

Criterios Homologados de Investigación en Psicología (CHIP) Investigaciones instrumentales Versión 1.0

Documento elaborado por la Comisión de Investigación
del Departamento de Psicología de la PUCP

Doris Argumedo,
Magaly Nóblega,
Patricia Bárrig y
Fiorella Otiniano

Los Criterios Homologados de Investigación en Psicología (CHIP) han sido elaborados por la Comisión de Investigación del Departamento de Psicología en concordancia con los estándares internacionales considerados en la publicación de artículos académicos, el Test Review Model de la European Federation of Psychologists Associations y la tercera edición en español del Manual de publicaciones de la American Psychological Association (APA) publicado el año 2010.

En este sentido, constituye una guía para los cursos de pregrado de la especialidad de Psicología que se encuentren relacionados con el eje de investigación y para el desarrollo de las tesis de pregrado de los alumnos.

El presente documento es una revisión de los criterios aplicables a las investigaciones instrumentales. La primera parte del documento presenta un resumen del proceso de su elaboración. Luego, se desarrollan los criterios para las distintas secciones del reporte de una investigación. Finalmente, se presenta un glosario que contiene las definiciones de la terminología más empleada en investigaciones instrumentales.

Proceso de creación de los Criterios Homologados de Investigación en Psicología (CHIP) Investigaciones instrumentales

En el año 2010, la Comisión de Investigación identificó la necesidad de establecer criterios homologados que guíen el desarrollo de las tesis e investigaciones en la especialidad de Psicología. De esta forma, la elaboración de dichos criterios se constituyó en una meta prioritaria de la comisión. Debido a ello, la Comisión de Investigación se planteó la publicación de CHIP para los diversos tipos de investigaciones.

Los criterios para las investigaciones instrumentales se formaron sobre la base de los estándares internacionales vigentes, con pertinencia local y con acuerdo de la comunidad de profesores de Psicología PUCP. En una primera fase, un equipo de la comisión se dedicó a la revisión y discusión de los diversos estándares internacionales existentes para diferentes componentes de estudios de esta naturaleza: la introducción, el planteamiento del problema, el diseño, el método, los resultados y la discusión. En una segunda fase, se presentaron los enfoques que iban a guiar la estructura lógica del documento y los estándares a los que se iban a adscribir las investigaciones. Estos criterios fueron discutidos en una reunión con profesores del Departamento ligados a la enseñanza y la práctica de la investigación en la especialidad de Psicología. En esta reunión, se elevaron sugerencias y acuerdos que culminaron con el desarrollo de una versión preliminar del documento. En una tercera fase, este documento preliminar fue puesto a consulta de los profesores del Departamento de Psicología que dictaban cursos vinculados al eje de investigación. Finalmente, tomando en consideración las sugerencias brindadas, se llegó a esta primera versión de los CHIP para investigaciones instrumentales.

Criterios Homologados de Investigación en Psicología Investigación Instrumental

Título

El título tendrá un máximo de 12 palabras. Se espera que incluya el constructo central o el nombre del instrumento en estudio, así como el objetivo (diseño/adaptación de una medición y/o propiedades psicométricas investigadas).

Validez factorial del Maslach Burnout Inventory Human Services (MBI-HSS) en profesionales chilenos

(Olivares-Faúndez, Mena-Miranda, Jélvez-Wilke, & Marcía-Sepúlveda, 2014)

Construcción y validación de una escala de autoeficacia para la actividad física

(Fernández, Medina, Herrera, Rueda, & Fernández, 2011)

Adaptación del Índice de Bienestar Personal para adolescentes en lengua española y portuguesa

(Alfaro et al., 2014)

Resumen y palabras clave

El resumen incluye de manera concisa el problema de investigación, el método, los resultados y las conclusiones. Respecto al formato, tiene un máximo de 250 palabras y el interlineado es a doble espacio¹.

Las palabras clave son los descriptores de la investigación, reflejan sus contenidos centrales y, en algunos casos, la población estudiada. Es recomendable utilizar aquellas que se mencionan en los estudios publicados y que resaltan las particularidades de la investigación instrumental (siglas del instrumento, constructo en medición, propiedades psicométricas implicadas). El número de palabras clave es de 3 a 5.

El resumen y las palabras clave tendrán una versión en inglés titulada *Abstract* y *Key words* respectivamente.

¹ Para especificaciones adicionales, se puede revisar el capítulo 2 del Manual de publicaciones de la American Psychological Association (APA, 2010).

