la prueba tenga valores tan bajos como 0.4 (Tristán, 1996-2010). Es posible establecer criterios objetivos para demostrar la pertinencia de ambos valores sin apelar a artificios en el diseño (incrementar el número de ítems o restringir la dificultad de los ítems alrededor del punto de corte) conduciendo a un instrumento con una alta confiabilidad a expensas de una pobre validez.

Modelos matemáticos y estadísticos más sofisticados favorecen la creación de herramientas que incorporan distintos supuestos sobre las variaciones en las puntuaciones (Shavelson y Webb, 2005; Ritter, 2010) en particular a través de modelos logísticos o multivariados para analizar el funcionamiento diferencial de cada ítem o de la prueba en su conjunto, con énfasis en reducir o corregir el sesgo inherente al diseño o relativo a la población evaluada (Bond y Fox, 2015; Fox y Glas, 2001,2003; Gómez y Hidalgo, 2003; Jiménez y Montero, 2013; Linacre y Wright, 1995; Prieto y Delgado, 2003; Wright y Stone, 1999; Wright y Mok, 2000). Tomar en cuenta el funcionamiento diferencial o la presencia de algún sesgo es fundamental al emitir juicios de valor sobre personas en forma individual o grupal, lo cual va más allá del interés estadístico por sus consecuencias éticas.

4. Objetividad y consideraciones éticas en las pruebas estandarizadas

El método científico tiene como característica inmanente (explícita o no) a la objetividad (Muñiz, 2010), porque se espera que las preferencias, actitudes, valores y prejuicios del investigador no afecten su trabajo. Se extrapola esta idea a las pruebas estandarizadas, al desarrollar instrumentos de medición en las ciencias sociales y de la salud perfeccionados con técnicas psicométricas y predictivas con rigor científico. Este desarrollo diluyó aparentemente la discusión sobre la relevancia, la utilidad y las implicaciones del uso ético de las pruebas (André y Loye, 2015), en parte por el tiempo que ha implicado desarrollar técnicas y software de análisis estadístico, así como enfrentar cierto rechazo a las pruebas estandarizadas, a la pertinencia de su uso y puesta a disposición de profesionales certificados para su administrarlas, interpretar los resultados y tomar decisiones prácticas dentro de un marco ético o de justicia para las personas evaluadas.

Los artículos de difusión de resultados, especialmente los de la segunda mitad del siglo XX en los Estados Unidos de América, trataban de convencer al lector de los beneficios de la estandarización desde el punto de vista positivista, vinculando el desempeño (intelectual, académico y laboral) con grupos de personas, mostrando diferencias entre géneros, etnias, culturas y niveles socioeconómicos, reforzando estereotipos y clasificaciones discriminatorias (Herrenstein y Murray, 1994; Bowen y Bok, 1998), provocando un impacto político y social resultante de algunas debilidades de estas pruebas. Las soluciones se concretaron de varias maneras: La primera fue criticando los defectos de las pruebas, promoviendo su erradicación en el ámbito de la educación y sugiriendo modelos de evaluación "auténtica" (Froese-Germain, 1999). Una segunda línea fue de tipo legal bajo sentencias judiciales y enmiendas del Congreso de los Estados Unidos (Enmienda Buckley de 1976 o FERPA) para supeditar el papel de las pruebas estandarizadas a los derechos civiles, durante la aplicación, la calificación y la utilización de los tests (Gómez, Hidalgo y Guilera, 2010; Nunnally y Bernstein, 1995). La tercera línea técnica construyó estándares para el diseño de pruebas por el Joint Commitee (AERA-APA-NCME, 2014), o estándares de buenas prácticas y equidad en las pruebas (Educational Testing Service, 1987; International Test Commission, 2014-2016). Una cuarta línea defendió las pruebas estandarizadas con base en argumentos objetivos, (curiosamente sin invocar a la objetividad) contrastando sus ventajas contra otras formas de evaluación (Phelps, 2005).

