

Principios y suposiciones de la evaluación psicométrica

Gavin T. L. Brown

The University of Auckland, New Zealand <http://orcid.org/0000-0002-8352-2351> gt.brown@auckland.ac.nz

Citar como: Brown, G. (2023). Principios y suposiciones de la evaluación psicométrica. *Revista Digital de Investigación en Docencia Universitaria*, 17(2), e2022. <https://doi.org/10.19083/ridu.2023.1834>

Recibido: 13/12/21. **Revisado:** 24/02/22. **Aceptado:** 21/06/22. **Publicado:** 30/06/2023.

Resumen

Introducción: la validez de las afirmaciones sobre los autoinformes depende de la credibilidad de las medidas utilizadas. Se espera que los inventarios desarrollados dentro de la tradición psicométrica proporcionen evidencia teórica y empírica de la validez y confiabilidad de las medidas para respaldar las interpretaciones y decisiones posteriores. Estas prácticas dependen de un conjunto de suposiciones basadas en la teoría del rasgo latente que son esenciales para comprender lo que pueden hacer las medidas psicométricas. Este documento describe las características esenciales del diseño y el informe de datos de autoinforme psicométrico.

Palabras clave: autoinforme, métodos psicométricos, validez, fiabilidad, evaluación

Principles and assumptions of psychometric measurement

Abstract

Introduction: the validity of claims about learner self-reports depends on the credibility of the measures used. Inventories developed within the psychometric tradition are expected to provide theoretical and empirical evidence for the validity and reliability of the measures to support subsequent interpretations and decisions. These practices depend on a set of assumptions based on latent trait theory that are essential to understanding what psychometric measures can do. This tutorial outlines essential characteristics of design and reporting of psychometric self-report data.

Keywords: self-report; psychometric methods; validity; reliability; evaluation

Introducción

La toma de decisiones sobre instrucción, intervención o tratamiento requiere información de alta calidad sobre el estado actual (fortalezas y debilidades), tendencias de cambio (si las hubiera) y el impacto del precursor y la introducción de factores causales. Asimismo, se debe entender el mérito, el valor o el valor de los valores obtenidos en la información para liderar decisiones, acciones y consecuencias apropiadas (Scriven, 1967). Las intervenciones educativas y clínicas requieren información sólida de parte de los supuestos beneficiarios de esas intervenciones (i.e., estudiantes, clientes). Si buscamos cambiar el conocimiento, las habilidades, las actitudes, las creencias y comportamientos, necesitamos medidas

***Correspondencia:**

Gavin T. L. Brown
gt.brown@auckland.ac.nz

Revista Digital de Investigación en Docencia Universitaria, 17(2)

e-ISSN: 2223-2516 © Universidad Peruana de Ciencias Aplicadas

<https://doi.org/10.19083/ridu.2023.1834>



precisas de dónde están las personas y adónde llegaron luego de la intervención (Messick, 1989). Por lo tanto, todos los métodos de muestreo sistemático de habilidades, actitudes, capacidades o cualquier otra característica de los individuos son considerados 'pruebas' (Cizek, 2020).

La psicometría, entonces, es una ciencia estadística detrás de la evaluación y medición de las emociones, pensamientos, actitudes, creencias, opiniones, ideas y conocimiento humanos latentes que moldean las conductas manifiestas. Debido a que estos fenómenos existen detrás de los ojos, entre las orejas, y al interior de la existencia humana, es esencial que el individuo brinde esta información al examinador o evaluador. Si bien esta información puede ser subjetiva debido a ilusiones optimistas sobre uno mismo (Dunning et al., 2004), en realidad, no existe ninguna opción sino obtener y explorar la información brindada por el individuo sobre el estado de sus sentimientos, cogniciones e intenciones. A diferencia de la investigación de fenómenos físicos o animales, el ser humano es capaz de conocer y comunicar rasgos y estados internos. Si bien las personas cercanas (padres, hermanos, hijos, parejas) pueden tener conocimientos privilegiados sobre la mente y la conducta de un individuo, quien posee la información más actualizada y completa es el individuo en sí. Los comportamientos observables pueden brindar información muy útil sobre la mente y la voluntad del individuo; sin embargo, las creencias intenciones o motivaciones opuestas pueden llevar a acciones observadas muy parecidas. De esa manera, entender lo que hay en la mente humana depende de la obtención de información del dueño de esa mente. A parte de los animales excepcionales que han aprendido a comunicarse a través del lenguaje de señas, solo el ser humano parece ser capaz de tener algo que decir que los demás no pueden saber, a menos que la propia persona revele esa información. Por lo tanto, la dependencia de la investigación psicométrica al autoinforme es necesaria e inevitable. Sin embargo, la confiabilidad y validez de herramientas que miden estos fenómenos psicológicos debe ser demostradas para que sus resultados se consideren creíbles.

En este artículo, ofreceré un panorama de alto nivel o una guía para la teoría y los principios psi-

cométricos utilizados para evaluar mediciones de conocimientos, creencias, y actitudes o intenciones humanas. Alentamos a los lectores interesados en entender de manera detallada los principios y prácticas psicométricas a leer manuales autorizados sobre psicometría (Rao & Sinharay, 2007), medición educativa (Brennan, 2006), o pruebas (Geisinger, 2013).

Teoría y práctica psicométricas

Teoría de las suposiciones latentes

La teoría psicométrica se basa en la suposición de que las conductas manifiestas son explicadas o causadas por la variación en los atributos psicológicos latentes (cognitivos, emocionales, volitivos, intencionales, etc.) en el individuo (Borsboom, 2005). Estos pueden ser denominados pensamientos, creencias, actitudes o cualquier número de factores invisibles, no materiales, pero causales, que generan comportamientos observables de manera consistente. Esta suposición surge de nuestra incapacidad para leer los fenómenos psicológicos de los demás sin la capacidad de comunicarnos verbalmente (Corballis, 2002) sobre estos temas. El hecho de poder hablar sobre fenómenos psicológicos (por ejemplo, sentimientos, pensamientos, creencias, etc.) y creer que entendemos lo que otros están hablando sugiere fuertemente que pensamos que estas cosas existen. Por lo tanto, la teoría del rasgo latente sobre los fenómenos psicológicos parece ser una conclusión lógica de la naturaleza de nuestra existencia compartida.

Los orígenes de las características estadísticas y matemáticas de la teoría del rasgo latente remonta a la aplicación de técnicas de correlación de Spearman (correlación de orden de rangos, corrección por atenuación y estimación de la confiabilidad de la prueba) a patrones en puntajes de evaluaciones para inferir la idea de habilidades mentales generales ('g') responsables por variaciones compartidas entre puntajes (Clauser, 2022). Se supone que la capacidad mental general o la inteligencia explican por qué los estudiantes que sobresalen en una materia tienden a sobresalir también en otras. No sorprende, entonces,

que la investigación sobre la naturaleza de la inteligencia haya avanzado hacia modelos multidimensionales y descrito la influencia causal de factores genéticos y ambientales (Deary, 2001).

El modelo matemático de la teoría del rasgo latente asume que la habilidad subyacente detrás del desempeño del examinado requiere solo una suposición. Lord (1953) describe esta suposición como: “el rasgo o habilidad en discusión puede ser concebido como una variable representada numéricamente en una sola dimensión” (cursivas en el original, p. 518). Por extensión, los investigadores que evalúan la orientación subyacente de un individuo hacia un fenómeno mental específico asumen que la causa latente tiene características unidimensionales para que los individuos puedan ser ubicados según su frecuencia, intensidad, importancia, probabilidad o valencia para ese constructo (Allport, 1935). El uso difundido de la escala de Likert's (1932) en la psicología social supone que existe una causa latente que explica la fortaleza y la covarianza de las respuestas a indicaciones, preguntas o ítems.

Se ha argumentado que los factores latentes no hayan sido comprobadamente cuantitativos (Michell, 1999). Debido a que la evidencia para constructos latentes solo se puede obtener de operaciones manifiestas (por ejemplo, respuestas a preguntas), la investigación sobre la forma en que las personas se perciben a sí mismas o a los fenómenos siempre debe basarse en la suposición de que el rasgo subyacente es algo que se puede cuantificar. Que nuestras experiencias del mundo y nuestras habilidades para responder a ese mundo sean diferentes no es una suposición, sino un hecho. Por lo tanto, es relativamente fácil aceptar que todas las personas varíen dentro de las poblaciones en torno a valores normativos para fenómenos psicológicos tan diversos como la naturaleza del amor de una persona por su propia madre, su capacidad para resolver problemas, su capacidad para usar lenguaje (s), etc. Consecuentemente, como indica Michell (2008), la investigación de la teoría del rasgo latente debe admitir que “Actualmente, no sabemos si esta hipótesis [modelos cuantitativos de entidades psicológicas] es verdadera, pero la asumiremos, reconociendo que en algún momento en el futuro se deberá investigar” (p. 12).

En consecuencia, las evaluaciones de un constructo latente constituyen un intermediario en procesos cognitivos que son psicológicamente multidimensionales (Rupp, 2008). Por lo tanto, tal como Rupp (2008) indica:

Podría ser realmente significativo, y no apenas estadísticamente conveniente, resumir información a través de un modelo de variable latente unidimensional. La variable latente resultante puede servir como una herramienta de comunicación más efectiva y pragmática que represente un compromiso decididamente consciente entre la fidelidad cognitiva, la viabilidad empírica y la practicabilidad utilitaria. (p. 122; énfasis añadido).

Esto demuestra que nuestra evaluación de un constructo psicológico, independiente de un enfoque matemático, siempre supone que hemos utilizado un conjunto de indicadores válidos o teóricamente sólidos que, de acuerdo con la falibilidad científica (Popper-, 1945), puede resultar equivocado.

Una ventaja de las suposiciones cuantitativas es que los valores numéricos, en especial cuando son obtenidos de grandes poblaciones, podrán comportarse como una variable aleatoria de valor real, de modo que las manipulaciones matemáticas (correlaciones, media, desviación estándar, etc.) proporcionan información sobre la variación dentro de las poblaciones con relación al rasgo latente. Deben tomarse decisiones sobre cómo se analizan estadísticamente los números creados en una evaluación, que varían desde la teoría clásica de las pruebas hasta la teoría de la respuesta al ítem (Embretson & Reise, 2000). Sin embargo, la posición de un individuo sobre la norma social es un factor importante en la teoría de la conducta planificada, para determinar la intencionalidad y el comportamiento (Ajzen, 1991).

Frente a la posibilidad de que los constructos latentes no sean cuantitativos, debemos aceptar que nuestras mediciones no poseen valores cero o incrementos de escala inherentes. La naturaleza arbitraria que utilizamos al medir fenómenos psicológicos no invalida la existencia del rasgo. Por ejemplo, las pruebas que miden la inteligencia tienden a establecer la media de la población en 100 y la desviación estándar en 15. Esta es una

forma conveniente de describir y ubicar sujetos con relación a otros. Este valor medio de 100 podría haber sido fácilmente 500 o 1000, con una desviación estándar concomitante de 50 o 100. El significado de cualquier puntaje relativo a esos valores normales nos permitiría planificar intervenciones educativas y clínicas de manera pragmática. De esa manera, las matemáticas de la teoría del rasgo latente cuando se combinan con un marco teórico para la naturaleza y función de ese rasgo latente nos permiten entender y responder a las necesidades y fortalezas.

Validación externa

El hecho de que la teoría latente de los fenómenos psicológicos sea pragmática no significa que el constructo latente importe. Es necesaria la observación de las relaciones sistemáticas entre la variación en el constructo latente y en comportamientos que resultados observables para establecer si un hipotético rasgo latente es importante. La evidencia de comportamientos observables o resultados generados independientemente supera el sesgo del autoinforme como el único recurso de información. La validación estándar de oro de los puntajes auto informados debe ser comparada con otras medidas que son teóricamente similares o diferentes (convergencia y divergencia; [Campbell & Fiske, 1959](#)). El puntaje de un autoinforme altamente correlacionado con una medida desarrollada previamente de un constructo relacionado ofrece evidencia convergente. En contraste, las correlaciones bajas con una medida de un constructo completamente diferente brindan evidencia divergente sobre la importancia de la medición. De la misma manera, la divergencia a lo través del tiempo o del informante puede cuestionar la validez del autoinforme. Idealmente, las medidas convergentes y divergentes evitarán la debilidad metodológica del autoinforme, que es el objetivo del enfoque de múltiples rasgos y múltiples métodos. La información de la evidencia biométrica (fMRI, por ejemplo), el comportamiento en línea, los puntajes de las pruebas, la asistencia y otras evaluaciones físicas mostrarán si la variación de los puntajes está significativamente relacionada (es decir, causal o explícitamente) con los constructos que deberían ser sensibles al

constructo ([Borsboom et al., 2004](#); [Zumbo, 2009](#)).

La neuropsicología moderna intenta encontrar órganos específicos del cerebro que mapeen los procesos psicológicos (por ejemplo, la activación de neuronas espejo se relaciona al entendimiento de los demás; Kaplan & Iacoboni, 2006). Una investigación fMRI reveló que, cuando los participantes reciben una retroalimentación falsa sobre su desempeño, las regiones del cerebro asociadas a emociones negativas (como la corteza cingulada posterior, el giro frontal medio y el lóbulo parietal inferior, por ejemplo) fueron activadas al brindar retroalimentación normativa referencial a participantes de baja competencia y también cuando se brindó retroalimentación referencial basada en criterios a participantes de alta competencia ([Kim, Lee, Chung, & Bong, 2010](#)). Asimismo, los puntajes objetivo de los enfoques de desempeño se correlacionaron positivamente al activarse las áreas cerebrales de emoción negativa durante la retroalimentación normativa referencial.

Una revisión de las investigaciones basadas en electroencefalogramas (EEG) identificó un fuerte énfasis en las áreas de emoción negativa del cerebro en torno a la negatividad relacionada con la retroalimentación ([Meyer et al., 2021](#)), pero sugirió que debido a que las múltiples regiones del cerebro están involucradas en esos procesos y a las limitaciones de los electroencefalogramas, se necesitan modificaciones para entender cómo la retroalimentación se relaciona con el cerebro. La complejidad del cerebro puede ser apreciada en la forma en que las representaciones visuales de objetos y personas familiares están ubicadas en múltiples áreas del cerebro ([Quiroga et al., 2005](#)). Si bien los estudios de laboratorio que fMRI o EEG pueden revelar la forma en que el cerebro interactúa con la mente, el desafío es conocer cómo esas investigaciones se relacionan con las complejidades del funcionamiento en los ambientes del mundo real. Los resultados difieren cuando los organismos son estudiados en probetas de laboratorio (in vitro, por ejemplo) y cuando son liberados a la población (in vivo) ([Autoimmunity Research Foundation, 2012](#)); sin mencionar cómo se comportarían en un entorno cibernético o simulado (in sílice). Aún falta evaluar hasta qué

punto las suposiciones humanas auto informadas sobre retroalimentación mapean la actividad cerebral antes o durante la retroalimentación.

Un enfoque bien establecido para determinar la validez de las medidas de un constructo psicológico es el análisis de cómo los autoinformes individuales se relacionan a los resultados de las evaluaciones. Esto está bien establecido en los sistemas de evaluaciones educativas, tales como las encuestas del Programa Internacional de Evaluación de Estudiantes (PISA) de la OCDE. [Marsh et al. \(2006\)](#) mostraron que los constructos psicológicos auto informados (intereses auto informados, auto concepto, auto eficacia) recopilados en la encuesta PISA aplicada a más de 100,000 estudiantes en el 2000 tenían una importancia estadística pero modesta ($.25 < r < .35$), con invarianza en 25 jurisdicciones diferentes. Dos aspectos necesitan ser mencionados sobre estos resultados. En primer lugar, las relaciones son consistentes con las expectativas teóricas de cómo funcionan estos constructos psicológicos. En segundo lugar, el efecto es modesto, en parte porque los contextos in vivo son muy complejos y porque los individuos pueden tener variabilidad en los constructos que respaldan. En efecto, el análisis de conglomerados de variables motivacionales en las evaluaciones de rendimiento en matemáticas mostraron que, en casi todos los 12 países analizados a través de tres oleadas de información, hubo individuos que no tenían puntajes auto informados consistentes en los tres puntajes motivacionales y el desempeño dependió más o menos de la combinación de motivaciones ([Michaelides et al., 2019](#)).

No obstante, los puntajes auto informados sobre los fenómenos psicológicos del individuo es un dominio tenso. El ser humano no solo sufre de problemas de memoria sobre sus experiencias ([Schacter, 1999](#)), sino también sufre de ignorancia sobre sí mismo y sus competencias ([Dunning, Heath & Suls, 2004](#)), en parte debido a que ser honesto sobre sus deficiencias o fracasos puede amenazar su ego ([Boekaerts & Corno, 2005](#)); o entre adolescentes puede ser 'divertido' subvertir los cuestionarios ([Fan et al., 2006](#)). Para superar estas amenazas, la psicometría propone un número de métodos que se describen a continuación.

Fundamento teórico

El campo de la psicometría ha debatido extensamente sobre cómo validar la evidencia en cualquier evaluación de los fenómenos psicológicos ([Kane & Bridgeman, 2022](#)). Antes de [Messick \(1989\)](#), la validez solía ser pensada en términos de tipos múltiples (es decir, rostro, contenido, concurrente, constructo y predictivo). Sin embargo, el entendimiento contemporáneo es que la validez es un concepto unitario entendido como 'validez de constructo' ([Cizek, 2020](#)). La evidencia de un grado de juicio de validez ('preponderancia de la evidencia', 'evidencia clara y convincente', 'evidencia substancial', por ejemplo; [Cizek, 2020, p. 26](#)) se adquiere considerando varios argumentos empíricos y teóricos para la interpretación propuesta de una medición ([Cizek, 2020; Kane, 2006; Messick, 1989](#)).

La evidencia de validación incluye cualidades explícitas y teóricas de la herramienta o instrumento de evaluación; los elementos del estímulo, las indicaciones o tareas presentadas para obtener respuestas necesitan alinearse teóricamente con las definiciones teóricamente informadas de los expertos de un dominio que incluye hipótesis sobre la manera en que el constructo influirá en la respuesta ([American Educational Research Association et al., 2014; Schmeiser & Welch, 2006](#)). Esto significa que se debe invertir un esfuerzo considerable para especificar el dominio, el desarrollo y la medición de la coherencia de los diferentes estímulos e indicaciones utilizados para obtener respuestas de parte de los individuos a esa especificación teórica. Estudios piloto ([International Test Commission, 2018](#)), paneles de juicio de expertos ([McCoach et al., 2013](#)), estudios 'think aloud' de los participantes ([van Someren et al., 1994](#)), y entrevistas cognitivas ([Karabenick et al., 2007](#)), entre otras, son utilizadas para demostrar que existe evidencia de que el instrumento está alineado, en principio, con lo que se pretende evaluar.

Los informes de cómo se administra esa prueba o conjunto de evaluaciones y los tipos de datos recogidos son esenciales para brindar seguridad de que los protocolos puedan ser replicados y teóricamente de acuerdo con el dominio. Las evaluaciones que miden el modelo teórico preferido contra explicaciones alternativas también of-

recen evidencia de que las escalas son eficaces ([Cronbach, 1988](#)). De esa manera, la evaluación de la estructura interna de un inventario debe incluir múltiples alternativas competentes.

Estructura interna

La industria psicométrica adopta un enfoque científico en el que la recolección de datos debe generar patrones consistentes de respuesta entre los individuos de acuerdo con sus diferentes respuestas frente a un fenómeno. Una restricción constante en la evaluación de constructos psicológicos es que, en sí mismos, no son directamente observables; son latentes ([Borsboom, 2005](#)). Como tales, los indicadores para múltiples causas (MIMIC) son utilizados para disminuir el error al calcular la fortaleza y la dirección de la actitud, la creencia o el valor, y para representar mejor el fenómeno de interés ([Jöreskog & Goldberger, 1975](#)). Con muestras suficientes y evaluaciones diseñadas teóricamente, el modelado matemático de patrones de respuesta MIMIC (estimación de confiabilidad interna, análisis factorial) es utilizado con el fin de crear evidencia para la estructura y confiabilidad de las escalas o factores propuestos ([Haertel, 2006](#)).

Los enfoques que proporcionan evidencia sobre los patrones de respuesta incluyen la estimación de la confiabilidad de la escala, que se puede calcular de diversas maneras. A pesar de que muchos investigadores están familiarizados con el alfa de Cronbach's (1951), una extensa investigación indica que el omega de [McDonald's \(1999\)](#) y el coeficiente H de [Hancock and Mueller's \(2001\)](#) son métodos superiores para establecer la confiabilidad de un conjunto de elementos. Dentro del marco del modelado de Rasch, la separación de elementos G de [Wright and Stone's \(1999\)](#) puede ser utilizado para afirmar que los ítems provocan respuestas acordes con la versión de la teoría de respuesta al ítem. Los investigadores deben ser conscientes que se pueden obtener resultados de confiabilidad muy alta (por ejemplo, donde $\alpha > .90$) a través de la escritura de ítems que son casi idénticos en el fraseo (es decir, tienen una alta homogeneidad) y producen un 'específico inflado' ([Cattell & Tsujioka, 1964](#)).

Mientras que el análisis de componentes principales puede identificar vectores subyacentes en una matriz de datos, la teoría psicométrica, con su énfasis en el error, se basa en el modelo de factor común para identificar factores compartidos y únicos que subyacen a los mismos datos ([Bryant & Yarnold, 1995](#)). Existen criterios convencionales para guiar la interpretación de resultados estadísticos ([Bandalos & Finney, 2010](#)) de modo que se pueda evaluar la calidad de la evidencia para la estructura interna de una herramienta de investigación o medición (es decir, se pueden ignorar las escalas con poca evidencia, mientras que se pueden usar aquellos con evidencia sólida). Una vez que se establece una medición sólida de los constructos latentes, se pueden derivar puntajes para cada factor ([DiStefano, Zhu, & Mindrila, 2009](#)).

Con puntajes de escala, se puede distribuir sujetos y grupos según sus puntajes, lo que permite comparar un constructo con otro y comparar las diferencias de puntajes entre grupos y a lo largo del tiempo. Estas técnicas analíticas están fuertemente asociadas con el campo de la psicometría debido a que abordan temas relacionados con demostrar estadísticamente que las expectativas teóricas han sido alcanzadas. Estas prácticas ayudan a crear un argumento sobre la confiabilidad de la información obtenida por humanos sobre sí mismos, como también para la interpretación y utilización de esos datos.

Replicabilidad

En el espíritu de la investigación científica, la replicación de estudios puede analizar la estabilidad de las propiedades psicométricas a través de muestras ([Makel et al., 2012](#)). Las técnicas estadísticas como el análisis factorial confirmatorio multigrupo permiten a los investigadores establecer si los modelos estadísticos de cómo los participantes responden a un instrumento varían dependiendo si la nueva muestra es obtenida desde la misma población o no. Idealmente, las características psicométricas del instrumento de medición deberían ser posibles cuando son aplicadas en una nueva muestra obtenida de la misma población. Claramente, se debe esperar la no invarianza cuando las muestras son de poblaciones

divergentes. Esto no significa que las mediciones estén rotas; por el contrario, ello sugiere que las mediciones funcionan de manera diferente o que el constructo que está siendo evaluado difiere según el idioma, la edad, la cultura, la prosperidad o los límites educativos. Considere la no invarianza encontrada en el inventario Teacher Conceptions of Feedback entre Nueva Zelandia y Lousiana, que cuentan con diferentes marcos de política ([Brown, Harris, O'Quin & Lane, 2017](#)). Las mediciones realizadas con nuevas muestras tienen más oportunidad de generar evidencia de validación, al superar artefactos aleatorios asociados con el desarrollo de la medición. Considere la similitud del inventario del Teacher Conceptions of Assessment en Nueva Zelandia y su falta de invarianza a lo largo de jurisdicciones y lenguajes ([Brown, Gebril, & Michaelides, 2019](#)).

Un efecto secundario desafortunado del énfasis en la complejidad de los modelos estadísticos y matemáticos de la estructura interna de una medición es que la evidencia de validación a partir de medidas externas tiende a ser ignorada. Claramente, es difícil recoger evidencia de muestras independientes, de pedir a los participantes que completen evaluaciones paralelas o divergentes al mismo tiempo como una nueva, y es aún más difícil recoger evidencia independiente de un comportamiento para construir un caso sólido en que la nueva medición no solo es psicométricamente robusta, sino que también posee evidencia teórica y empírica de la relación entre lo que la mente reporta y lo que se puede observar desde el exterior. Sin embargo, sin la evidencia estadística y matemática de cómo funciona una nueva medición, no será posible evaluar las afirmaciones teóricas sobre cómo piensan, sienten y se comportan los humanos. Asimismo, dada la complejidad de factores que inciden en el rendimiento, el efecto real de cualquier percepción específica puede ser muy pequeño. Cuando los efectos son pequeños y responden a las condiciones ambientales, son intrínsecamente difíciles de replicar ([Lindsay, 2015](#)).

Desafíos

Un problema fundamental dentro de la psicología es que todo en las experiencias de vida, ambientes y la fisiología de los individuos influye

en lo que ellos piensan, sienten, creen, dicen o hacen. Por lo tanto, no es de sorprender que el impacto de cualquier constructo psicológico único sea relativamente pequeño por interactuar con otras cosas que también importan. Desafortunadamente, los esfuerzos por aislar y entender factores psicológicos importantes en la vida humana han llevado, a generalizar el 'tintineo' (o sea, las mismas palabras con diferentes significados o los mismos significados con diferentes palabras) en el ámbito ([Flake & Fried, 2020](#)). Sin embargo, la teoría del rasgo latente nos ofrece un camino hacia la mente, el corazón, y las representaciones mentales de los individuos.

Una buena evidencia psicométrica para las mediciones de cualquier constructo psicológico debería ser capaz de:

1. Demostrar fidelidad a una descripción teóricamente sólida de lo que es el constructo, cómo funciona, y cómo debe hacer;
2. Ofrecer evidencia de que la operacionalización propuesta tiene credibilidad, en principio, frente a esa teoría;
3. Demostrar evidencia estadística sólida de la coherencia de los elementos contra el diseño del constructo de un modelo de medición;
4. Brindar evidencia de que las mediciones pueden ser reproducidas a partir de muestras adicionales;
5. Ofrecer evidencia de que las mediciones producen efectos en otras mediciones (incluyendo a los auto-informes), comportamientos o resultados alineados con las expectativas teóricas; y
6. Proporcionar evidencia de que el constructo puede ser manipulado y que los puntajes de evaluación cambian y tienen efectos propuestos teóricamente.

Los lectores pueden sentirse consternados ante la idea de que cualquier informe debe necesariamente lograr todas estas cosas. Sin embargo, un excelente ejemplo de cómo se pueden abordar estas preocupaciones es visible en [Thielsch and Hirschfeld \(2019\)](#) quienes en siete investigaciones ofrecieron: un marco teórico, desarrollo de conjuntos de elementos, demostración estadística de propiedades de escala o factor, confiabilidad demostrada de

evaluaciones y reevaluaciones. También validaron las escalas contra otras mediciones convergentes y divergentes, puntajes manipulados experimentalmente, y crearon normas de muestras más amplias. En realidad, los usuarios de cualquier evaluación psicométrica deben esperar evidencia de este tipo antes de decidirse por la utilización de una nueva herramienta o inventario.

References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A Handbook of Social Psychology* (pp. 798-844). Clark University Press.
- American Educational Research Association, American Psychological Association, National Council for Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Autoimmunity Research Foundation. (2012). *Differences between in vitro, in vivo, and in silico studies*. The Marshall Protocol Knowledge Base. Retrieved 12 November from <http://mpkb.org/home/patients/assessing-literature/in-vitro-studies>
- Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 93-114). Routledge.
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An international review*, 54(2), 199-231. <https://doi.org/10.1111/j.1464-0597.2005.00205.x>
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Praeger.
- Brown, G. T. L., Gebriel, A., & Michaelides, M. P. (2019). Teachers' Conceptions of Assessment: A Global Phenomenon or a Global Localism. *Frontiers in Education*, 4(16). <https://doi.org/10.3389/feduc.2019.00016>
- Brown, G. T. L., Harris, L. R., O'Quin, C., & Lane, K. E. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research & Method in Education*, 40(1), 66-90. <https://doi.org/10.1080/1743727X.2015.1070823>
- Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). APA.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin*, 56(2), 81-105.
- Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, 24(1), 3-30. <https://doi.org/10.1177/001316446402400101>
- Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.
- Clauser, B. E. (2022). A history of classical test theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 157-180). Routledge.
- Corballis, M. C. (2002). *From Hand to Mouth: The Origins of Language*. Princeton University Press.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Lawrence Erlbaum.
- Deary, I. J. (2001). *Intelligence: A Very Short Introduction*. Oxford University Press.
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20). <https://doi.org/10.7275/da8t-4g52>
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and

- the workplace. *Psychological Science in the Public Interest*, 5(3), 69-106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. LEA.
- Fan, X., Miller, B. C., Park, K.-E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, 18, 223-244. <https://doi.org/10.1177/152822X06289161>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>
- Geisinger, K. F. (Ed.). (2013). *APA Handbook of Testing and Assessment in Psychology*. American Psychological Association.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Praeger.
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking Construct Reliability Within Latent Variable Systems. In R. Cudeck, S. d. Toit, & D. Sorbom (Eds.), *Structural Equation Modeling: Present and Future - A Festschrift in Honor of Karl Joreskog* (pp. 195-216). Scientific Software International Inc.
- International Test Commission. (2018). ITC Guidelines for Translating and Adapting Tests (Second Edition). *International Journal of Testing*, 18(2), 101-134. <https://doi.org/10.1080/15305058.2017.1398166>
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a), 631-639. <https://doi.org/10.1080/01621459.1975.10482485>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Praeger.
- Kane, M., & Bridgeman, B. (2022). The evolution of the concept of validity. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 181-203). Routledge.
- Kaplan, J. T., & Iacoboni, M. (2006). Getting a grip on other minds: mirror neurons, intention understanding, and cognitive empathy. *Social neuroscience*, 1, 175-183. <https://doi.org/10.1080/17470910600985605>
- Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazeviski, J., Bonney, C. R., De Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, 42(3), 139-151. <https://doi.org/10.1080/00461520701416231>
- Kim, S.-I., Lee, M.-J., Chung, Y., & Bong, M. (2010). Comparison of brain activation during norm-referenced versus criterion-referenced feedback: The role of perceived competence and performance-approach goals. *Contemporary Educational Psychology*, 35(2), 141–152. <https://doi.org/10.1016/j.cedpsych.2010.04.002>
- Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22, 5–55.
- Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, 26(12), 1827-1832. <https://doi.org/10.1177/0956797615616374>
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), 517-549. <https://doi.org/10.1177/001316445301300401>
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research. *Perspectives on Psychological Science*, 7(6), 537-542. <https://doi.org/10.1177/1745691612460688>
- Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6(4), 311-360. https://doi.org/10.1207/s15327574ijt0604_1
- McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument Development in the Affective Domain: School and Corporate Applications*. Springer.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). MacMillan.
- Meyer, G. M., Marco-Pallarés, J., Boulinguez, P., & Sescousse, G. (2021). Electrophysiological underpinnings of reward processing: Are we exploiting the full potential of EEG? *NeuroImage*, 242, 118478. <https://doi.org/10.1016/j.neuroimage.2021.118478>
- Michaelides, M., Brown, G. T. L., Eklöf, H., & Papanastasiou, E. (2019). *Motivational Profiles in TIMSS Mathematics: Exploring Student Clusters across Countries and Time* (Vol. 7). Springer Open & IEA.
- Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.

- sity Press <https://doi.org/10.1017/CBO9780511490040>
- Michell, J. (2008). Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research and Perspectives*, 6(1-2), 7-24. <https://doi.org/10.1080/15366360802035489>
- Popper, K. (1945). *The open society and its enemies: The high tide of prophecy: Hegel, Marx, and the aftermath* (Vol. II). George Routledge & Sons.
- Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435, 1102-1107. <https://doi.org/10.1038/nature03687>
- Rao, C. R., & Sinharay, S. (Eds.). (2007). *Psychometrics* (1st ed.). Elsevier North-Holland.
- Rupp, A. A. (2008). Lost in Translation? Meaning and Decision Making in Actual and Possible Worlds. *Measurement: Interdisciplinary Research and Perspectives*, 6(1-2), 117-123. <https://doi.org/10.1080/15366360802035612>
- Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, 54(3), 182-203. <https://doi.org/10.1037/0003-066X.54.3.182>
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Greenwood.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39-83). Rand McNally.
- Thielsch, M. T., & Hirschfeld, G. (2019). Facets of Website Content. *Human-Computer Interaction*, 34(4), 279-327. <https://doi.org/10.1080/07370024.2017.1421954>
- van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The Think Aloud Method: A practical guide to modeling cognitive processes*. Academic Press.
- Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. MESA press.
- Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. Lissitz (Ed.), *The concept of validity* (pp. 65-82). Information Age Publishers.