The 62-item Multidimensional Experiential Avoidance Questionnaire (MEAQ) was recently developed to assess a broad range of experiential avoidance (EA) content. However, practical clinical and research considerations made a briefer measure of EA desirable. Using items from the original 62-item MEAQ, a 15-item scale was created that tapped content from each of the MEAQ's six dimensions. Items were selected on the basis of their performance in 3 samples: undergraduates ($n = 363$), psychiatric outpatients ($n = 265$), and community adults ($n = 215$). These items were then evaluated using 2 additional samples (314 undergraduates and 201 psychiatric outpatients) and cross-validated in 2 new, independent samples (283 undergraduates and 295 community adults). The resulting measure (Brief Experiential Avoidance Questionnaire; BEAQ) demonstrated good internal consistency. It also exhibited strong convergence with respect to each of the MEAQ's 6 dimensions. The BEAQ demonstrated expected associations with measures of avoidance, psychopathology, and quality of life and was distinguishable from negative affectivity and neuroticism.

Keywords: experiential avoidance, avoidance, psychological flexibility, assessment, Multidimensional Experiential Avoidance Questionnaire

The Brief Experiential Avoidance Questionnaire: Development and Initial Validation
(Gámez, Chmielewski, Kotov, Ruggero, Suzuki, & Watson, 2014)

Tabla de contenidos

La tabla de contenidos incluye las secciones del trabajo y los números de página. Se usan tabulaciones que diferencien las secciones y las partes de las mismas. Es importante recalcar que no se emplean las palabras “capítulo” o “subcapítulo” antes de las secciones y que el espaciado interlineal recomendado es de 2.

Introducción

La primera sección del texto se denomina Introducción. Esta comprende el estado de la cuestión y el planteamiento del problema.

El objetivo de las investigaciones instrumentales es proponer la creación de un nuevo instrumento o la adaptación de una herramienta ya existente a un nuevo contexto de aplicación pertinente al momento actual. La introducción debe contener suficiente evidencia teórica y empírica que permita fundamentar las razones por las que el estudio es relevante, y explicitar el aporte que significaría una nueva escala o adaptación con respecto a mediciones ya existentes. Para ello, es necesario presentar el estado del arte con respecto a: (1) la definición conceptual del constructo que quiere evaluarse, y (2) la medición del mismo en el momento actual.

La sistematización de información actual, coherente y significativa que permite definir cuidadosamente el constructo, si bien es un elemento característico en toda

investigación, es de especial consideración en el proceso de creación/adaptación de una prueba.

Esto supone, por un lado, (1) presentar claramente el marco epistemológico/conceptual en el que el instrumento se inserta, (2) desarrollar el constructo en profundidad en términos conceptuales y en sus facetas o dimensiones, y (3) exponer la variedad de sus manifestaciones o expresiones considerando el contexto en el que el fenómeno se desea evaluar. Es decir, se debe proveer información para garantizar un proceso de construcción/adaptación fundamentado en términos de pertinencia y relevancia, y, a la vez, proporcionar evidencias suficientes, especialmente en caso de requerir sustentar la validez del contenido.

Por otra parte, implica hacer explícitas las relaciones esperadas con otros constructos y/o variables sociodemográficas. Así, en función de la revisión bibliográfica y de los modelos teóricos de referencia, debe proponerse una red significativa de relaciones para el constructo, es decir, un marco que otorgue significado a las puntuaciones derivadas de la escala. Especificar y justificar esta matriz relacional resulta esencial e ineludible para los estudios dirigidos a obtener evidencias externas de validez del instrumento.

En relación con el estado de la medición del constructo de interés, la sistematización se orientará a analizar críticamente la calidad de los instrumentos ya existentes y/o del instrumento seleccionado para su estudio, sobre la base de las evidencias sobre sus cualidades psicométricas en la versión original y otras adaptaciones, así como en las diversas poblaciones en las que han sido utilizados. Esta revisión permitirá: (1) evaluar la idoneidad de las mediciones existentes y/o de la medición en estudio y su pertinencia para el nuevo contexto y población; (2) identificar los consensos, contradicciones y vacíos en las evidencias de validez y confiabilidad recolectadas hasta el momento.

La última parte de la Introducción corresponde a la explicitación de los objetivos de la investigación que se desprenden del estado del arte. Es decir, se explicitará si se va a crear o adaptar un instrumento; y se describirá el conjunto de evidencias que se proveerán para respaldar la consistencia y estabilidad de las mediciones desarrolladas o adaptadas, y las inferencias que se desprenden de las puntuaciones de las mismas. Las

fuentes de validez que se planteen recolectar deben ser apropiadas para sostener las interpretaciones y acciones que se derivan de los puntajes de la prueba en cuestión.

Finalmente, incluirá una breve descripción del diseño de estudio como un modo de anticipar el método utilizado para la investigación sin que este sea nombrado o categorizado de acuerdo con algún autor.

This study assesses the basic psychometric properties of the RCMAS in a South African community sample, using data collected in a large study examining the mental health of children and adolescents living in poor urban townships of Cape Town (Cluver et al., 2007). Reliability of the RCMAS was assessed by examining item-total correlations for individual RCMAS items, as well as Cronbach's α for the full scale and the three narrow anxiety factors identified Reynolds and Paget (1981). Validity of the RCMAS was assessed by examining the associations between total scores and previously identified correlates of anxiety (sex, age, symptoms of depression, posttraumatic stress, somaticism, and delinquent behavior). The underlying factor structure of the RCMAS was assessed using a combination of confirmatory and exploratory factor analyses.

Performance of the Revised Children's Manifest Anxiety Scale in a sample of children and adolescents from poor urban communities in Cape Town
(Boyes & Cluver, 2013)

The present study has two aims: to obtain a relevant Turkish SAM factor structure and to test whether this factor structure is stable across samples of university students and nonstudent adults. This study was conducted in three phases: First, the SAM was adapted into Turkish to determine the best factor structure as well as its internal consistency and item-total correlation coefficients; second, the factor structure was tested by using a confirmatory factor analysis (CFA) across two studies; and third, the validity of the SAM was assessed by a model examining sex differences and by correlating the SAM with measures of related (state anxiety) and unrelated (social desirability) constructs.

The development and psychometric properties of the Turkish Version of the Stress Appraisal Measure
(Durak & Senol-Durak, 2013)

Para la elaboración y redacción de la introducción, es necesario tomar en cuenta las siguientes consideraciones formales:

- La elaboración de esta sección supone la revisión de textos clásicos, revisiones de las teorías que se hayan realizado posteriormente, así como artículos actualizados que representen consensos y puntos de controversia sobre el tema.
- Es importante que la bibliografía utilizada sea actualizada, es decir, predominantemente de los últimos cinco años, en especial para el caso de las investigaciones.

- Es importante que incluya solamente la información que sustente el tema del problema de investigación.
- Los contenidos presentarán una adecuada integración. Esta se refleja tanto en el orden lógico de los argumentos presentados como en el uso de una buena redacción de estilo académico.
- Las citas en el texto y referencias bibliográficas siguen el formato APA.
- Para el caso del documento de tesis de pregrado de la especialidad de Psicología, la sección de la introducción tiene como máximo diez páginas. No obstante, de ser necesario, puede complementarse la información a través del uso de los apéndices al final del texto. También se puede usar la totalidad o una parte de las cinco páginas adicionales disponibles.

Método

La segunda sección del documento se denomina método. El sentido de esta sección es describir la forma en que se ha desarrollado la investigación instrumental de modo que (1) brinde información suficiente para evaluar la veracidad y generalización de los resultados, y (2) viabilice replicarla. Este es un aspecto altamente relevante en estos estudios, pues el sentido de los mismos es ir aportando acumulativamente evidencias de confiabilidad y validez para el instrumento en diversas poblaciones y contextos.

La propuesta metodológica del estudio responderá al problema planteado en la introducción. Cuenta con los siguientes acápites: participantes, medición, procedimiento y análisis de datos.

En el caso de los documentos de tesis de pregrado, la extensión de esta sección es de cinco páginas como máximo. No obstante, de ser necesario, puede complementarse la información a través del uso de los apéndices al final del texto o de las cinco páginas adicionales disponibles para el documento completo.

Participantes

El objetivo de este acápite es identificar las características del grupo de estudio, teniendo en cuenta la contribución de esta descripción para la posibilidad de generalización de los resultados. Esta descripción contiene los detalles relevantes para realizar la comparación con posteriores réplicas del estudio. Sin embargo, se evita brindar

detalles anecdóticos o poco relevantes (por ejemplo, datos sociodemográficos poco relevantes, coordinaciones, envío de cartas).

Esta sección incluye lo siguiente:

- *Las características sociodemográficas de los participantes:* se presenta la descripción de las principales características sociodemográficas, en especial las que se van a tomar en cuenta para los objetivos generales y específicos. Todos los datos son redactados en párrafo, inclusive el promedio, la desviación estándar, el intervalo de confianza, mínimo y máximo, u otros descriptivos de las características descritas. Es importante considerar que solo se recaban y reportan los datos que sean pertinentes para el estudio.
- *La descripción del proceso de conformación del grupo de participantes:* se presenta el proceso seguido para lograr la colaboración de los participantes, los lugares en los que se obtuvo su participación, los criterios de inclusión y de exclusión, el porcentaje de participantes que aceptaron y el porcentaje de autoseleccionados. Se evita brindar detalles anecdóticos o clasificar la técnica de selección empleada de acuerdo con algún autor.
- *La descripción de los estándares éticos seguidos:* se presentan los acuerdos, el uso del consentimiento informado, los pagos o las retribuciones no económicas, las acciones que se incluyen en el proceso para garantizar el cuidado ético a los participantes, la revisión del proyecto por parte de una comisión especializada en los aspectos éticos del estudio, entre otros.

La determinación del tamaño muestral depende de los requerimientos de los diversos procedimientos estadísticos implicados en la evaluación de la calidad de los ítems, estimación de la confiabilidad, estudio de la dimensionalidad de un instrumento, obtención de evidencias de validez externa, construcción de normas, entre otros. A continuación, se presentarán algunos lineamientos generales que pueden servir como referencia para la toma de decisiones respecto al número de participantes.

Para los procesos de construcción/adaptación de un instrumento, en los que los ítems elaborados/traducidos, luego de ser sometidos a una evaluación cualitativa, deben ser analizados cuantitativamente en sus propiedades métricas (media, varianza, índices de discriminación, correlación media inter-ítem), se plantean dos momentos de evaluación: el prepiloto y el piloto. El primero de ellos consiste en la administración

preliminar de la prueba con la finalidad de detectar, por ejemplo, la presencia de ítems que muestren un mal funcionamiento o que sean difíciles de comprender debido a una mala redacción (Martínez-Arias, 1995). Este primer análisis suele basarse, según Osterling (1989), en la administración de estos a un grupo compuesto por entre 50 y 100 participantes. En el caso de que el número de ítems sea demasiado elevado, se recomienda que estos sean divididos y pasados a muestras diferentes. Con los resultados de este primer estudio, se lleva a cabo el piloto con los ítems seleccionados, se sugiere repetir el proceso con la intención de obtener más garantías sobre estos, pero con una muestra de mayor tamaño: 300 participantes como mínimo, o entre 5 y 10 participantes por ítem (Martínez-Arias, 1995).

Con respecto a la confiabilidad de las puntuaciones, la evaluación de la consistencia y estabilidad de las mediciones es sensible a las características de la muestra en términos de su homogeneidad y tamaño (Kieffer & MacDonald, 2011). Para maximizar la variabilidad, se recomienda, por tanto, muestras heterogéneas. En cuanto al tamaño de la muestra, es altamente recomendable que esta se sitúe entre los 200 y 300 participantes, aunque ello se flexibiliza dependiendo de la estrategia de cálculo (índices de consistencia, correlación media inter-ítem, correlación entre mediciones repetidas) (Clark & Watson, 2003). Por ejemplo, el mínimo requerido por el European Federation of Psychologists Associations (EFPA) Test Review Model (2013) para definir como adecuados los estudios de equivalencia, consistencia interna y estabilidad entre puntuaciones es 100 participantes.

Para los estudios instrumentales cuyo objetivo sea recolectar evidencias de validez, se recomienda que los lineamientos para determinar el tamaño de la muestra estén basados en el análisis de la potencia estadística². En general, el análisis del poder puede ser utilizado para calcular el tamaño mínimo de la muestra necesario para detectar un efecto moderado (EFPA, 2013).

Adicionalmente a esta guía general de determinación de tamaño de muestra, se pueden encontrar otros criterios. Los estudios basados en las relaciones con otras variables, ya sea con un criterio (validez de criterio) o con otros constructos que miden lo mismo o con los que deberían diferenciarse (evidencias convergentes y discriminantes),

² Este proceso no es necesario para los estudios conducentes a la obtención del grado de licenciatura.

requieren entre 100 y 200 participantes para ser valorados por el EFPA Test Review Model (2013) como adecuados.

Un caso especial lo constituyen los estudios focalizados en investigar la estructura interna de la prueba a través del análisis factorial exploratorio. Sobre el tamaño muestral necesario para poder aplicar cualquier procedimiento factorial, habría que decir que la respuesta no es única y que las recomendaciones abordan tanto el tamaño muestral total del estudio como el número de casos por variable (N / p).

Hair, Anderson, Tatham y Black (2004) señalan que es preferible trabajar con 100 o más participantes. Tabachnick (2007), por su parte, sugiere al menos 300 casos para obtener soluciones fiables. Recientemente, De Winter, Dodou y Wieringa (2009) señalan que un $N=50$ es un valor mínimo razonable, cuando las comunalidades son altas (mayores que .60) y los factores están compuestos por varios ítems.

Cuando se trata de considerar la proporción de número de casos por variable medida, Stevens (1992) aconseja que al menos se cuente con 5 participantes por cada variable (ítem). Los criterios en este caso también son diversos: 3:1, 6:1, 10:1, 20:1 (Williams, Brown, & Osnman, 2010). Hogarty et al. (2005) señalan, luego de estudiar los distintos referentes, que no existe un número mínimo de proporción N/p que garantice un buen análisis factorial³.

Finalmente, cuando el estudio incluye la determinación de normas o puntajes estandarizados, el EFPA Test Review Model (2013) distingue criterios distintos para normas clásicas y continuas. La diferencia entre ambos tipos de normas es que las continuas se emplean cuando se requiere especificar normas de acuerdo con ciertos subgrupos, como, por ejemplo, diversos rangos de edad o grado escolar. Sin embargo, no se emplea la data de los participantes individuales que componen los grupos; por el contrario, se utilizan las estadísticas descriptivas del grupo como unidad de análisis en los modelos de curvas de mejor ajuste. Para el caso de las normas clásicas, se consideran adecuadas muestras de entre 200 y 300 participantes; su equivalente en normas continuas es de 70 a 90 participantes considerando 8 grupos. A medida que el número de grupos aumenta, tiende a disminuir el número de participantes mínimo requerido por grupo.

³ Para especificaciones adicionales, se puede revisar EFPA (2013).

Medición

El objetivo de la sección es presentar y demostrar la calidad de los instrumentos utilizados.

En el caso de las investigaciones instrumentales, el estado del arte del instrumento en estudio es objeto de presentación y análisis en la introducción; por tanto, solo se consigna en esta sección una breve descripción del mismo. Las evidencias de confiabilidad y validez investigadas en el estudio en cuestión deben ser reportadas en resultados.

Los instrumentos adicionales que forman parte del diseño se describen de modo resumido con respecto a sus características generales (dimensiones, número de ítems, alternativas de respuesta, calificación), y las evidencias de confiabilidad y de validez acordes al estado del arte del instrumento y la población en el que será usado, así como el uso de los instrumentos en poblaciones similares.

Finalmente, para cada uno de estos instrumentos, se presentan brevemente los indicadores de confiabilidad y/o resultados de las evidencias de validez obtenidos en la investigación siempre que se hayan evaluado.

The scale was mainly based on the theory of hardiness of Kobasa (1979, 1982). A review of existing hardiness questionnaires was used to develop an initial item pool representing the classical dimensions of challenge, commitment, and control. As previous research has shown that complex response scales add construct-irrelevant variance (Harvey, Billings, & Nilan, 1985), we decided to use a simple response scale. Thus, items were rated from 1 (*completely disagree*) to 4 (*completely agree*). Furthermore, because negatively worded items in measures of hardiness have been shown to overlap with neuroticism (Sinclair & Tetrick, 2000), we worded items positively. We followed two processes of item refinement. Firstly, four experts from the area of personality established content validity (DeVellis, 2001) and reviewed the initial pool of 30 items. Four responses were used as follows: 1 = *not relevant*, 2 = *unable to assess relevance without item revision*, 3 = *relevant but needs minor revisions* and 4 = *very relevant and succinct*. The number of items rated between 3 and 4 by experts divided by the total number of items was calculated as the Content Validity Index (CVI) of scale. Twenty-two items remained for analysis. The CVI was .85 for the 22 items, higher than the recommended cut-off of .80 (Waltz, Strickland, & Lenz, 1991). Secondly, with an exploratory sample of 200 participants, we selected for analysis only the items with a corrected item-total correlation higher than .40 (DeVellis, 1991). Five items were eliminated, so the initial version of the OHQ consisted of a total of 17 items.

Development and validation of the Occupational Hardiness Questionnaire
(Moreno-Jiménez, Rodríguez-Muñoz, Garrosa, & Blanco, 2014)

The Values in Action Inventory of Strengths for Youth (VIA-Youth; Park & Peterson, 2006b) consists of 198 items for the self-assessment of the 24 character strengths (7–9 items per strength) included in the classification of Peterson and Seligman (2004). About one third of the items are reverse coded. The VIA-Youth uses a 5-point Likert-style format (from 1 = *Not like me at all* to 5 = *Very much like me*)...

The Students' Life Satisfaction Scale (SLSS; Huebner, 1991) adapted to German by Weber, Ruch, and Huebner (2013), is a 7-item self-report measure of satisfaction with life (as a global cognitive judgment of adolescents life). Two of the items are reverse coded. It uses a 6-point answer format (from 1 = *Strongly disagree* to 6 = *Strongly agree*). A sample item is "I have what I want in life." In several studies across cultures (e.g., Huebner, 1991; Weber et al., 2013), the SLSS has been shown to be a reliable and valid measurement. The internal consistency (α coefficient) of the present study was .88.

Character Strengths in Children and Adolescents: Reliability and Initial Validity of the German Values in Action Inventory of Strengths for Youth (German VIA-Youth)
(Ruch, Weber, Park, & Peterson, 2014)

En el caso de que el objetivo central del estudio implique la construcción de un instrumento, se coloca en los apéndices la versión final de este, así como su sistema de observación/calificación completo. Los instrumentos que se encuentran publicados no son colocados en los apéndices con la finalidad de respetar los derechos de autoría o copyright.

Procedimiento

El objetivo de esta sección es reportar información necesaria para la adecuada réplica del estudio que no esté descrita en la introducción y en las partes anteriores de la sección de método.

En el procedimiento, se mencionarán los pasos seguidos para construir o adaptar la prueba en estudio. Esta descripción debe ser breve y suficiente para que el lector pueda entender claramente la metodología usada para dicho proceso. Algunos de estos procedimientos pueden incluir:

- Proceso de selección y/o creación de ítems (inductivo o deductivo)
- Procedimientos de traducción⁴
- Proceso de validación por jueces o expertos⁵
- Estudio piloto⁶

⁴ Para especificaciones adicionales, se puede revisar Peña (2007).

⁵ Para obtener mayor información respecto al tema, se puede revisar Merino y Livia (2009).

⁶ Para especificaciones adicionales, se puede revisar Martínez-Arias (1995).

Items for the initial ELOCS scale were generated by taking into consideration definitions and items employed in other measures (EDE—Fairburn & Cooper, 1993; EDE-Q—Fairburn & Beglin, 1994; QEWP—Spitzer et al., 1993; BES—Gormally et al., 1982), clinical observations of patients' reports of LOC-related feelings and behaviors, as well as multiple discussions with researchers and clinicians familiar with eating disorders. The initial scale was composed of 20-items with two parts. The structure of the ELOCS was modeled after the EDE-Q and therefore each question begins by asking respondents, "During the past 4 weeks, how many times did you . . . ?" Participants were asked to provide an estimate of the number of times in the past 28 days (4 weeks) they experienced an eating episode characterized by a LOC-related feeling or behavior. After answering an open-ended frequency question, participants were prompted with the phrase, "On average, during these times, how much did you . . . ?" and then asked to provide a rating on an 11-point Likert-type scale ranging from 0 (not at all) to 10 (extremely or completely). These questions enabled participants to indicate the degree to which they experienced different feelings or behaviors related to a LOC. These item scores were averaged to produce a total scale score (Item 6b is reverse scored); higher total scale scores reflect greater LOC.

Development and validation of the Eating Loss of Control Scale
(Blomquist, Roberto, Barnes, White, Masheb, & Grilo, 2014)

Análisis de datos

El objetivo de esta sección es describir los procedimientos estadísticos realizados justificando el razonamiento que subyace a la elección de estos. La descripción incluye nombrar el paquete estadístico usado y la versión del mismo, así como detallar las técnicas estadísticas utilizadas y el cumplimiento o no de los supuestos necesarios para el empleo de estas.

Los supuestos por evaluar y la cantidad de detalles por ser explicitados dependen de la técnica particular empleada. Así, por ejemplo, para las evidencias de confiabilidad por formas paralelas, es necesario explicitar los supuestos de paralelismo que permiten definir si las diferentes versiones de una prueba son equivalentes. Ello implica analizar, entonces, si las puntuaciones medias, varianzas y correlaciones con otros instrumentos alcanzadas por las pruebas asumidas como equivalentes en un mismo grupo son iguales⁷.

Si se toma el caso del análisis factorial exploratorio (AFE), en tanto que este plantea como supuestos medidas por lo menos en escala de intervalo y relación lineal entre variables (ítems o áreas del instrumento), se requiere desarrollar un análisis

⁷ Para especificaciones adicionales, se puede revisar el Glosario que se encuentra en este documento. Asimismo, para obtener mayor información respecto a criterios sugeridos y puntos de corte, se puede revisar EFPA (2013) y Evers et al. (2013).

exploratorio de la data que dé cuenta mínimamente de la simetría de la distribución de las puntuaciones de las variables, así como evaluar la matriz de correlaciones o covarianzas de las variables en estudio a través de pruebas como el test de esfericidad de Barlett y el índice de medida de adecuación de la muestra KMO. Resulta necesario igualmente explicitar el tipo de método de estimación o extracción de factores por emplear (factor común o componentes principales) y el tipo de rotación de factores (ortogonal u oblicua), señalando los elementos en los que se sustenta la elección (objetivo del investigador, cumplimiento de supuestos, uso o no de estadísticos de bondad de ajuste). En necesario indicar, igualmente, el criterio o criterios a partir de los cuales se identificará el número de factores por retener (el criterio de Kaiser, la prueba de scree y/o el análisis paralelo)⁸.

In order to establish the factor structure of the OHQ, the 17 items were subjected to EFA, using principal component analysis with varimax rotation. Following the recommendations of Dziubna and Shirkey (1974), before the analysis, we explored the psychometric adequacy of the items. We used multiple criteria to determine the number of factors to retain (Ford, MacCallum, & Tait, 1986; Gorsuch, 1983; Stevens, 1992), the Kaiser criterion (Eigenvalue higher than 1) to extract the number of factors and, to assign items to factors, we considered factor loadings equal to or higher than .40 (Cliff & Hamburger, 1967).

Development and validation of the Occupational Hardiness Questionnaire
(Moreno-Jiménez, Rodríguez-Muñoz, Garrosa, & Blanco, 2014)

The descriptive statistics related to the mean, standard deviation, minimum and maximum scores, skewness, and kurtosis were calculated. We excluded the items whose values in skewness and kurtosis were outside the range -1 to 1. Of the various available strategies for assessing construct validity, we decided to apply factor analysis with the maximum likelihood method and oblique rotation, calculated from a Pearson correlation matrix. For this type of analysis, we used a cross-validation procedure, the first random sample consisting of 1,581 people; it was on this sample that we carried out exploratory factor analysis (EFA). The analysis was initially conducted without specifying a certain number of factors, and then extracting a single factor. As measures of sampling adequacy of the data for conducting factor analysis, we applied the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's sphericity test. We selected the items with a factor loading equal to or greater than .40...

Lastly, we carried out an analysis of the items, calculating the discrimination indexes, we estimated reliability with Cronbach's (1951) alpha coefficient, and we analyzed the differential item functioning (DIF), in which the focus group was men and the reference group was women... The items that showed DIF in both statistics and a standardized mean difference index with negative values were removed (Zwick, Donoghue, & Grima, 1993).

Assessing organizational climate: Psychometric properties of the CLIOR scale
(Peña-Suárez, Muñiz, Campillo-Álvarez, Fonseca-Pedrero, & García-Cueto, 2013)

Resultados

⁸ Para especificaciones adicionales, se puede revisar el Glosario que se encuentra en este documento. Asimismo, para obtener mayor información respecto a criterios sugeridos y puntos de corte, se puede revisar EFPA (2013) y Evers et al. (2013).

En la sección resultados, se presentan únicamente los resultados que responden al problema de investigación planteado al final de la introducción. En ese sentido, solo incluyen los datos que contribuyan a las conclusiones. Para el caso de los resultados no significativos, estos no se reportan, salvo que sean contrarios a la evidencia empírica previa.

Los resultados obtenidos se presentan de forma precisa y sin sesgo, es decir:

- El texto solo describe y considera los resultados encontrados, evitando circunloquios y digresiones
- Se presentan los resultados de forma concreta y precisa omitiendo comentarios.
- Se utiliza un lenguaje científico.
- Se asume que el lector tiene conocimientos profesionales de estadística, por lo que no se revisan conceptos básicos de estadística; es decir, no se citan los procedimientos estadísticos comúnmente utilizados. Asimismo, se justifica el uso de métodos polémicos o novedosos.
- Cuando se reporta estadística inferencial, se provee la información suficiente para que el lector entienda los resultados y analice posibles explicaciones alternativas.
- Los datos por reportar dependen del análisis utilizado.

Cuando se decide llevar a cabo procesos de construcción de ítems o adaptación de un instrumento a un nuevo contexto o población, resulta necesario proporcionar evidencias cualitativas de que la definición semántica del constructo quedó bien representada en los ítems formulados o adaptados. Lo anterior es evidencia de validez de contenido. El investigador debe revelar en los resultados la valoración de los ítems por parte de los jueces, informando con claridad qué ítems originalmente propuestos o de la versión original han sido eliminados o modificados, y por qué. A la vez debe especificar cuál es finalmente la batería de ítems conservada, especificando para cada reactivo la valoración obtenida a partir de los jueces con respecto a su pertinencia y calidad de redacción. Tras el análisis cualitativo de los ítems, resulta relevante reportar el análisis de las propiedades métricas de los mismos, a partir de la presentación de sus distintos índices estadísticos (correlación ítem test corregido y correlación media inter ítem).

Para el caso de las evidencias de confiabilidad por consistencia interna, deben consignarse los coeficientes de confiabilidad de las pruebas, escalas y subescalas según sea el caso. Si la prueba tiene varias escalas o subescalas se coloca el rango de la

confiabilidad desde el valor más bajo al más alto. Si se obtienen evidencias de confiabilidad test-retest, resulta necesario indicar el intervalo de tiempo entre aplicaciones, así como la correlación obtenida y su significancia. Para las evidencias de confiabilidad por formas paralelas, deben explicitarse las correlaciones obtenidas y su grado de significancia, siempre y cuando se hayan alcanzado los supuestos de equivalencia. Si la puntuación de una prueba involucra toma de decisiones y juicios de valoración de las respuestas, se consignan las evidencias de confiabilidad intercalificadores reportando el coeficiente elegido (porcentaje de acuerdo, coeficiente Kappa, correlación intraclase) según la naturaleza de la puntuación que el juez otorga (categórica o continua). Para la correlación intraclase, es necesario indicar, además, la significancia.

Para proveer alguna idea de la distribución de valores de los indicadores de confiabilidad cuando varias escalas componen un instrumento, el EFPA Test Review Model (2013) recomienda que se consigne en una tabla resumen el número de escalas por cada rango de magnitud de los coeficientes ($r < 0.60$, $0.60 \leq r < 0.70$, $0.70 \leq r < 0.80$, $r \geq 0.80$). El juicio respecto a la adecuación de la confiabilidad en el caso de múltiples escalas se basará en el rango en el que se ubique la mediana del número de escalas.

En caso de reportar evidencias de validez sobre la estructura interna del constructo a través de análisis factorial exploratorio, resulta necesario consignar el número de factores del modelo y el porcentaje de varianza explicada por los factores. La matriz factorial final se presenta en una tabla con formato APA⁹ en esta sección, mientras que las comunalidades de los ítems se colocan en un apéndice del documento. Para las evidencias de validez convergente o discriminante, se reporta la correlación entre instrumentos y la significancia de dicha correlación.

Cuando las evidencias de validez se basan en la comparación de puntuaciones medias entre grupos en los que las diferencias de puntajes eran esperables, se describen los datos vinculados a la prueba ANOVA (por ejemplo, $F(6, 543) = 71.778$, $p < .001$ es decir $F(\text{gl}, \text{error}) = F \text{ modelo corregido}, p$).

Si el interés es explorar evidencias de validez predictiva o concurrente, corresponde reportar las correlaciones entre el instrumento y el criterio. Si bien el uso de la correlación es lo más común, la capacidad predictiva puede ser evaluada de modo más

⁹ Para una descripción de las características de la tabla, se puede revisar Nicol y Pexman (2010).

pertinente a través del coeficiente de determinación y los estadísticos vinculados a modelos de regresión lineal. Finalmente, en pruebas de uso clínico, la relación entre el instrumento y el criterio podría implicar obtener información sobre la sensibilidad y especificidad de la prueba. En este caso se reportan las curvas ROC¹⁰.

Las tablas de resultados siguen el formato APA¹¹. Para el caso de los documentos de tesis, la extensión máxima de esta sección es de cinco páginas. No obstante, de ser necesario, puede complementarse la información a través del uso de los apéndices al final del texto de una parte de las cinco páginas disponibles.

Discusión

En esta sección, se retoman brevemente los resultados, se discute la relación de los mismos con estudios previos y se propone una conclusión integradora. Asimismo, se describen las consecuencias prácticas y/o teóricas de los resultados, y la relación de estos con temas prioritarios de la agenda social, esbozando la relación entre la investigación y tales temas sociales. De la misma manera, en esta sección, se explicitan las limitaciones específicas del estudio y se reflexiona sobre su importancia y los estudios que continuarían con la línea de investigación.

En cuanto al contenido, las interpretaciones presentadas en la discusión deben presentar coherencia con los resultados encontrados. Asimismo, a nivel formal, la redacción empleada debe reflejar el carácter especulativo de los resultados obtenidos en la investigación.

La discusión consta de tres partes principales:

1. Una discusión de los resultados obtenidos:
 - Una afirmación inicial que apoya o no las hipótesis planteadas en el estudio, o el resultado obtenido sobre un objetivo.
 - Si las hipótesis no se cumplen, se ofrecen explicaciones *post-hoc*.

¹⁰ Si bien en las investigaciones instrumentales publicadas se reportan las curvas ROC, en el caso de las tesis de licenciatura se pueden reportar solo la sensibilidad y especificidad.

¹¹ Para una descripción de las características de las tablas de acuerdo con el tipo de análisis, se puede consultar Nicol y Pexman (2010).

- Los resultados se comparan con los de anteriores investigaciones. En ese sentido, las diferencias y similitudes con otros estudios son usadas para contextualizar, confirmar y esclarecer las conclusiones.
 - No se repite lo previamente afirmado; cada frase contribuye a la interpretación y a la comprensión del problema.
2. Una discusión sobre las limitaciones o debilidades del estudio:
- Se incluyen potenciales fuentes de sesgo y otras variables que puedan afectar la validez interna del estudio.
 - Se incluyen explicaciones alternativas a los resultados.
 - Se discute la posibilidad de generalización de los resultados.
3. Comentario final sobre la importancia de sus resultados:
- Se retoma brevemente la justificación presentada en la introducción.
 - Se considera el significado teórico, clínico o práctico de los resultados y la base de esas interpretaciones.
 - Se incluye qué problemas se mantienen sin resolver o surgen sobre la base de los resultados.

Para el caso de los documentos de tesis, la extensión máxima de esta sección es de siete páginas y sigue el formato APA.

Referencias¹²

En esta sección, solamente se incluyen las referencias citadas. Estas deben seguir las normas APA.

Apéndices¹³

En los apéndices, se presenta la información que permite profundizar en los datos o examinar interpretaciones alternativas. