ISSN 1669-2438
ISSN-electrónico 2469-2050
Revista de Psicología. Año 2023. Vol. 19, Nº 38, pp. 121-148

Pasos Fundamentales Para Realizar Adaptaciones de Pruebas Psicológicas

Fundamental Steps for Making Psychological Test Adaptations

María Elena Brenlla[12] ORCID: 0000-0003-2536-9499
Mariana Soledad Seivane[13] ORCID: 0000-0002-9162-6935
Rocío Giselle Fernández da Lama[13] ORCID: 0000-0003-1529-2926
Guadalupe Germano[1] ORCID: 0000-0003-2896-6272

Resumen

Este artículo examina los retos y la importancia de adaptar las pruebas psicológicas. La psicología, como ciencia que estudia la mente y el comportamiento, se enfrenta a la singular complejidad de evaluar constructos intangibles como las emociones, los pensamientos y las actitudes. A diferencia de otras disciplinas científicas, las mediciones psicológicas suelen ser indirectas y estar influidas por errores de medición. En consecuencia, es crucial garantizar su fiabilidad y validez. Este artículo profundiza en los pasos fundamentales para realizar adaptaciones de pruebas psicológicas. Dado que la mayor parte de la investigación psicológica procede de países anglosajones, es esencial modificar las pruebas para adaptarlas a poblaciones e idiomas diversos. Por ello, un aspecto crítico es la adaptación lingüística y cultural de los instrumentos y pruebas psicológicas. Se resalta la importancia de las perspectivas émicas y éticas para comprender los matices culturales y lingüísticos y se abordan los posibles sesgos y heurísticos que pueden influir en los resultados de las pruebas. Además, se destaca el papel de la adaptación para promover una mejor comprensión del comportamiento humano en diversos contextos culturales. Por último, se presenta una síntesis clara de los pasos para la adaptación de las pruebas, siguiendo las directrices de la International Test Commission (ITC). La incorporación de consideraciones culturales y lingüísticas en la adaptación de las pruebas mejorará sin duda la eficacia y aplicabilidad de las

[1]Pontificia Universidad Católica Argentina. Centro de investigaciones en Psicología y Psicopedagogía (CIPP).
[2]Universidad a Distancia de Madrid (UDIMA)
[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Mail de contacto: bren@uca.edu.ar
DOI: https://doi.org/10.46553/RPSI.19.38.2023.p121-148
Fecha de recepción: 23 de octubre de 2023 - Fecha de aceptación: 30 de octubre de 2023

evaluaciones psicológicas en poblaciones diversas.
Palabras clave: psicometría; escala; confiabilidad; validez; prueba psicológica

Abstract

This article examines the challenges and importance of adapting psychological tests. Psychology, as a science that studies the mind and behavior, faces the unique complexity of assessing intangible constructs such as emotions, thoughts, and attitudes. Unlike other scientific disciplines, psychological measurements are often indirect and influenced by measurement errors. Consequently, it is crucial to ensure their reliability and validity. This article delves into the fundamental steps in making adaptations of psychological tests. A critical aspect of cross-cultural research is linguistic and cultural adaptation. Given that most psychological research comes from Anglo-Saxon countries, it is essential to modify tests to adapt them to diverse populations and languages. The authors highlight the importance of emic and ethical perspectives in understanding cultural and linguistic nuances. In addition, they address potential biases and heuristics that may influence test results. The role of adaptation in promoting a better understanding of human behavior in diverse cultural contexts is highlighted. Finally, a clear synthesis of the steps for test adaptation is presented, following the guidelines of the International Test Commission (ITC). The incorporation of cultural and linguistic considerations in test adaptation will undoubtedly improve the effectiveness and applicability of psychological assessments in diverse populations.
Key words: psychometrics; scale; reliability; validity; psychological testing

Introducción

La Psicología es la ciencia que se encarga de estudiar la mente y el comportamiento (American Psychological Association [APA], 2015). Como toda rama de la ciencia debe tener instrumentos válidos para poder evaluar su objeto de estudio. Así, por ejemplo, la medicina se vale de análisis de sangre y estudios por imágenes para estudiar el cuerpo humano; la química utiliza balanzas, termómetros y probetas para abordar la materia y la energía; la biología para aproximarse a los seres vivos utiliza microscopios, probetas y pipetas. Entonces, cabe preguntarse ¿cómo estudian los psicólogos la mente y el comportamiento humanos? Tal como sucede en las otras disciplinas, no existe una única forma de abordar el objeto de estudio. Sin embargo, a diferencia de otras áreas científicas, especialmente las exactas, la psicología debe afrontar la dificultad que conlleva el estudio del ser humano. En numerosas ocasiones se enfrenta al estudio de aspectos que se pueden explicar, pero no se pueden medir con exactitud por ser invisibles, tales como las emociones, el pensamiento, las actitudes, el estilo de personalidad, entre otros. A diferencia de una hoja de una planta que podría introducirse en un microscopio para ser estudiada por un biólogo, un psicólogo no podría introducir el pensamiento de una persona en una herramienta para ser analizado. Ese pensamiento no es tangible en sí mismo, pero sí existe y es objeto de estudio. Así sucede con la mayoría de los aspectos

abordados por la psicología, lo cual suscita un gran desafío que viene siendo abordado desde hace tiempo. Este tema puede sintetizarse en la frase "Medimos a los hombres por sus sombras", atribuida a Thurstone, que señala que medir en psicología no es una tarea sencilla y, por ende, la evaluación tampoco lo es. No obstante, el progreso tecnológico de las últimas dos décadas puede resultar fundamental para el avance significativo de la medición y evaluación psicológica.

En este punto vale la pena introducir la distinción de dos conceptos estrechamente ligados, pero distintos: la medición y la evaluación en psicología. En primer lugar, medir implica básicamente asignar números a los fenómenos siguiendo ciertas reglas (Martínez Arias, 1995). Sin embargo, esto no es sencillo en psicología debido a que, tal como se mencionó previamente, las mediciones son indirectas y están influenciadas por el error de medida. En otras palabras, los constructos psicológicos no son accesibles directamente, sino que se infieren a partir de manifestaciones que pueden ser recolectadas mediante diferentes técnicas de evaluación. Esto implica que la mayoría de las mediciones no se refieren a cantidades, como lo exige el modelo clásico de medida. En cambio, se refieren a relaciones entre cantidades, según lo establece el modelo representacional, que pueden ser de tipo nominal (distintividad), ordinal (orden), de intervalos (aditividad) y de razón (proporcionalidad; Stevens, 1946).

Además, la medición en psicología no es exacta. Como postuló Spearman (1905), toda medición está sujeta al error de medida, lo que significa que la puntuación observada está compuesta por una puntuación verdadera —que es desconocida— y el error de medida. Por lo tanto, las mediciones no solo son indirectas, sino que, por definición, son inexactas. Por esta razón, es crucial que las puntuaciones derivadas de técnicas y pruebas psicológicas presenten evidencias satisfactorias de fiabilidad y validez. Estos conceptos serán desarrollados en profundidad posteriormente. Solo es necesario recordar aquí que la fiabilidad se refiere a la consistencia, estabilidad y objetividad de las medidas, mientras que la validez se relaciona con su significación y pertinencia empírica y conceptual. Ambas características son indispensables en la evaluación de cualquier medida o prueba psicológica.

En segundo lugar, la evaluación en psicología implica un proceso más amplio que consiste en un procedimiento sistemático para observar y describir la conducta utilizando escalas o categorías establecidas. Como señaló Cronbach (1972), la evaluación va más allá de la mera aplicación de pruebas, ya que implica la integración y valoración de la información recopilada. Mientras que la medición responde a la pregunta "¿cuánto?", un proceso de evaluación se centra en la pregunta "¿qué significa o implica ese resultado?". En este sentido, la evaluación puede entenderse como un juicio de valor asociado a un desempeño o resultado.

Garaigordobil (1998) define la Evaluación Psicológica (EP) como

> aquella disciplina que explora y analiza el comportamiento de un individuo o grupo con diversos objetivos (descripción, diagnóstico, selección/predicción, explicación, cambio y/o valoración) a través

> de un proceso de toma de decisiones en el que se utilizan diversos dispositivos (pruebas y técnicas de medida y/o evaluación), tanto para evaluar aspectos positivos como patológicos (p. 22).

De esta manera, la EP abarca tareas como el psicodiagnóstico, la evaluación de potencialidades y capacidades, y la valoración de programas e intervenciones, que se clasifican según el objetivo específico de la evaluación (Casullo, 2009; Fernández Ballesteros, 2013). Un caso especial es el psicodiagnóstico, que a menudo incluye pruebas proyectivas. Si bien se reconoce el valor de estas pruebas, este artículo se centrará únicamente en las pruebas psicométricas. Así, se introduce un concepto clave en esta área, la Psicometría.

La Psicometría, en un sentido amplio, es la disciplina que se encarga de la medición en psicología. Según Martínez Arias (1995) la Psicometría "aglutina todo el conjunto de modelos formales que posibilitan la medición de variables psicológicas, centrándose en las condiciones que permiten llevar a cabo todo proceso de medición en psicología" (p. 21). Por su parte, Muñiz (1998) aclara que pueden distinguirse cinco grandes áreas en la Psicometría: (a) la Teoría de la medición, que incluye a las teoría clásica o de la medición conjunta, teoría operacional y teoría representacional de la medida (Martinez Arias, 1995); (b) la Teoría de los tests —Teoría Clásica de los Tests (TCT) y Teoría de Respuesta al ítem (TRI)—; (c) el escalamiento psicofísico, cuantificación de la percepción de estímulos físicos, que dan lugar a la formulación de las leyes de Weber, Fechner y Stevens; (d) el escalamiento psicológico, cuantificación del continuo psicológico, por ejemplo las escalas de Thurstone, Likert y Guttmann; y (e) el desarrollo y aplicación de técnicas multivariadas para la construcción y análisis de los tests.

En este sentido, la *construcción* de un test[1] implica —en forma expresa o tácita— la asunción de una teoría de la medida, una tipología de test, un escalamiento y el uso de técnicas estadísticas para obtener evidencias de fiabilidad y validez. A su vez, la *adaptación* de tests, que es el objeto de este artículo, refiere a la constatación de que un test construido en una cultura sea igualmente pertinente en otra; aún más, la adaptación de tests constituye, en parte, una prueba de la universalidad del constructo medido y coadyuva al desarrollo de la ciencia psicológica.

Relevancia de la Adaptación de Pruebas Psicológicas

En los párrafos precedentes se ha hecho un recorrido por los conceptos más relevantes a tomar en consideración en relación con las pruebas psicológicas. Se introducirá entonces la cuestión específica de la adaptación de pruebas. Así, vale la pena preguntarse ¿por qué resulta importante adaptar las pruebas psicológicas? Tal como sucede en la mayoría de las disciplinas científicas, casi la totalidad de las investigaciones y progresos en ciencia y tecnología psicológica se dan en países desarrollados y del hemisferio norte, generalmente de habla inglesa. Esto puede observarse a simple vista en las bases de datos de alto impacto tales como Scopus y PubMed y, fundamentalmente, en los rankings de revistas científicas que

clasifican y evalúan la calidad e impacto de las publicaciones tales como Web of Science Journal Citation Reports (JCR) y Scimago Journal Rank (SJR). Por ejemplo, al ingresar en SJR al área de psicología se visualiza que de las diez primeras revistas del ranking, ocho son estadounidenses y dos son inglesas (ver: https://www.scimagojr.com/journalrank.php?area=3200). De este modo, se puede asumir que la gran mayoría de las pruebas de medición en psicología se construyen en inglés, es decir, un idioma que no es el utilizado en Latinoamérica en general ni en Argentina en particular. Además, en el desarrollo de estos instrumentos se utilizan muestras que viven en lugares con pautas culturales que difieren en gran medida de las locales. Esto responde, en buena parte, a la pregunta acerca de la importancia de la adaptación de técnicas: la adaptación lingüística y la consideración de diferencias culturales son cruciales en la adaptación porque influyen decisivamente en la fiabilidad y validez de las medidas.

En las décadas de 1980 y de 1990, algunos investigadores latinoamericanos como Casullo (1999) y Marín (1986) y otros provenientes de la psicología transcultural (Berry, 1980; Triandis y Brislin., 1984) pusieron de relieve la importancia de analizar la influencia de los factores culturales en la adaptación de tests. Para ello, tuvieron en cuenta el enfoque transcultural (Berry, 1980) que clasifica a los constructos en *etic,* constructos universales, considerados como tales por la comunidad científica; *emic,* constructos dependientes de una cultura determinada y *pseudoetic,* que toman como universales a constructos que, en realidad, son propios de una cultura. En concreto, lo émico se refiere al enfoque interno, desde dentro de la cultura o grupo social en estudio. Se centra en las creencias, valores, normas y significados que son importantes para los individuos dentro de un contexto cultural específico. Por otro lado, el enfoque ético se refiere a un enfoque externo o desde fuera de la cultura en estudio. Este enfoque busca establecer parámetros y normas universales para evaluar el comportamiento humano y, en definitiva, su propósito es tener un marco de referencia objetivo para analizar y juzgar la conducta (Mostowlansky y Rota, 2020). Por ello, He y Van de Vijver (2012) señalan que, para maximizar la validez en el desarrollo y adaptación de constructos y medidas, hay dos temas esenciales en estudio: el sesgo y la equivalencia. El sesgo se refiere a cualquier error sistemático que atente contra la comparabilidad de los datos transculturales. Por ello, la demostración de equivalencia es previa. En general, se distinguen tres tipos de sesgos: los de constructo, de método y de ítem, que se estudian y neutralizan mediante investigaciones acerca de las equivalencias de constructo, las cuales analizan si existen diferencias en la definición del constructo y de sus conductas representativas; las métricas —que estudian la comparabilidad de las muestras, la familiaridad con los estímulos y los procedimientos de respuesta, los estilos de respuesta y la concordancia de consignas e instrucciones— y las escalares, que se enfocan en la adaptación lingüística de ítems, consignas y escalas de medidas.

Este último punto atañe a la adaptación lingüística que no se restringe, como es de esperar, a la mera traducción de los ítems. Si así fuera, se podría utilizar una escala adaptada en español en cualquier población con el mismo idioma. Por ejemplo, una prueba desarrollada en Estados Unidos

es adaptada en España o en Argentina y esa adaptación es utilizada en cualquier país de Latinoamérica. Esto ocurre con frecuencia, pero no es una práctica correcta. Para entender esto es necesario puntualizar en las diferencias de semántica y de habla (pragmática del lenguaje) aún dentro del mismo idioma. Baste recordar que el número de las zonas dialectales del español en las Américas oscila entre tres y 16 según el autor (Quesada Pacheco, 2014) y que ello implica diferencias en el significado de las palabras (Por ejemplo, en España "cartera", designa una pieza de marroquinería para llevar dinero, pero en Argentina, al bolso de mujer) y el uso de distintos términos para designar los mismos conceptos (por ejemplo, el trozo de papel de variadas formas que los niños remontan los días de viento, en España se dice "cometa", en Argentina, "barrilete" y en México, "papelote"). Por eso, es importante sopesar adecuadamente las particularidades del habla en español al analizar la pertinencia de una adaptación realizada en una zona para ser utilizada en otra. En la Tabla 1 se puede visualizar un breve listado de palabras que difieren en dos países diferentes hablantes del español. Esta síntesis permite entender

Tabla 1

Ejemplos de Palabras en Español de España y en Español de Argentina (Distintas Palabras para el Mismo Significado y las Mismas Palabras con Distinto Significado)

Categoría Gramatical	Español de España	Español de Argentina
Adjetivo	Enfadado	Enojado
Sustantivo	Falda	Pollera
Sustantivo	Saco (Receptáculo de tela o cuero)	Saco (Blazer)
Verbo	Tomar/Coger	Agarrar
Expresión	Me hace ilusión	Me entusiasma
Pronombre	Vos: tratamiento de máxima solemnidad (Vos, majestad, sabéis de nosotros)	Vos: tratamiento de máxima intimidad (Vos sabés lo que te espera)
Expresión	El día me cundió	Aproveché el día
Expresión	Igual voy a tu casa (Quizá iré a tu casa)	Igual voy a tu casa (Seguro iré a tu casa)
Adjetivo	Prolijo (Largo, dilatado con exceso)	Prolijo (Ordenado, pulcro, esmerado)
Adjetivo	Constipado (Resfriado)	Constipado (Estreñido)
Expresión	Tener apuro (Tener vergüenza)	Tener apuro (Tener prisa)

rápidamente la importancia de realizar una adaptación lingüística incluso en pruebas a utilizarse en países con el mismo idioma.

Finalmente, otro punto para destacar es la posible influencia de los sesgos y heurísticos. Los heurísticos son procedimientos de estimación —"atajos cognitivos"— y redundan en respuestas intuitivas. Se utilizan no solo para los problemas de alta complejidad, sino también para cuestiones simples de verosimilitud, frecuencia y predicción y se clasifican en heurísticos de anclaje, de representatividad y de disponibilidad (Kahneman & Tversky, 1979). En particular, en relación con las adaptaciones de pruebas psicológicas, hay que considerar el efecto marco y los heurísticos de disponibilidad.

El efecto marco refiere a las variaciones de respuesta que producen los sujetos según el modo en que se presente la información de una tarea. La información puede presentarse desde un marco positivo o bien uno negativo y ello es decisivo para definir la dirección de la respuesta (Kahneman & Tversky, 1976). Por ejemplo, en un estudio con el BDI-II se observa que la presencia de títulos negativos en los grupos de ítems (por ejemplo, "Desvalorización"; "Pesimismo"; "Fracaso") influye en la respuesta al inventario. Si bien la correlación entre la administración de una versión con títulos y otra sin títulos en la misma muestra es alta (Brenlla y Rodríguez, 2006), no obstante, se registra un efecto principal significativo, ya que las puntuaciones son mayores en la versión con títulos (Brenlla et al., 2023).

Los heurísticos de disponibilidad refieren a cuán disponibles, cuan rápidamente vienen a la mente los ejemplos de algo y está asociado con el efecto de recencia y con la memoria. Los hechos que se recuerdan mejor se utilizan para establecer frecuencias o probabilidades. Así, un mismo ítem puede tener diferentes significados en distintas culturas. Esto lleva a que un grupo puede obtener puntajes significativamente distintos en un ítem determinado a pesar de obtener un puntaje total similar en la puntuación total. Por ejemplo, en la adaptación argentina del WISC-IV, en el estudio piloto, se constató que los niños obtenían puntuaciones más bajas en ítems gráficos que contenían dibujos de trineos y bellotas. Al realizar entrevistas cognitivas con los niños, se notó que estos dibujos no les eran familiares, ya que aludían a objetos poco usuales. Se los reemplazó por hamacas y zanahorias y las puntuaciones concordaron con las esperadas para la edad y con las del país de origen (Taborda et al., 2011), lo cual señala la importancia de la relevancia cultural en la adaptación de ítems (Mikulic, 2007), tanto en las pruebas de rendimiento como en las evaluaciones neuropsicológicas y en las de personalidad.

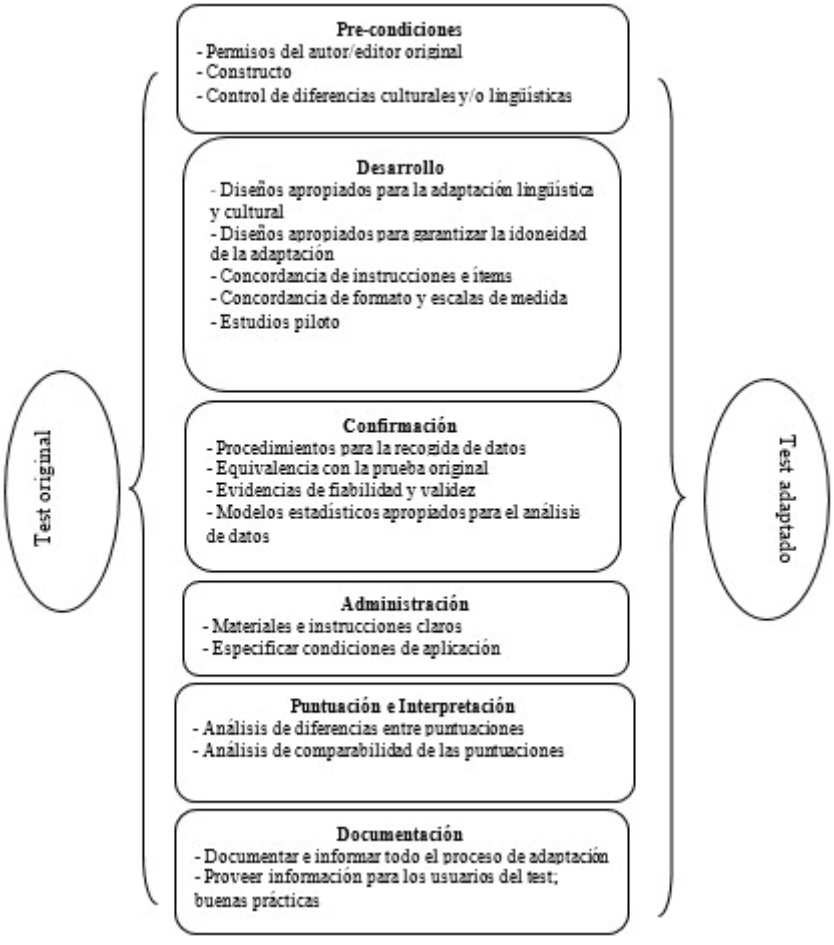
En síntesis, a diferencia de otras disciplinas científicas, la psicología enfrenta el desafío de abordar aspectos intangibles del ser humano, como las emociones, el pensamiento y las actitudes. A pesar de la dificultad que implica medir estos constructos, los psicólogos han desarrollado técnicas de evaluación y adaptación de pruebas para comprender y valorar la complejidad de la conducta humana en distintos contextos culturales y lingüísticos. Si bien la medición en psicología puede ser inexacta, la búsqueda constante de fiabilidad y validez en las pruebas psicológicas impulsa el avance en esta área de investigación.

Por ello, la adaptación de pruebas

juega un papel fundamental para asegurar que los instrumentos utilizados en estudios y evaluaciones sean apropiados y significativos para la diversidad de poblaciones y culturas. A medida que la tecnología avanza, la Psicología se beneficia de nuevos enfoques que permiten un mejor entendimiento de la mente y el comportamiento humanos, promoviendo el progreso en el conocimiento y la práctica psicológica. A continuación, se presentará una síntesis de los pasos para adaptación de pruebas siguiendo los lineamientos de la Comisión Internacional de Pruebas o *International Test Comission* en inglés (International Test Comission [ITC], 2017) que se resumen en la Figura 1. A su vez, se invita al lector a considerar la lista de verificación de estos lineamientos

Figura 1

Pasos en la Adaptación de un Test de Acuerdo a los Lineamientos de la ITC (2017)



que realizan Hernández et al. (2020).

Pasos para Realizar Adaptación de Pruebas Psicométricas

Directrices Previas (3)

Las directrices previas incluyen tres elementos fundamentales: la obtención del permiso del autor o editor original del test para su adaptación; las características y equivalencia del constructo a medir y la minimización de la influencia de diferencias culturales y lingüísticas (ITC, 2017).

Permiso del Autor. Se torna fundamental antes de iniciar cualquier proceso de adaptación de un test el obtener el permiso de los autores o portadores de la propiedad intelectual de este. Estas prácticas se reflejan en la autenticidad de la versión final de una prueba y dificulta la adaptación no autorizada de tests para los distintos campos que existen (Muñiz et al., 2013). Asimismo, este paso remite a las dimensiones éticas de la evaluación psicológica, principalmente desde la dimensión de la integridad del evaluador al presentar una comunicación y espíritu colaborativos con otros profesionales para el enriquecimiento del conocimiento (Muñiz, 1998). A la vez promueve los ideales de la Ciencia Abierta al generar un espacio colaborativo y transparente entre investigadores (Becerril-García et al., 2018).

Constructo a Medir/ Diferencias Culturales y Lingüísticas. El uso de instrumentos diseñados y validados para otras culturas entraña ciertos aspectos a considerar de manera concienzuda por el investigador. Si bien la aplicación de tests diseñados en otras culturas reducen la inversión de recursos humanos, económicos y de tiempo y que emplear un mismo instrumento posibilita unificar el conocimiento derivado del estudio sobre una temática en particular; no obstante, no es correcto utilizar un instrumento extranjero, con normas realizadas en otro país sin realizar la tarea de adaptación correspondiente. Este proceso implica distintos niveles de modificación del instrumento original. Así, se diferencia entre la mera aplicación —administrar una prueba habiendo realizado previamente la traducción literal de los ítems originales al idioma de la población objeto—, adaptación —tiene lugar un proceso de mayor transformación del instrumento original, pero en su mayor parte su estructura y contenido se conservan—, y ensamble —las modificaciones realizadas son profundas y el resultado final diverge del instrumento original grandemente, por lo que ya se trata de un test nuevo (Van de Vijver y Hambleton, 1996).

Cuando se verifican diferencias culturales y/o lingüísticas importantes en el constructo o entre la población objeto y la originaria, no es recomendable realizar un proceso de adaptación. Por caso, en el estudio del comportamiento económico del ahorro en población argentina, el emplear instrumentos que hubiesen sido concebidos y diseñados para economías de otras características (más estables, con menor inflación, entre otros factores) o que representen una realidad socioeconómica e histórica muy distinta, acarrearían posiblemente mediciones y resultados poco representativos de la realidad local (Fernández Da Lama y Brenlla, 2023ª, 2023b).

Directrices de Desarrollo (5)

Las directrices de desarrollo incluyen cinco temas referidos a (a) los procedimientos para la adaptación lingüística y cultural de los elementos del test; (b) el

uso de diseños apropiados para garantizar la idoneidad de la adaptación de la prueba; (c) brindar evidencias acerca de la concordancia de las instrucciones, las consignas del test, con los ítems; (d) del formato de los ítems y escalas de medida y © proveer de información detallada acerca de los estudios piloto realizados.

Diseños Apropiados para la Adaptación Lingüística y Cultural/ Para Garantizar la Idoneidad de la Adaptación. Un aspecto para destacar como fundamental en todo estudio transcultural es definir la perspectiva del proyecto investigativo. En este sentido, la investigación transcultural basada en una perspectiva operacional tenderá a determinar si el concepto, fenómeno o constructo existe en una cultura determinada, mientas que, si el interés del investigador está centrado en la comparación entre grupos culturales, el producto adaptado final deberá reflejar de manera sensible la integración en las distintas culturas de interés (Irvine y Caroll, 1980).

La literatura clásica sobre la metodología de investigación a aplicar en estudios transculturales delimita tres aspectos fundamentales (Berry, 1980):

El primero son estudios de equivalencia cultural, en tanto que el constructo a medir presente un significado y relevancia similar entre las distintas poblaciones, proceso que deberá incluir distintas traducciones y retro traducciones realizadas de manera independiente, estudios de análisis de la población objetivo, estudios piloto para el testeo de las versiones generadas; y monitoreo del flujo de trabajo. Asimismo, vale mencionar que, si bien el proceso de retro traducción —o *back-translation* en inglés— es de uso muy frecuente en investigación, distintas problemáticas se han asociado al mismo. En particular, Behr (2017) encuentra incongruencias cuando las versiones re-traducidas se contrastan con comprobaciones adicionales por parte de hablantes nativos y que puede ocasionar falsas alarmas y dejar ocultos problemas de adaptación. Por ejemplo, en la adaptación argentina de los subtests verbales del WAIS-III (Brenlla, 2004), mediante la técnica de back-translation, se llegó a la palabra "audaz" como traducción de "audacious". No obstante, en el estudio piloto, esta palabra presentó un índice de dificultad mucho menor al encontrado en inglés. Al analizar la frecuencia de uso de la palabra audacious, se constató que ésta es una palabra de baja frecuencia de uso en inglés; en cambio, en español, la palabra "audaz" es de una frecuencia de uso medio-alta. Por ende, la adaptación lingüística —basada en el significado de la palabra y su frecuencia de uso— llevó a cambiarla por "intrépido" que funcionó de acuerdo con lo esperado en el estudio final.

El segundo es equivalencia funcional, en tanto el grado en que un concepto genera una actitud o respuesta similar o es concebido de manera semejante invariablemente en las poblaciones objeto. Puede citarse el concepto "saudade", término de origen portugués incorporado al español, que remite a un estado emocional profundo de nostalgia y felicidad (Bulat Silva, 2012; Farrell, 2006; Vasconcelos, 1996). Este término no cuenta con una traducción literal a otras lenguas, por lo que no contaría con una equivalencia funcional entre distintas culturas a diferencia de otros conceptos como amor o compromiso (Neto y Mullet, 2014).

Finalmente, la perspectiva émica y ética a la hora de abordar un concepto incluye la comprensión de la existencia de diferencias en los significados atribuibles al constructo de interés dentro de un entorno cultural específico, tal como se puntualizó en la introducción.

En virtud de estos antecedentes se recomienda que el equipo orientado en el proceso de traducción se encuentre constituido por distintos profesionales con experticia relevante al constructo a evaluar, por profesionales bilingües, expertos en la cultura diana y expertos en construcción de tests. Los procedimientos y diseños para la adaptación lingüística pueden incluir la traducción directa, la re-traducción o la traducción simultánea según el objetivo de la investigación (Hernández et al., 2020), pero en todos los casos, es conveniente trabajar con traducciones independientes y juicio de expertos para la obtención de la versión adaptada definitiva del test.

Concordancia de Instrucciones e Ítems/ Concordancia de Formato y Escalas de Medida. Otro aspecto de importancia es proveer evidencias de que las instrucciones del test y de los ítems tengan un significado similar en la cultura origen y en la cultura donde se adapta el test. Para ello, es de utilidad trabajar con grupos a los que se les administra la prueba para indagar acerca de la claridad y pertinencia de los ítems, las consignas e instrucciones. El mismo procedimiento puede ser utilizado para analizar las escalas de medida, las opciones de respuesta y el formato de los ítems así como asegurar que la población diana comprende y está familiarizada con la administración de pruebas.

Estudios Piloto. Por último, se remarca la importancia de llevar adelante estudios piloto que permitan detectar errores o dificultades de comprensión de la escala por parte de la muestra objetivo, así como también emprender análisis preliminares a nivel psicométrico de la versión adaptada experimental. De acuerdo con Martínez Arias (1995) una muestra piloto debería incluir al menos 120 casos para una primera aplicación. El análisis de los datos del estudio piloto incluye revisar la calidad psicométrica de la adaptación —análisis de ítems, fiabilidad y validez— y la calidad de la adaptación lingüística realizada (Hernández et al., 2020). En caso de detectar errores sistemáticos que sugieran dificultad en la adaptación psicométrica y/o lingüística de algún ítem, se recomienda elaborar reactivos alternativos y realizar un nuevo estudio piloto y un re-análisis de los datos con una nueva muestra piloto.

Directrices de Confirmación (4)

Esta etapa incluye (a) los procedimientos para la recogida de datos, (b) la equivalencia con la prueba original, (c) las evidencias de fiabilidad y validez y (d) el uso de modelos estadísticos apropiados para el análisis de datos. Esto es, se profundiza en el análisis de las propiedades psicométricas del test adaptado en relación con el test original.

Muestra. Debe iniciarse definiendo las características de los sujetos que conformarán la muestra, así como la técnica de muestreo y la suficiencia, relevancia y representatividad de la población en estudio. El tamaño muestral tiene un impacto en la precisión de las estimaciones estadísticas que se realicen en el estudio de adaptación ya sea un análisis factorial exploratorio, confirmatorio o de ecuaciones estructurales

(Kyriazos, 2018; Thompson, 2004). A la hora de definir el tamaño muestral existen distintos métodos, en términos generales, muestras más grandes son preferibles a muestras más reducidas, especialmente dada la mayor estabilidad en la solución factorial arribada con las primeras (DeVellis, 2017). Una regla clásica indica como deseable el contar entre 5 y 10 casos por ítem del instrumento (Everitt, 1975; Gorsuch, 1983), entre al menos 200 (Cattell, 1978), 250 (Comrey y Lee, 1992) o varios cientos más (Thompson, 2004). No obstante, aspectos como las comunalidades, cargas factoriales, cantidad de factores, naturaleza de las variables, y el número de ítems por factor también influyen en la definir el tamaño muestral requerido (Costello y Osborne, 2005; DeVellis y Thorpe, 2017; MacCallum et al., 1999), con lo cual, no se trata de una regla univariante en todo estudio. De hecho, recolectar una muestra "demasiado grande" no solo representa un sobre esfuerzo a nivel investigativo, junto con el uso extensivo de recursos de tiempo y dinero, sino que, además, podría remarcar efectos espurios o poco relevantes.

Otro aspecto importante, especialmente dado el desarrollo de programas y aplicativos, es el análisis de potencia estadística como R Studio (2015), G*Power (Kang, 2021) y calculadoras online (Preacher y Coffman, 2006). Este tipo de análisis aplicados al cálculo del tamaño muestral permiten detectar la probabilidad de detectar un efecto real presente (Coolican, 2018).

Equivalencias. Un aspecto fundamental en este punto es explorar los aspectos técnicos relacionados a las propiedades psicométricas de la prueba adaptada y el grado de equivalencia con la prueba original. Para esto existen métodos basados en los modelos de ecuaciones estructurales (SEM; Garnier-Villarreal y Jorgensen, 2020; Cui et al., 1998; Rigo y Donolo, 2018), en la Teoría de Respuesta al Ítem (TRI; Auné et al., 2020; Attorressi et al., 2009; Bean y Bowen, 2021; Choi y Asilkalkan, 2019; Muñiz, 2010; Toland, 2013), y métodos de detección del funcionamiento diferencial de los ítems (DIF; Zumbo, 2003, 2007; Zumbo et al., 2015).

En cuanto al análisis de la equivalencia de los ítems adaptados y de la prueba original, se recomiendan los estudios de Ferrando y Lorenzo-Seva (2014, 2018) y Ferrando et al. (2022). A modo de resumen, el investigador deberá considerar la adecuación de los datos y la muestra al estudio de adaptación a realizar, el cálculo de estadísticos descriptivos univariados, el análisis de la varianza común explicada, así como determinar qué conjunto de ítems serán analizados. Deberá definir qué tipo de modelo factorial se empleará, junto con el tipo de solución factorial, el cálculo de los distintos parámetros que den cuenta del ajuste del modelo y la adecuación de la solución factorial obtenida, la coherencia sustantiva del modelo ajustado, y por último, la selección del conjunto de ítems con mejores propiedades psicométricas que conformen la versión final del test.

Fiabilidad y Modelos Estadísticos Apropiados. En términos generales, la fiabilidad de un instrumento remite a la consistencia en las puntuaciones obtenidas tras su uso repetido (Muñiz, 2010). Como se ha mencionado en apartados anteriores, la validez y la fiabilidad de una prueba son

dimensiones de interés para la evaluación educativa y psicológica (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [NCME], 2014). En la Tabla 2 se exponen de manera resumida las diferentes dimensiones que componen el estudio de la confiabilidad de un instrumento.

Puede suceder también que la fuente de imprecisión en las puntuaciones arrojadas por la prueba provenga de cambios personales aleatorios en el examinado y en las condiciones de administración que ocurran a lo largo del tiempo. Esto podría ocurrir en el caso de evaluar una variable que se viera modificada a medida que el sujeto de desarrolla (relación entre inteligencia y edad hasta cierto punto de la vida de la persona) o donde la mediación del aprendizaje es de relevancia. Sin embargo, si esos cambios que se dan en la variable no se encuentran sustentados por la teoría del investigador, bien puede estarse frente a un funcionamiento inadecuado de la prueba adaptada. Este fenómeno remite a la estabilidad temporal de las puntuaciones. La estimación de la confiabilidad en este caso requerirá de al menos dos conjuntos de medidas paralelas que difieran en contenido de la prueba o el tiempo de administración (método de formas alternas o equivalentes o método test-retest)

Tabla 2

Procedimientos Empíricos para Determinar la Confiabilidad de un Instrumento

Dimensiones de la confiabilidad	**Métodos**	**Cantidad de aplicaciones**	**Criterios teóricos y estadísticos**
Estabilidad	Test-retest Formas paralelas	2	Coeficiente de Pearson
Consistencia interna	División por mitades Formas paralelas Coeficiente Alfa de Cronbach, Ordinal, u Omega de McDonald Alfa de Cronbach (escalas dicotómicas) Kuder Richardson (escalas dicotómicas)-	1	Coeficiente de Pearson Alfa, KR 20, Ω
Objetividad	Acuerdo entre examinadores (control de fluctuaciones en las puntuaciones según el evaluador)	1	Kappa/w de Kendall Coeficiente intraclase

de los mismos sujetos examinados. Se deberá calcular un coeficiente de correlación de Pearson para evaluar la estabilidad entre administraciones de una prueba (Guttman, 1945; Polit, 2014).

Por su parte, el estudio de la consistencia interna estará referido a en qué medida la elección de la muestra de ítems que componen la prueba resulta una fuente de error en la medición realizada. Representa una tarea vital en la labor científica el considerar y controlar la presencia de errores aleatorios que puedan alterar la medición (Cronbach y Shavelson, 2004).

La fórmula KR-20 (Kuder y Richardson, 1937), antecedente del Alfa de Cronbach, tradicionalmente se ha empleado para estimar la consistencia interna de escalas dicotómicas. Se torna importante destacar que el KR-20 no es recomendable si se trabaja con ítems con un nivel heterogéneo de dificultad (Merino-Soto y Charter, 2010), por lo que se sugiere emplear un método de corrección en tal caso (Horst, 1953).

El uso del Alfa de Cronbach ha contado con gran popularidad y adhesión por parte de investigadores (Doval et al., 2023; Ursachi et al., 2015) a la hora de estimar la consistencia interna en escalas ordinales. En el contexto de su uso en investigación, valores por encima del coeficiente 0.70 para la escala total o una subescala son interpretados como aceptables (Taber, 2018; Van Griethuijsen et al., 2015). En cambio, para la toma de decisiones en el ámbito de la psicología aplicada, se recomiendan valores entre 0.80 y hasta 0.95, ya que coeficientes más altos podrían indicar redundancia entre los ítems. No obstante, con el tiempo han surgido varios cuestionamientos sobre la potencia estadística de esta medida (Christmann y Van Aelst, 2006; Sijtsma, 2009), hallando incluso una infravaloración de la consistencia interna de escalas de menos de 10 ítems (Herman, 2015), y se ha criticado su uso en escalas ordinales, especialmente en aquellas de menos de 5 opciones de respuesta (Elosua Oliden y Zumbo, 2008; Espinoza y Novoa-Muñoz, 2018).

En la actualidad, se recomienda el uso de otros indicadores, tales como los coeficientes alfa ordinal (Gadermann et al., 2012; Zumbo et al., 2007), omega (McDonald, 1999; Viladrich et al., 2017), beta (Revelle, 1979), theta (Armor, 1973), y H (Hancock y Mueller, 2001).

A la hora de determinar qué coeficiente de fiabilidad emplear es fundamental tener en cuenta el nivel de medición de la variable en cuestión (Doval et al., 2023). Esto implica distinguir si se trabaja con datos dicotómicos, ordinales, o continuos. En este sentido, se sugiere el uso de matrices de correlaciones o covarianzas de Pearson para variables continuas, mientras que el uso de matrices de correlaciones policóricas y tetracóricas se ha establecido para variables ordinales y dicotómicas, respectivamente (Ferrando et al., 2022). Vale destacar que, el uso de un coeficiente de confiablidad estará sujeto también a las características del modelo de la variable a medir. Por tanto, en el caso de contar con un modelo unifactorial, se deberá estimar el alfa o el omega para el total de la escala, mientras que, si cuenta con dos o más factores, se deberá calcular el coeficiente para cada una de ellas.

El estudiar la dimensión de la objetividad en la confiabilidad de un instrumento estriba en qué grado la medición de un rasgo es independiente de

la subjetividad del evaluador (Tristán-López y Pedraza Corpus, 2017). En este sentido, existen distintos métodos conocidos que dependerá de la naturaleza de la variable que se mida.

El uso de Kappa de Cohen está sujeto a la evaluación del acuerdo entre dos examinadores de una variable medida de manera nominal y en el caso de contar con más de dos examinadores, se debería usar el coeficiente Kappa de Fleiss (Cohen, 1968; Fleiss y Cohen, 1973; McHugh, 2012). En el caso de una variable medida de manera ordinal, se recomienda el uso del coeficiente Tau de Kendall (Hays, 1960; Jinyuan et al., 2016; Lapata, 2006). Por último, para variables continuas, se recomienda el cálculo de coeficientes intraclase (Bartko, 1966; Bland y Altman, 1990; Weir, 2005), los cuales difieren de su interpretación de la mera estimación de una correlación de valores, ya que una correlación establece asociación entre estos, pero no mide acuerdo. Vale mencionar también el método *Detection of Multiple Examiners Nor In Consensus* (DOMENIC; Baca-García et al., 2001; Cicchetti y Showalter, 1997) que permite calcular el acuerdo entre evaluadores en base a un promedio global de nivel de acuerdo inter-examinador y a la vez, un nivel de acuerdo de cada examinador individualmente (García-Nieto et al., 2012).

Validez y Modelos Estadísticos Apropiados. Dentro del ámbito de la adaptación de pruebas psicológicas, el concepto de validez ha sido objeto de estudio y debate durante mucho tiempo debido a su naturaleza compleja en términos de definición e interpretación. Actualmente, la validez es entendida como la adecuación, significación y utilidad de las inferencias específicas hechas a partir de las puntuaciones de los tests (APA, 2014). Las puntuaciones de un test evidencian propiedades de validez cuando se comprueba que el test realmente mide el constructo que pretende evaluar.

Antes de entrar en más detalles, cabe realizar una breve reseña histórica del concepto, la cual puede dividirse en tres etapas (Messick, 1995). En sus inicios, el concepto de validez estaba estrechamente asociado a la correlación entre las puntuaciones de un test y alguna medida externa que el test buscaba predecir. Esta perspectiva enfatizaba la importancia de establecer relaciones entre el test y criterios externos para validar su uso y asegurar su precisión en la medición. Para ese entonces, autores como Garrett (1937) y Guildford (1946; citados en Ventura-León, 2016) asociaban la validez con la medición de coeficientes estadísticos que cuantificaban la relación entre el test y una variable de referencia.

En una etapa posterior, en la década de los cincuenta, el concepto de validez experimentó cambios significativos y evolucionó para incluir otras dimensiones. Surgió la validez de contenido, entendida como la medida en que el contenido del test refleja adecuadamente el constructo que se pretende evaluar (Pedrosa et al., 2014) y se introdujo por primera vez el concepto de validez de constructo, especialmente en relación con el análisis factorial. Esta forma de validez se centra en examinar la estructura interna del test y su capacidad para medir el constructo subyacente. Se busca identificar los factores latentes que sustentan el test y su relación con el constructo que se desea medir (Ferrando et al., 2022). En esta segunda etapa, se destaca el modelo tripartito de validez basado en tres tipos: de contenido, de

constructo y de criterio, introducido en el año 1966 por la *American Educational Research Association* (AERA). Dicho modelo ha significado una importante contribución en el campo de la evaluación y medición de variables psicológicas y educativas, teniendo relevancia científica hasta el día de hoy.

Actualmente, la comprensión de la validez se ha ido transformando y ampliando, pasando de una visión limitada centrada en coeficientes a un enfoque más integral y basado en múltiples fuentes de evidencia. Se entiende a la validez como un concepto unitario y refiere al grado en que la evidencia respalda las inferencias realizadas. En este sentido, la validez no es una característica del instrumento, sino una cualidad asociada al uso del instrumento en un contexto particular (Elosua Oliden, 2003). Un test puede ser validado para una población y propósito específico, pero eso no garantiza la validez del test en todas las poblaciones y para todos los propósitos (Knekta et al., 2019). Por ejemplo, no sería correcto asumir que un cuestionario validado para evaluar el nivel de satisfacción laboral en trabajadores de empresas tecnológicas sea igualmente válido para medir el nivel de satisfacción en empleados de una industria diferente, como la agricultura. Los factores y dinámicas que influyen en la satisfacción laboral pueden ser muy diferentes en cada industria, por lo cual se necesitaría validar específicamente ese cuestionario para los agricultores. Entonces, para referirnos a la validez de un test resulta necesario considerar el propósito o interpretación propuesta, así como también el contexto de aplicación.

Desde esta perspectiva, ya no se habla de distintos tipos de validez, sino de diversas fuentes de evidencia. En específico, se proponen cinco tipos de evidencias de validez, las cuales se basan en: el contenido del test, es decir, en qué medida el contenido del test refleja adecuadamente el constructo que se está evaluando; la estructura interna del test, que se refiere a la organización y coherencia de los ítems del test; el proceso de respuesta al test, que busca comprender cómo los individuos responden a las preguntas y tareas del test; las relaciones con otras variables externas, es decir, cómo se correlaciona el test con otras medidas o comportamientos relevantes; y finalmente, las consecuencias de la aplicación del test, evaluando el impacto y las implicaciones del uso del test en las personas evaluadas (AERA et al., 2014).

Las evidencias basadas en el contenido hacen referencia al grado en el que el contenido de un test (sus ítems) reflejan una muestra representativa y relevante del constructo que pretende medir. Este análisis involucra tres aspectos principales: la definición del ámbito temático, el análisis de cómo se representa dicho ámbito y la evaluación de su relevancia (Sireci, 1998). Siguiendo los lineamientos de la APA (APA et al.,1999), el método por excelencia para reunir evidencias de contenido de un test es el juicio de expertos, en el cual un grupo de expertos en el constructo que se busca medir evalúa la relevancia, suficiencia, claridad y coherencia del contenido. Se deben considerar varios aspectos tales como el formato de los ítems, el tipo de tareas requeridas, la claridad de la consigna, la familiaridad con la situación propuesta, el tipo de material utilizado y las posibles diferencias en motivación o ansiedad.

Aunque la evidencia basada en contenido suele ser mayormente cualitativa

y se fundamenta en razonamientos lógicos, en ocasiones puede incorporar medidas empíricas de concordancia, especialmente en pruebas de rendimiento y criterio, utilizando índices como la congruencia interjueces o técnicas de escalamiento unidimensional y multidimensional (Elosua Oliden, 2003). Hay estudios que se han encargado de sintetizar los principales avances teóricos y metodológicos referidos a las evidencias de validez de contenido, entre los cuales se destaca el de Pedrosa y cols. (2014).

En relación con las evidencias de constructo, hay que recordar que refiere a en qué medida el test representa la teoría psicológica sobre la que se fundamenta y si permite interpretar las puntuaciones de acuerdo con ello (AERA et al., 2014). Entre los procedimientos más utilizados se cuenta el análisis factorial exploratorio y confirmatorio para el análisis de la estructura interna del instrumento; las matrices multimétodo-multirrasgo (Campbell y Fiske, 1959) que proveen evidencias de validez convergente —correlaciones del mismo constructo evaluado con distintos instrumentos— y discriminante —correlaciones de distintos constructos evaluados con el mismo tipo de instrumento y el análisis de redes para el estudio de la dimensionalidad de los tests (Christensen y Golino, 2021).

En cuanto a la comparación con criterios externos, estos difieren en función de si las evidencias son predictivas, es decir, si el criterio se evalúa luego de haber aplicado el test; concurrentes, cuando el

Tabla 3

Procedimientos Empíricos para Determinar las Evidencias de Validez de un Instrumento

Tipo de evidencia de validez	Subtipos	Criterios teóricos y estadísticos
Validez empírica o de criterio	Validez Concurrente Validez Predictiva Validez Retrospectiva	Correlación bivariada r de Pearson o Rho de Spearman; Grupos contrastados (t de student, Chi cuadrado)
Validez de constructo o teórica	Validez Convergente Validez Discriminante	Estudios evolutivos y clínicos; Análisis Factorial Exploratorio y Confirmatorio; Correlación bivariada r de Pearson o Rho de Spearman; Estudios de metaanálisis; Estudios pretest-postest; Análisis de Redes (*Network Analysis*) Matrices multi-método/multi-rasgo
Validez de contenido		Juicio experto; Cálculo de V de Aiken;

criterio se evalúa al mismo tiempo que el test; o retrospectivas, si el criterio se valora antes de aplicar el test. En todos los casos, los análisis más recomendados son los análisis de correlación y de regresión, así como los de diferencias de medias con cálculo del tamaño del efecto (Cortada de Kohan y Macbeth, 2007).

Directrices Sobre Aplicación (2)

Las condiciones de la administración de un test pueden influir en la validez y la confiabilidad de sus puntuaciones (Muñiz et al., 2013). Respecto a este punto, la ITC propone dos procedimientos para minimizar cualquier sesgo relacionado con la cultura y el idioma causado por los procedimientos de administración y los formatos de respuesta: (a) preparar materiales e instrucciones claras sobre la aplicación del test adaptado y (b) especificar las condiciones de aplicación del test adaptado que deben seguirse en todas las poblaciones a las que va dirigido (Hernández et al., 2020).

Directrices Sobre Puntuación e Interpretación (2)

Las dos directrices incluidas en este apartado refieren a que (a) si se observan diferencias en las puntuaciones de los grupos, es menester analizar toda la información relevante disponible incluyendo valores culturales, religiosidad, posición socioeconómica, entre otras y (b) solamente comparar las puntuaciones entre poblaciones cuando el nivel de invarianza ha sido bien establecido.

Un paso esencial en el proceso de adaptación es establecer normas de puntuación específicas para cada población o contexto cultural. Esto permitirá comparar las puntuaciones de los participantes con un grupo de referencia adecuado y obtener una interpretación más precisa de los resultados.

La comparación directa de puntuaciones obtenidas en contextos culturales o lingüísticos diferentes utilizando escalas adaptadas puede ser problemática y poco confiable debido a las diferencias en motivación, trayectorias escolares, valores culturales, nivel de vida, políticas educativas y oportunidades de acceso a la educación entre diferentes grupos o comunidades (Muñiz et al., 2013). Por lo tanto, se recomienda utilizar los estudios comparativos únicamente para comprender las similitudes y diferencias entre los grupos analizados, pero no para establecer comparaciones directas sin considerar los factores contextuales. Resulta necesario demostrar la equivalencia psicométrica y empírica de las escalas para permitir la comparación de puntuaciones.

Directrices Sobre Documentación (2)

Estas directrices incluyen (a) proveer toda la documentación técnica referida a las modificaciones, cambios respecto del original, así como toda la evidencia que garantice la equivalencia de las medidas y (b) brindar información clara a los usuarios del test para asegurar las buenas prácticas en la profesión. Como última instancia, tal y como recomienda la ITC, una parte esencial del proceso de adaptación de tests es la creación de una documentación exhaustiva que describa en detalle todo el proceso llevado a cabo, incluyendo los procedimientos de administración, adaptación y validación realizados (Hernández et al., 2020). Además, esto implica dejar disponible información minuciosa acerca de las alteraciones y

ajustes efectuados en comparación con el test original. De esta manera, otros investigadores y profesionales podrán evaluar la calidad de la adaptación y replicar los estudios de ser necesario.

Conclusiones

El objetivo de este artículo es proveer una guía para la realización de adaptaciones de pruebas a la Argentina que respete fundamentalmente dos aspectos. Por un lado, los criterios internacionales para la adaptación de tests (ITC, 2007) y, por otro, la toma de conciencia acerca de que la adaptación realizada en un país de habla española no garantiza su pertinencia cultural, psicométrica y lingüística en otro país o región hispanohablante. Por ello, en el trabajo se hizo alusión a dos cuestiones fundamentales, la distinción entre constructos etic, emic y pseudoetic provenientes de la psicología transcultural y la importancia de considerar los conocimientos de la ciencia psicológica sobre la influencia de los heurísticos cognitivos en las respuestas a situaciones cotidianas y cómo ello puede influir en la adaptación de tests. En definitiva, la adaptación de tests coadyuva tanto al desarrollo de la ciencia psicológica como al proceder ético en la evaluación y la práctica profesional. Esperamos que esta guía colabore para afianzar una actitud científica y ética en la toma de decisiones y propender las buenas prácticas.

Referencias

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

American Psychological Association. (2015). *Dictionary of Psychology* (2nd edition). American Psychological Association (APA). http://dx.doi.org/10.1037/14646-000

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological test y manuals.* Washington, DC: American Psychological Association.

Armor, D. J. (1973). Theta Reliability and Factor Scaling. *Sociological Methodology, 5,* 17-50. https://doi.org/10.2307/270831

Auné, S. E., Abal, F. J. P., & Attorressi, H. F. (2020). Análisis psicométrico mediante la Teoría de la Respuesta la Ítem: modelización paso a paso de una Escala de Soledad. *Ciencias Psicológicas, 14*(1), e-2179. https://doi.org/10.22235/cp.v14i1.2179

Attorressi, H. F., Lozzia, G. S., Abal, F. J. P., Galibert, M. S., & Aguerri, M. E. (2009). Teoría de Respuesta al ítem: Conceptos básicos y aplicaciones para la medición de constructos psicológicos. *Revista Argentina de Clínica Psicológica, 18*(2), 179-188. https://www.redalyc.org/pdf/2819/281921792007.pdf

Baca-García, E., Blanco, C., Sáiz-Ruiz, J., Rico, F., Diaz-Sastre, C., & Cic-

chetti, D. V. (2001). Assessment of reliability in the clinical evaluation of depressive symptoms among multiple investigators in a multicenter clinical trial. *Psychiatry Research, 102*(2), 163-173. https://doi.org/10.1016/S0165-1781(01)00249-9

Bartko, J. J. (1966). The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports, 19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Bean, G. J., & Bowen, N. K. (2021) Item Response Theory and Confirmatory Factor Analysis: Complementary Approaches for Scale Development. *Journal of Evidence-Based Social Work, 18*(6), 597-618, https://doi.org/10.1080/26408066.2021.1906813

Becerril-García, A., Aguado-López, E., Batthyány, K., Melero, R., Beigel, F., Vélez Cuartas, G., Banzato, G., Rozemblum, C., Amescua García, C., Gallardo, O., & Torres, J (2018). *AmeliCA : Una estructura sostenible e impulsada por la comunidad para el Conocimiento Abierto en América Latina y el Sur Global*. México: Redalyc; Universidad Autónoma del Estado de México ; Argentina : CLACSO; Universidad Nacional de LaPlata ; Colombia : Universidad de Antioquia. En Memoria Académica. Disponible en: http://www.memoria.fahce.unlp.edu.ar/libros/pm.693/pm.693.pdf

Behr, D. (2017). Assessing the use of back translation: the shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology, 20*(6), 573-584. https://doi.org/10.1080/13645579.2016.1252188

Berry, J. W. (1980). Acculturation as varieties of adaptation. En A. M. Padilla (Ed.), *Acculturation: Theory, models and some new findings* (pp. 9-25). Westview.

Bland, J. M. & Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methos of measurement. *Computers in Biology and Medicine, 20*(5), 337-340. https://doi.org/10.1016/0010-4825(90)90013-F

Bulat Silva, S. (2012). Saudade: A key Portuguese emotion. *Emotion Review, 4*(2), 203-211. https://doi.org/10.1177/1754073911430727

Brenlla, M. (2004). Aspectos socio culturales y métricos en la adaptación de tests: un estudio en base al test de inteligencia para adultos de Wechsler III (WAIS III). *XI Jornadas de Investigación de la Facultad de Psicología de la Universidad de Buenos Aires, Buenos Aires.* Recuperado de http://docplayer.es/23453801-Xi-jornadas-de-investigacion-facultad-de-psicologiauniversidad-de-buenos-aires-buenos-aires-2004

Brenlla, M. E.; Fernández Da Lama, R.G.; Otero, A. & Filgueira, P. (2023). *The influence of titles on test validity: exploring the frame effect on the Beck Depression Inventory-second edition* (manuscrito sin publicar).

Brenlla, M. E. & Rodríguez, C. M. (2006).

Adaptación argentina del Inventario de Depresión de Beck (BDI-II). En *BDI-II. Inventario de depresión de Beck* (pp. 11-37). Paidós.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Carpenter, S. (2018). Ten Steps in Scale Development and Reporting: A Guide for Researchers. *Communication Methods and Measures, 12*(1), 25-44. https://doi.org/10.1080/19312458.2017.1396583

Casullo, M. M. (1999). La evaluación psicológica: Modelos, técnicas y contexto sociocultural. *Revista Iberoamericana de diagnóstico y evaluación psicológica*, *1*(1), 97-113. https://aidep.org/03_ridep/R07/R077.pdf

Casullo, M. M. (2009). La evaluación psicológica: Modelos, técnicas y contextos. *Revista iberoamericana de diagnóstico y evaluación psicológica, 1*(27), 9-28. https://www.redalyc.org/pdf/4596/459645443002.pdf

Cattell, R. B. (1978). Factor Measures: Their Construction, Scoring, Psychometric Validity, and Consistency. *The scientific use of factor analysis in behavioral and life sciences*, 273-320. https://link.springer.com/chapter/10.1007/978-1-4684-2262-7_11

Choi, Y.-J. & Asilkalkan, A. (2019). R Packages for Item Response Theory Analysis: Descriptions and Features. *Measurement: Interdisciplinary Research and Perspectives, 17*(3), 168–175. https://doi.org/10.1080/15366367.2019.1586404

Christensen, A. P., & Golino, H. (2021). On the equivalency of factor and network loadings. *Behavior Research Methods, 53*(4), 1563–1580. https://doi.org/10.3758/s13428-020-01500-6

Christmann, A. & Van Aelst, S. (2006). Robust estimation of Cronbach's alpha. *Journal of Multivariate Analysis, 97*(7), 1660-1674. https://doi.org/10.1016/j.jmva.2005.05.012

Cicchetti, D. V. & Showalter, D. (1997). A computer program for assessing interexaminer agreement when multiple ratings are made on a single subject. *Psychiatry research, 72*(1), 65-68. https://doi.org/10.1016/S0165-1781(97)00093-0

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Comrey, A. L. & Lee, H. B. (1992). Interpretation and application of factor analytic results. En A. L. Comrey & H. B. Lee (Eds.), *A first course in factor analysis.* Hillsdale, NJ: Lawrence Eribaum Associates. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=2335989

Contreras Espinoza, S. & Novoa-Muñoz, F. (2018). Ventajas del alfa ordinal respecto al alfa de Cronbach ilustradas con la encuesta AUDIT-

-OMS. *Revista Panamericana de Salud Pública, 42,* e65. https://doi.org/10.26633/RPSP.2018.65

Cortada de Kohan, N. & Macbeth, G. (2007). El tamaño del efecto en la investigación psicológica*. Revista de Psicología, 3*(5) 25-31. http://bibliotecadigital.uca.edu.ar/repositorio/revistas/efecto-investigacion-psicologica-kohanmacbeth.pdf

Coolican, H. (2018). *Research methods and statistics in psychology.* Routledge.

Costello, A. B. & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation, 10*(1), 7. https://doi.org/10.7275/jyj1-4868

Cronbach, L. J. (1972). *Fundamentos de la exploración psicológica*. Biblioteca Nueva.

Cronbach, L. J. & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement, 64*(3), 391-418. https://doi.org/10.1177/0013164404266386

Cui, G., van den Berg, S., & Jiang, Y. (1998). Cross-cultural adaptation and ethnic communication: Two structural equation models. *Howard Journal of Communication, 9*(1), 69-85. https://doi.org/10.1080/106461798247122

DeVellis, R. F. & Thorpe, C. T. (2017). *Scale Development: Theory and Applications*. SAGE Publications.

Doval, E., Viladrich, C.,ânguloulo-Brunet, A. (2023). Coefficient alpha: the resistance of a classic. *Psicothema*, *35*(1), 5-20. https://www.psicothema.com/pdf/4785.pdf

Elosua Oliden, P., (2003). Sobre la validez de los tests. *Psicothema, 15*(2), 315-321. https://www.redalyc.org/articulo.oa?id=72715225

Elosua Oliden, P. & Zumbo, B. D. (2008). Coeficientes de confiabilidad para las escalas de respuesta categórica ordenada. *Psicothema, 20*(4), 896-901. https://www.psicothema.com/pi?pii=3572

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *The British Journal of Psychiatry, 126*(3), 237-240. https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/abs/multivariate-analysis-the-need-for-data-and-other-problems/4226888B99AA7C3F861B3B203050AC17

Farrell, P. (2006). Portuguese saudade and other emotions of absence and longing. In B. Peeters (Ed.), *Semantic primes and universal grammar: Empirical evidence from the Romance languages* (pp. 235- 258). John Benjamin.

Fernández Ballesteros, R. (2013). *Evaluación psicológica. Conceptos, métodos y estudio de casos*. Madrid: Síntesis. Síntesis.

Fernández Da Lama, R. G. & Brenlla, M.ªE. (2023a). Resultados preliminares en la evaluación de actitudes hacia el ahorro en economías inestables: importancia de los factores contextuales y socioeconómicos. Revista PUCE Pontificia Universidad Ca-

tólica de Ecuador, 116. https://doi.org/10.26807/revpuce.vi

Fernández Da Lama, R. G., & Brenlla, M. E. (2023b). Attitudes towards saving and debt-taking behavior during first major flexibility on pandemic restrictions in Argentina. *International Journal of Economic Behavior*, *13*(1), 51-70. https://doi.org/10.14276/2285-0430.3716

Ferrando, P. J. & Lorenzo-Seva, U. (2014). El análisis factorial exploratorio de los ítems: algunas consideraciones adicionales. *Anales de Psicología, 30*(3), 1170-1175. http://dx.doi.org/10.6018/analesps.30.3.199991

Ferrando P. J., & Lorenzo-Seva, U. (2018). Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis. *Journal of Educational and Psychological Measurement, 78*(5), 762-780. https://doi.org/10.1177/0013164417719308.

Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decálogo para el Análisis Factorial de los ítems de un test. *Psicothema, 34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

Fleiss, J. L. & Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement, 33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation, 17,*1-13. https://doi.org/10.7275/N560-J767

Garaigordobil, M. (1998). *Evaluación Psicológica: Bases teórico-metodológicas, situación actual y directrices de futuro*. Amarú.

García-Nieto, R., Parra Uribe, I., Palao, D., Lopez-Castroman, J., Sáiz, P. A., García-Portilla, M. P., Sáiz Ruiz, J., Ibañez, A., Tiana, T., Durán Sindreu, S., Pérez Sola, V., de Diego-Otero, Y., Pérez-Costillas, L., Fernámdez García-Andrade, R., Saiz-González, D., Jiménez Arriero, M. A., Navío Acosta, M., Giner, L., Guija, J. A., … & Baca-García, E. (2012). Protocolo breve de evaluación del suicidio: fiabilidad interexaminadores. *Revista de Psiquiatría y Salud Mental, 5*(1), 24-36. https://doi.org/10.1016/j.rpsm.2011.10.001

Garnier-Villarreal, M. & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods, 25*(1), 46–70. https://doi.org/10.1037/met0000224

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment*, *5*(2), 141-154. https://psycnet.apa.org/record/1983-31741-001

van Griethuijsen, R. A. L. F., van Eijck, M. W., Haste, H., den Brok, P. J., Skinner, N. C., Mansour, N., Savran Gencer, A., & BouJaoude, S.

(2015). Global Patterns in Students' Views of Science and Interest in Science. *Research in Science Education, 45*(4), 581-603. https://doi.org/10.1007/s11165-014-9438-6

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika 10,* 255–282. https://doi.org/10.1007/BF02288892

Hancock, G. R. & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In *Structural equation modeling: Present and future* (pp. 195-216).

Hays, W. L. (1960). A note on average tau as a measure of concordance. *Journal of the American Statistical Association, 55*(290), 331-341. https://doi.org/10.2307/2281746

Herman, B. C. (2015). The Influence of Global Warming Science Views and Sociocultural Factors on Willingness to Mitigate Global Warming. *Science Education, 99*(1), 1-38. https://doi.org/10.1002/sce.21136

Hernández, A., Hidalgo, M. D., Hambleton, R. K., & Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema, 32*(3), 390-398. https://doi.org/10.7334/psicothema2019.306

He, J. & van de Vijver, F. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture, 2*(2), Article 8. https://doi.org/10.9707/2307-0919.1111

Horst, P. (1953). Correcting the Kuder-Richardson reliability formula for dispersion of item difficulties. *Psychological Bulletin, 50*(5), 371- 374. https://doi.org/10.1037/h0062012

International Test Comission. (2017). *The ITC Guidelines for Translating and Adapting Tests* (Second edition). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Irvine, S. H. & Caroll, W. K. (1980). Testing and assessment across cultures: Issues in methodology and theory. En H. D. Trandis & J. W. Berry (Eds.), *Handbook of Cross-culture Psychology Vol. 2: Methodology* (pp. 181-244). Allyn & Bacon.

Jinyuan, L. I. U., Wan, T. A. N. G., Guanqin, C. H. E. N., Yin, L. U., & Changyong, F. E. N. G. (2016). Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai archives of psychiatry, 28*(2), 115-120. https://doi.org/10.11919/j.issn.1002-0829.216045

Kang, H. (2021). Sample size determination and power analysis using the G* Power software. *Journal of educational evaluation for health professions, 18.* https://doi.org/10.3352/jeehp.2021.18.17

Kahneman D. & Tversky A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291. https://doi.org/10.2307/1914185

Knekta, E., Runyon, C., & Eddy, S. (2019). One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research. *CBE—Life Sciences Education, 18*(1), 1-17. https://doi.org/10.1187/cbe.18-04-0064

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://link.springer.com/article/10.1007/BF02288391

Kyriazos, T. A. (2018). Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology, 9*(08), 2207. https://www.scirp.org/html/15-6902564_86856.htm

Lapata, M. (2006). Automatic evaluation of information ordering' Kendall's tau. *Computational Linguistics, 32(*4), 471-484. https://doi.org/10.1162/coli.2006.32.4.471

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989x.4.1.84

Marín, G. (1986): Consideraciones metodológicas básicas para conducir investigaciones en América Latina. *Acta Psiquiátrica y Psicologica de America Latina, 32*(3), 183-192. https://pesquisa.bvsalud.org/portal/resource/pt/lil-44521?lang=es

Martínez Arias, M. del R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Síntesis.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates Publishers.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

Merino Soto, C. & Charter, R. (2010). Modificación Horst al Coeficiente KR-20 por Dispersión de la Dificultad de los Ítems. *Revista Interamericana de Psicología/Interamerican Journal of Psychology, 44*(2), 274-278. https://www.redalyc.org/pdf/284/28420641008.pdf

Messick, S. (1995). Validity of psychological assessment: Validation of inferences fro' persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Mikulic, I. M. (2007). *Construcción y adaptación de Pruebas Psicológicas*. https://acortar.link/xMByre

Mostowlansky, T., & Rota, A. (2020). Emic and Etic. En F. Stein, S. Lazar, M. Candea, H. Diemberger, J. Robbins, A. Sanchez, & R. Stasch (Eds.). *The Cambridge Encyclopedia of Anthropology* (pp. 1-16). University of Cambridge. https://boris.unibe.ch/154189/

Muñiz, J. (1998). La Medición de lo Psicológico. *Psicothema, 10*(1), 1-21. https://reunido.uniovi.es/index.php/PST/article/view/7442

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo, 31*(1), 57-66. https://digibuo.uniovi.es/dspace/bitstream/handle/10651/10994/?sequence=1

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25*(2), 151-157. https://doi.org/10.7334/psicothema2013.24

Muñiz, J., & Fonseca-Pedrero, E. (2019).

Diez pasos para la construcción de un test. *Psicothema, 31*(1), 7-16. https://digibuo.uniovi.es/dspace/bitstream/handle/10651/51958/Diez.pdf?sequence=1

Neto, F. & Mullet, E. (2014). A Prototype Analysis of the Portuguese Concept of Saudade. *Journal of Cross-Cultural Psychology, 45*(4), 660–670. https://doi.org/10.1177/0022022113518370

Pedrosa, I., Suárez-Álvarez, J., & García-Cueto, E. (2014). Evidencias sobre la Validez de Contenido: Avances Teóricos y Métodos para su Estimación. *Acción Psicológica, 10*(2), 3-20. http://dx.doi.org/10.5944/ap.10.2.11820

Polit, D. F. (2014). Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Quality of Life Research: An international journal of quality of life aspects of treatment, care and rehabilitation, 23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

Preacher, K. J. & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA [Computer software]. http://quantpsy.org/.

Quesada Pacheco, M. Á. (2014). División dialectal del español de América según sus hablantes Análisis dialectológico perceptua. *Boletín de filología*, *49*(2), 257-309. http://dx.doi.org/10.4067/S0718-93032014000200012

Revelle, W. (1979). Hierarchical Cluster Analysis and The Internal Structure Of Tests. *Multivariate Behavioral Research, 14*(1), 57-74. https://doi.org/10.1207/s15327906mbr1401_4

Rigo, D. Y. & Donolo, D. (2018). Modelos de Ecuaciones Estructurales usos en investigación psicológica y educativa. *Revista Interamericana de Psicología/Interamerican Journal of Psychology, 52(*3), 345-357. https://doi.org/10.30849/ripijp.v52i3.388

RStudio. (24 de Abril de 2015). RStudio. http://www.rstudio.com/about/

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107-120. https://doi.org/10.1007/s11336-008-9101-0

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299-321. https://doi.org/10.1207/s15326977ea0504_2

Spearman, C. (1905). Proof and Disproof of Correlation. *The American Journal of Psychology, 16*(2), 228-231. https://doi.org/10.2307/1412129

Stevens, S. S. (1946). On the Theory of Scales of Measurement | Science. *Science, 103*(2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education, 48*(6), 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Taborda, A. R., Brenlla, M. E., & Barbenza, C. (2011). Adaptación argentina de la Escala de Inteligencia de Wechsler para Niños IV (WISC-

IV). En D. Wechsler. *Escala de Inteligencia de Wechsler para Niños IV (WISC-IV)* (pp. 37-55). Paidós.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. *Applied Psychological Measurement, 31*(3), 245-248. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1a0656fae7e2bed422fed08bedf0dab73203f325

Toland, M. D. (2013). Practical Guide to Conducting an Item Response Theory Analysis. *The Journal of Early Adolescence, 34*(1), 120–151. https://doi.org/10.1177/0272431613511332

Triandis, H. C., & Brislin, R. W. (1984). Cross-cultural psychology. *American psychologist, 39*(9), 1006-1016. https://doi.org/10.1037/0003-066X.39.9.1006

Tristán-López, A., & Pedraza Corpus, N. Y. (2017). La objetividad en las pruebas estandarizadas. *Revista Iberoamericana de evaluación educativa, 10*(1), 11-31. https://doi.org/10.15366/riee2017.10.1.001

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How Reliable are Measurement Scales? External Factors with Indirect Influence on Reliability Estimators. *Procedia Economics and Finance, 20*, 679-686. https://doi.org/10.1016/S2212-5671(15)00123-9

Vasconcelos, M. C. (1996). *A saudade portuguesa*. Guimarães Editors.

Ventura-León, J. L. (2016). Breve historia del concepto de validez en Psicometría. *Revista peruana de historia de la Psicología, 2*, 89-92. https://historiapsiperu.org.pe/wp-content/uploads/2021/08/Version-completa-del-volumen-2.pdf#page=89

van de Vijver, F. & Hambleton, R. K. (1996). Traslating tests: some practical guidelines. *European Psychologist, 1*(2), 89-99 https://doi.org/10.1027/1016-9040.1.2.89

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). Un viaje alrededor de alfa y omega para estimar la fiabilidad de consistencia interna. *Anales de Psicología/Annals of Psychology*, *33*(3), 755-782. https://revistas.um.es/analesps/article/view/analesps.33.3.268401

Weir, J. (2005). Quantifying test-retest reliability using the instraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*(1), 231-240. https://doi.org/10.1519/15184.1

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*(2), 136–147. https://doi.org/10.1191/0265532203lt248oa

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly, 4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal Versions of

Coefficients Alpha and Theta for Likert Rating Scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21-29. https://doi.org/10.22237/jmasm/1177992180

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding. *Language Assessment Quarterly, 12*(1), 136–151. https://doi.org/10.1080/15434303.2014.972559

Notas al Final

[1]Para una guía actualizada sobre el desarrollo de escalas se sugiere revisar las publicaciones de Carpenter (2018); DeVellis y Thorpe (2017); Muñiz y Fonseca-Pedrero (2019), y los estándares de la *American Educational Research Association* (dhttps://www.aera.net/).