La defensa de las pruebas estandarizadas ha implicado aportar elementos para corregir deficiencias reveladas por las críticas de sus detractores con un impacto ético. Estos elementos agregados sobre todo a la validez y a la confiabilidad las convierten en atributos "ómnibus" que absorben todo lo que permita reforzar a las pruebas, pensando que enderezan el camino de las pruebas estandarizadas pero que enturbian su existencia, complicando su vulnerabilidad en el campo ético frente a una mirada inquisitiva y crítica. Toda proporción guardada, son empeños similares a los que defendían el modelo geocéntrico de Tolomeo, agregando elementos complicados y tortuosos para explicar la cinemática de los cuerpos celestes, frente al modelo heliocéntrico de Copérnico, simple, claro y preciso. Las implicaciones éticas de la objetividad se relacionan con las propiedades de neutralidad, imparcialidad e impersonalidad del observador-evaluador.

La impersonalidad hace explícitas y conscientes las representaciones culturales y sociales implicadas en una prueba estandarizada y, por lo tanto, bajo la responsabilidad

de las personas que la desarrollan, desde los consejeros que determinan el objeto de medida, hasta los responsables de su utilización e interpretación, pasando por los diseñadores de ítems y los encargados del procesamiento estadístico. Es fundamental definir claramente el objeto de medida, sus interacciones con factores psicológicos, biológicos, ambientales y de experiencias previas que puedan afectar o condicionar la obtención de evidencias sobre el objeto, especialmente cuando es un rasgo latente. La representación debe explicitar cómo el objeto es compartido en el grupo social, cultural, étnico, en un momento dado o en su devenir temporal y contextual (etario, regional, socioeconómico). La impersonalidad obliga a adaptar una prueba creada en un idioma o país para aplicarse en otro, no solamente como traducción sino como concepción del objeto, definiendo las situaciones o casos que describen y aclarándolas para cada contexto. Esto requiere de un arduo trabajo de interpretación de la prueba, de validación para cada población y el establecimiento de criterios de corte y baremos para los diversos grupos poblacionales (Muñiz, Elosua y Hambleton, 2013; Sattler, 2010).

La neutralidad requiere que no haya injerencia externa en los juicios de valor que emite un evaluador con los resultados de una prueba estandarizada, haciéndola aplicable a todas las personas, en todos los ambientes y condiciones, obteniendo medidas libres de otras características ajenas al objeto. Por ejemplo, se tiene un problema de neutralidad en una prueba aplicada por un sindicato para clasificar personal en un puesto de trabajo, si el resultado que se emite es distinto cuando las personas están sindicalizadas o no. En el caso de la prueba PISA se tiene un problema de falta de neutralidad, si los textos utilizados como situación para derivar los ítems hacen referencia a objetos comunes en un país y que no son comprensibles para los estudiantes de otro.

Una prueba de comprensión lectora sobre el tópico central de un texto y diversos aspectos gramaticales concibe que ambos son constructos neutrales y no personalizados. De hecho, se puede plantear sobre un texto que describa la belleza del campo (neutral y no personalizada), o sobre un texto que detalle una situación de violencia social (personaliza aunque puede ser neutral si no toma una postura) o un relato que ridiculice a los seguidores de una religión (personaliza y no es neutral por demeritar al grupo en cuestión). La respuesta ante esos estímulos será diferente porque movilizará en cada persona sentimientos y reacciones ajenas al propósito de medida.

La imparcialidad pretende garantizar que la prueba estandarizada sea justa, sin prejuicios ni sesgos (Gómez, Hidalgo y Guilera, 2010), de tal modo que las medidas que se obtienen de ella sean resultado de la comparación de un rasgo en condiciones de equidad contextual (Nunnally y Bernstein, 1995). El análisis de imparcialidad o carencia de sesgo, hace indispensable el reconocimiento escrupuloso de todas las variables que pueden inducir a respuestas no objetivas, con las que se producen medidas erróneas y apreciaciones injustas a personas de un grupo específico, en función de género, grupo etario, nivel socioeconómico, antecedentes culturales, pertenencia religiosa o étnica, entre otras. En ese sentido, los investigadores deben cuidar que el lenguaje, las situaciones y el contexto de los ítems no vulneren la dignidad de las personas, que no induzcan la movilización de rasgos latentes no previstos que pudieran favorecer que se movilicen actitudes positivas o negativas en ciertos grupos o individuos.

El análisis de sesgo debe hacerse a priori, al definir el objeto y las especificaciones de diseño de la prueba y a posteriori con técnicas estadísticas avanzadas para detectarlo, medirlo y realizar ajustes matemáticos de cambio de escala e igualación de los resultados obtenidos por los grupos potencialmente afectados por dicho sesgo. Es muy

acostumbrado entrar en un proceso tautológico utilizando un discurso subjetivo para explicar la falta de imparcialidad con base en valores de comparación o puntos de corte sin justificación objetiva, haciendo que las conclusiones estén igualmente sesgadas y, por lo tanto, carezcan también de imparcialidad.

Al ignorar que la objetividad requiere satisfacer estas propiedades se transfiere el problema a decidir si es válido utilizar un instrumento para fines distintos a los que motivan su diseño, si los resultados son válidos para determinado grupo, o si es válido hacer dictaminar a un individuo con los resultados de una prueba independientemente de sus consecuencias. Obsérvese que se acostumbra usar coloquialmente la palabra "válido" pero no en el sentido estricto de "validez", con lo que se confunden los propósitos y conceptos, haciendo que la validez -y no la objetividades- se asocie con el contexto cultural, con los usos y las consecuencias de la interpretación de los resultados (Messick, 1993-1995; Prieto y Delgado, 2003). Es de esperarse que la triada objetividad-validez-confiabilidad oriente el interés de los evaluadores hacia las implicaciones éticas, de equidad y de justicia. Como apuntan Kovač-Šebart y Krek (2009): "objetividad, validez y confiabilidad son categorías interconectadas e interdependientes, y todas ellas están incluidas en la percepción de la justicia".

5. Conclusiones

La objetividad incide, como se ha visto, en todos los factores y las etapas de la evaluación en general y en el desarrollo de una prueba estandarizada en particular. Puede decirse que, junto con la validez y la confiabilidad, forma una cadena interactiva, donde intervienen simultáneamente. Sin embargo debido a la necesidad de definir objetivamente el objeto de medida como primer elemento en el proceso de evaluación y como auxiliar en el desarrollo de la prueba, la objetividad es el primero de los atributos, solo a partir de ella es posible cuestionar si el instrumento es válido y confiable.

La objetividad debe verse como una brújula que orienta el desarrollo de un proyecto de evaluación, siendo al mismo tiempo la línea de horizonte hacia la cual debe caminarse de forma continua, debido a que es la única manera de garantizar que se cumple con los propósitos científicos de las pruebas estandarizadas. Negar la objetividad o relegarla a una posición diferente a ésta, genera confusión y ambigüedad en el desarrollo de una prueba, redundando en medidas con una validez potencialmente dudosa y una confiabilidad de interpretación poco clara, además de contribuir a configurar un contexto que puede incidir en uso inadecuado y poco ético de los resultados.

Las propiedades que resultan de los tres ejes teóricos utilizados en este trabajo permiten identificar los elementos indispensables de la objetividad, con ellos se puede llevar a cabo una vigilancia práctica en cada etapa del desarrollo de una prueba estandarizada. La tabla 2 incluye un ejemplo correspondiente a una prueba olímpica (patinaje artístico) que el lector podrá adaptar a otras aplicaciones.

Tabla 2. Propiedades de la objetividad en las pruebas estandarizadas (I)

PROPIEDAD	**1. ESPECIFICIDAD**
La prueba tiene este atributo si:	*Cuenta con una definición completa, pertinente, precisa del objeto, que lo distingue de otros*
Propósito en las pruebas estandarizadas	**Ejemplo**
1.1 Definir el objeto, modelo de medición, registro de los rasgos, análisis de datos y resultados del instrumento para que no se vean influidos por cualidades ajenas al objeto mismo. La aprehensión del objeto debe ser hecha con base en cualidades inherentes, en función de sus características, categorías, comparaciones, ponderaciones, valoraciones o medidas y arquetipos, entre otras formas.	Fuera de los aspectos reglamentarios y de la organización por categorías, la calificación debe hacerse con criterios asociados a la ejecución artística (belleza, gracia, estética de movimiento…) y los aspectos técnicos (cualidades de la carrera de frente, de espaldas, de los saltos…), pero no debe considerar nacionalidad, religión, grupo étnico o edad de los patinadores como criterio para ser asignada.
1.2 Distinguir claramente entre dos ideas contrastantes o hasta en conflicto respecto del objeto.	Dos jueces pueden explicar y justificar las calificaciones respecto de un patinador, reconociendo sus aciertos o errores.
1.3 Distinguir entre las características inherentes medibles del objeto y los requisitos no medibles construidos alrededor del mismo.	El reglamento establece claramente las categorías por género o por tipo de discapacidad para las competencias de patinaje.
1.4 Comparar las cualidades métricas de varios instrumentos, incluyendo el error de medida y la consistencia de resultados que se obtienen con una población focal dada.	Un modelo de facetas múltiples puede brindar medidas de habilidad de los patinadores en diversas ejecuciones de dificultad dada, de la severidad de los jueces y del error de medida de cada caso.

Fuente: Elaboración propia.

Tabla 3. Propiedades de la objetividad en las pruebas estandarizadas (II)

PROPIEDAD	**2. NEUTRALIDAD**
La prueba tiene este atributo si:	*No hay injerencia externa en los juicios de valor que hace un evaluador u otras personas con los resultados de una prueba estandarizada.*
Propósito en las pruebas estandarizadas	**Ejemplo**
2.1 Reducir o evitar la interpretación subjetiva del evaluador en un momento dado o a lo largo del tiempo, inducida por la fatiga o el número de juicios emitidos en una población numerosa).	El juez dispone de criterios para asignar calificaciones iguales al principio y al final de la competencia, comparables con calificaciones de otros patinadores en eventos previos.
2.2 Evitar o reducir la diferencia de apreciación entre dos evaluadores o entre el evaluador y el evaluado.	Las discrepancias entre jueces ante el desempeño de un patinador deben reducirse al mínimo. El patinador y su entrenador (u otra persona experta) deben percibir que la calificación emitida no difiere de lo que ellos mismos pueden juzgar.
2.3 Evitar que grupos específicos puedan verse favorecidas o perjudicadas por el diseño de la prueba o la apreciación del evaluador.	Un juez califica de forma más benévola a los patinadores de su mismo país para ayudarlos. Otro juez es más severo con los patinadores de su país para evitar que piensen que hace favoritismo.
2.4 Eliminar la discrepancia de opinión respecto de lo que se considera la respuesta correcta o la más aceptable, facilitando la calificación por personal no experto o por medio de un programa informático.	Las puntuaciones emitidas por los jueces deben ser verificables dentro de su orden de error. El público (persona no experta) puede reconocer que la calificación del patinador es aceptable siguiendo los mismos criterios y emitir calificaciones equiparables.

Fuente: Elaboración propia.

Tabla 4. Propiedades de la objetividad en las pruebas estandarizadas (III)

PROPIEDAD	**3. INDEPENDENCIA**
La prueba tiene este atributo si:	*Las medidas y juicios de valor no se ven influidas por otros rasgos, instrumentos o agentes, personales o contextuales.*
Propósito en las pruebas estandarizadas	**Ejemplo**
3.1 Permitir que la medida de cada persona no se vea influida por las medidas de las otras personas a las que se administra la prueba, ni tampoco por las características propias del instrumento utilizado.	Las calificaciones de los patinadores no deben darse en comparación con otro patinador sino respecto de los atributos de su desempeño.
3.2 Favorecer que la medida de cada ítem no se influya por las medidas de otros ítems incluidos en el instrumento, ni por las características de grupos específicos en los que se administra la prueba.	Las calificaciones de los desempeños artístico y técnico del patinador deben ser independientes entre sí.
3.3 Garantizar que el juicio que emite un evaluador no refleje la influencia u opinión de otro evaluador.	Cada juez emite la calificación del patinador sin ver las de los otros jueces.
3.4 Garantizar que el juicio que emite un evaluador no se vea influido por datos previos de cualidades del sujeto o del conjunto de personas a evaluar.	Cada patinador debe ser calificado sin tomar en cuenta su desempeño en un evento anterior.

Fuente: Elaboración propia.

Tabla 5. Propiedades de la objetividad en las pruebas estandarizadas (IV)

PROPIEDAD	**4. INDEPENDENCIA**
La prueba tiene este atributo si:	*Las medidas y juicios de valor no se ven influidas por otros rasgos, instrumentos o agentes, personales o contextuales.*
Propósito en las pruebas estandarizadas	**Ejemplo**
4.1 Emitir juicios de valor libres de sesgo sobre uno o varios rasgos o características inherentes del objeto mismo.	Los jueces emiten su calificación basados en el desempeño de los patinadores sin importar su género, país de procedencia, pertenencia étnica u otro aspecto ajeno al patinaje.
4.2 Eliminar en el evaluador el efecto de halo, de prejuicios o estereotipos.	El juez emite una calificación más favorable a los patinadores procedentes de países con mayor tradición en esta disciplina.
4.3 Otorgar a todas las personas evaluadas las mismas oportunidades para mostrar su desempeño ante un instrumento dado, previas adaptaciones por discapacidades u otra característica justificada.	Las reglas para calificar los elementos de una rutina de patinaje de pareja deben ser las mismas independientemente del género de los patinadores.

Fuente: Elaboración propia.

Tabla 6. Propiedades de la objetividad en las pruebas estandarizadas (V)

PROPIEDAD	**5. IMPERSONALIDAD**
La prueba tiene este atributo si:	*Explicita la forma en que el objeto es compartido en el grupo social, cultural, étnico u otro al que pertenece en un momento dado, considerando su evolución en el tiempo y en cada contexto.*
Propósito en las pruebas estandarizadas	**Ejemplo**
5.1 Evitar que personas específicas puedan verse favorecidas o perjudicadas en la prueba.	El juez no emite su calificación a partir de la trayectoria deportiva del patinador sino sobre el desempeño concreto observado.
5.2 Plasmar las características o rasgos de un objeto transparentando su análisis y aprehensión por diversas personas, incluyendo el evaluador y el evaluado, o un público independiente.	La apreciación del juez sobre las características técnicas de las piruetas está plenamente descrita en las reglas disponibles por el patinador, su entrenador y los diferentes jueces.
5.3 Validar los usos e interpretaciones a nivel contextual, cultural, grupal, o de otra índole, que se postulen a partir de datos obtenidos de la realidad.	La apreciación del juez sobre las características técnicas de una pirueta no debe verse modificada en función del origen étnico del patinador.

Fuente: Elaboración propia.

Incorporar la objetividad como atributo principal del proceso de evaluación es particularmente imprescindible en educación y ciencias sociales, no solamente para definir objeto a evaluar, sino por el uso de las pruebas estandarizadas de selección para ingreso a universidad o certificación profesional. Pocas veces se cita la objetividad junto con validez y confiabilidad en las pruebas estandarizadas, proliferando los detractores que objetan que sean "válidas" para evaluar a los estudiantes de ambiente rural, de etnias monolingües que no dominan el idioma nacional o que pertenecen a zonas deprimidas del país, sobre la base de que están en desventaja respecto de los estudiantes urbanos y de alto nivel socioeconómico, haciendo que la interpretación de sus resultados tenga implicaciones y consecuencias negativas para ellos. Debe quedar claro, por lo tanto, que no se trata de un asunto que pueda resolver la validez sino la objetividad, porque al usar una prueba en toda la población focal se tiene la ventaja establecer comparativos útiles para las políticas educativas y sociales del país, así como hacer interpretaciones diferenciadas entre grupos poblacionales.

La prueba PISA, promovida por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), cumple con altos criterios de validez y de confiabilidad, pero su objetividad es cuestionable debido a que, fuera de que usa ítems objetivos, no hace explicita su relación con este atributo. Entre las versiones de 2003 a 2015 (OECD, 2005-2016), solo se menciona en dos reportes nacionales (Eslovaquia y República Checa) vinculándola con la neutralidad y la imparcialidad para garantizar medidas objetivas sobre el desempeño (Santiago, Halász, Levacíc y Shewbridge, 2016; Shewbridge, Herczyński, Radinger y Sonnemann, 2016).

Alcanzar la objetividad en el proceso de evaluación junto con la validez y la confiabilidad permite disponer de pruebas mejor diseñadas, más robustas, donde las perfeccionadas herramientas de medición facultan tomar decisiones en beneficio de los individuos y de la sociedad en su conjunto.

Referencias

American Educational Research Association, American Psychological Association, National Council on Measurement in Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. Washington D. C.: Autor.

American Psychological Association. (1954). *Technical recommendations for psychological test and diagnostic techniques*. Washington D. C.: Autor.

American Psychological Association. (1966). *Standards for Educational and Psychological Test and Manual*. Washington D. C.: Autor.

American Psychological Association. (2010). *Ethical Principles for Psychologists and Code of Conduct*. Washington D. C.: Autor.

André, N. y Loye, N. (2015). La validité psychologique: Un regard global sur le concept centenaire sa genése ses avatars. *Mesure et Évaluation en Éducation, 37*(3), 125-148. doi:10.7202/1036330ar

Argibay, J. (2006). Técnicas psicométricas: Cuestiones de validez y confiabilidad. *Subjetividad y Procesos Cognitivos, 8*, 15-33.

Binet, A. (1910). Qu´est-ce qu´une émotion? Qu´est-ce qu´un acte intellectuel? *L´Année Psycologique, 17*, 1-47.

Blanco-Villaseñor, Á. (1991). La teoría de la generalizabilidad aplicada a los diseños observacionales. *Revista Mexicana de Análisis de la Conducta*, *17*(3), 23-53.

Bond, T. y Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Borsboom, D., Mellenbergh, G. J. y Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071.

Bowen, W. G. y Bok, D. (1998). *The shape of the river.* Princeton, NJ: Princeton University Press.

Campbell, D. (1960). Recommendations for APA test standards regarding construct, trait or discriminate validity. *American Psychologist*, *15*(8), 546-553.

Carvajal-Carrascal, G. (2012). Medición de fenómenos de enfermería: El reto de la validez y la confiabilidad en la investigación cuantitativa. *Aquichan*, *12*(1).

Céspedes, V. (2009). *Modelo conceptual del manejo del síntoma: Clasificación por percepción, evaluación y respuesta de mujeres con síndrome coronario agudo; originada por la construcción de un instrumento validado en Bogotá, Colombia* (Tesis doctoral, Universidad Nacional de Colombia, Bogotá).

Chan, E. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. En B. Zumbo y E. Chan (Eds.), *Validity and validation in social, behavioral and health sciences* (pp. 9-24). Nueva York, NY: Springer.

Cupani, A. (2011). Acerca de la objetividad científica. *Scientiae Studia*, *9*(3), 501-525. doi:10.1590/S1678-31662011000300004

Educational Testing Service. (1987). *Standards for quality and fairness. Adopted by the Board of Trustees.* Princeton, NJ: Autor.

Embretson, S. (2007). Construct validity: A universal validity system or just another test evaluation procedure. *Educational Researcher*, *36*(8), 449-455. doi:10.3102/0013189X07311600

Fox, J. y Glas, C. (2001). Bayesian estimation of a multilevel in model using Gibbs sampling. *Psychometrika*, *66*(2), 271-288. doi:10.1007/BF02294839

Fox, J. y Glas, C. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*(2), 169-191. doi:10.1007/BF02294796

Froese-Germain, B. (1999). *Standardized testing: Undermining equity in education. Report prepared for the National Issues in Education Initiative.* Ottawa: Canadian Teachers' Federation.

García, M. (1955). Objetividad en el conocimiento científico. *Revista Cubana de Filosofía*, *3*(12), 21-26.

Gaukroger, S. (2012). *Objectivity. A very short introduction.* Oxford: Oxford University Press.

Gómez, J. e Hidalgo, M. (2003). Desarrollos recientes en Psicometría. *Avances en Medición*, *1*(1), 17-36.

Gómez, J. e Hidalgo, M. (2005). La validez de los test, escalas y cuestionarios. *La Sociología en sus Escenarios*, *12*, 1-14.

Gómez, J., Hidalgo, D. y Guilera, G. (2010). El sesgo de los instrumentos de medición. Test justos. *Papeles del Psicólogo*, *31*(1), 75-84.

Guilford, J. P. (1987). Validity of measurements. En J. P. Guilford (Ed.), *Fundamental statistics in psychology and education* (pp. 424-458). Tokyo: McGraw-Hill - Kogakusha.

Herrenstein, R. J. y Murray, G. (1994). *The Bell Curve. Intelligence and class structure in American life.* Nueva York, NY: Simon y Schuster.

International Test Commission. (2014). *International guidelines on the security of tests, examinations, and other assessments.* Recuperado de www.intestcom.org

International Test Commission. (2016). *The ITC guidelines for translating and adapting tests.* Recuperado de www.intestcom.org

Jeffrey, M. (2003). *Test validation: A literature review,* 1-46. Florida, CA: University of Florida.

Jiménez, K. y Montero, E. (2013). Aplicación del modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemáticas. *Revista Digital Matemáticas, Educación e Internet, 13*(1), 1-24. doi:10.18845/rdmei.v13i1.1628

Jornet, J. y Suarez, R. (1996). Pruebas estandarizadas y evaluación del rendimiento: Usos y características métricas. *Revista de Investigación Educativa, 14*(2), 141-163.

Kane, M. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*(2), 76-82. doi:10.3102/0013189X08315390

Kelley, T. L. (1927). Proposes served by educational test. En T. Kelley (Ed.), *Interpretation of educational measurements* (págs. 18-43). Nueva York, NY: World Book Company Yorkers on Hudson.

Kerlinger, F. y Howard, L. (2008a). Confiabilidad. En F. Kerlinger y L. Howard (Eds.), *Investigación del comportamiento. Métodos de investigación en ciencias sociales (*pp. 581-602). Ciudad de México: McGraw Hill.

Kerlinger, F. y Howard, L. (2008b). Validez. En F. Kerlinger y L. Howard (Eds.), *Investigación del comportamiento. Métodos de investigación en ciencias sociales* (pp. 603-628). Ciudad de México: McGraw Hill.

Kimberlin, C. y Winterstein, A. (2008). Validity and reliability of measurement instruments used in research. *American Journal Health-System Pharmacy, 65*(1), 2276-2284. doi:10.2146/ajhp070364

Kovač-Šebart, M. y Krek, J. (2009). *Justice in the assessment of knowledge: The opinions of teachers and parents.* Cracovia: AFM Publishing House.

Lane, R. (1999). *Validity evidence for assessments. Reidy interactive lecture series.* Pittsburgh, PA: University Pittsburgh.

Larroyo, F. (1968). *El positivismo lógico. Pro y contra.* Ciudad de México: Editorial Porrúa.

Linacre, J. y Wright, B. (1995). *How do Rasch and 3P differ? MESA Laboratory.* Chicago, IL: Kimbark.

Linacre, J. (1994*). Many-facet, Rasch measurement, MESA Press.* Chicago, IL: Kimbark.

Lissittz, R. y Samuelsen, K. (2007). Dialogue on validity. A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448. doi:10.3102/0013189X07311286

Messick, S. (1993, abril). *Foundation of validity: Meaning and consequences in psychological assessment.* Comunicación presentada en el Second Conference of the European Association of Psychological Assessment, Groningen, Netherlands.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 1-8.

Mislevy, R. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469. Doi: 10.3102/0013189X07311660

Morales de Barbenza, C. (2001). Consideraciones acerca de la objetividad en evaluación psicológica. *Interdisciplinaria*, *18*(2), 169-178.

Moss, P. (2007). Reconstructing validity. *Educational Researcher*, *36*(8), 470-476.

Muñiz, J. (2010). Las teorías de los test: Teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, *31*(1), 57-66.

Muñiz, L., Elosua, P. y Hambleton, R. (2013). Directrices para la traducción y adaptación de los test: Segunda edición. *Psicothema*, *25*(2), 151-157.

Myers, A. y Hansen, C. H. (2002). *Experimental psychology*. Belmont, CA: Wadsworth Thomson Learning.

Newton, P. y Baird, J. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 173-177. doi:10.1080/0969594X.2016.1172871

Nunnally, J. y Bernstein, I. (1995). *Teoría psicométrica*. Ciudad de México: McGraw-Hill.

Organización para la Cooperación y el Desarrollo Económicos. (2005). *PISA 2003 technical report*. París: OECD Publishing.

Organización para la Cooperación y el Desarrollo Económicos. (2009). *PISA 2006 technical report*. París: OECD Publishing.

Organización para la Cooperación y el Desarrollo Económicos. (2010). *PISA 2009 results: Learning to learn – Student engagement, strategies and practices (Volume III)*. París: OECD Publishing. doi:10.1787/9789264083943-en

Organización para la Cooperación y el Desarrollo Económicos. (2016). *Equations and inequalities: Making mathematics accessible to all*. París: OECD Publishing. doi:10.1787/9789264258495-en.

Organización para la Cooperación y el Desarrollo Económicos. (2016). *Low-performing students: Why they fall behind and how to help them succeed*. París: OECD Publishing. doi:10.1787/9789264250246-en.

Padilla, J., Gómez, J., Hidalgo, M. y Muñiz, J. (2006). La evaluación de las consecuencias del uso de los test en la teoría de la validez. *Psicothema*, *18*(2), 307-312.

Phelps, R. P. (2005). *Defending standardized testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Prieto, G. y Delgado, A. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, *15*(1), 94-100.

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, *31*(1), 67-74.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

Real Academia Española. (2016). *Diccionario de la Lengua Española*. Recuperado de http://dle.rae.es/?id=QmvS5XH

Ritter, N. (February, 2010). *Understanding a widely misunderstood statistic: Cronbach´s alpha*. Comunicación presentada en el Annual Meeting of the Southwest Educational Research Association, Nueva Orleans. Recuperado de http://files.eric.ed.gov/fulltext/ED526237.pdf

Sánchez-Meca, J., López-Pina, J. y López, J. (2009). Generalización de la fiabilidad: Un enfoque meta analítico aplicado a la fiabilidad. *Fisioterapia*, *31*(6), 262-270.

Santiago, P., Halász, G., Levacíc, R. y Shewbridge, C. (2016) *Reviews of school resources: Slovak Republic*. París: OECD Publishing. doi:10.1787/9789264247567-en

Sattler, J. (2010). Niños de minorías étnicas. En J. Sattler (Ed.), *Evaluación infantil. Fundamentos cognitivos* (pp. 134-181). Ciudad de México: Manual Moderno.

Shavelson, R. y Webb, N. (2005). *Generalizability theory*. Newbury Park, CA: Sage Publications.

Shewbridge, C., Herczyński, J., Radinger, T. y Sonnemann, J. (2016). *OECD reviews of school resources: Czech Republic 2016*. París: OECD Publishing. doi:10.1787/9789264262379-en

Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, *9*, 167-194. doi:10.1080/15305050903106883

Sireci, S. (2007). On validity theory and test validation. *Educational Research*, *36*(8), 477-481. doi:10.1002/9781118445112.stat06403

Sireci, S. y Padilla, J. (2014). Validating assessment: Introduction to the special section. *Psicothema*, *26*(1), 97-99. doi:10.7334/psicothema2013.255

Tristán, A. (1996). *Nota 5: Contribución al estudio del error de medida. Kalt Criterial. Un programa de la familia Kalt. Versión 2. Guía de usuario*. San Luis Potosí: IEIA.

Tristán, A. (2010). *Theoretical Alpha values for objective test*. Recuperado de https://www.coreprojects.org/PROMIS/PROMIS2/Sandbox/Presentations/Tristan-AlphaPresentation.pdf

Wright, B. y Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement*, *1*(1), 84-106.

Wright, B. y Stone, M. (1999). *Measurement essential*. Wilmington, DE: Wide Range.

Zamora, E. (2007). *Evaluación objetiva de la calidad sensorial de los alimentos procesados*. La Habana: Editorial Universitaria.

Zumbo, B. (2009). Validity as contextualized and pragmatic explanation, and implication for validation practice. En R. Lissitz (Ed), *The concept of validity, revisions, new directions and applications* (pp. 65-82). Charlotte, NC: Information Age Publishing.

Zúñiga, M. y Montero, E. (2007). Teoría G. Un futuro paradigma para el análisis de pruebas psicométricas. *Anualidades en Psicología*, *21*(108), 117-144. doi:10.15517/ap.v21i108.29

Breve CV de los autores

Agustín Tristán López

Doctor en Ingeniería por la *École Nationale des Ponts et Chaussées*, París, Francia. Director General del Instituto de Evaluación e Ingeniería Avanzada, S.C. Asesor en psicometría y evaluación educativa, responsable de proyectos de certificación en docencia y en el área profesional para Colegios de Profesionales en las áreas de la Salud e Ingenierías. Su interés principal se centra en desarrollo de sistemas de medición, diseño de modelos matemáticos y estadísticos con teoría clásica y modelos logísticos y de Rasch. Autor de más de 30 productos de software para evaluación en educación y salud. Cuenta con más de 40 publicaciones en el tema de evaluación educativa. Email: atristan@ieia.com.mx. Sitio web: www.ieia.com.mx.

Nancy Yahibé Pedraza Corpus

Doctora en Estudios de Población, Centro de Estudios Demográficos, Urbanos y Ambientales, de El Colegio de México. Responsable Psicopedagógica del Instituto de Evaluación e Ingeniería Avanzada, S.C. Asesora y da seguimiento a diversos sistemas de evaluación educativos y procesos de certificación profesional. Especializada en diseño de pruebas e ítems objetivos para evaluar competencias con énfasis en aspectos sociales, actitudinales y psicológicos. Su interés principal se centra en el desarrollo de sistemas de medición, y en el análisis del comportamiento y la evaluación de sistemas educativos. Email: nancypedraza@ieia.com.mx. Sitio Web: www.ieia.com.mx.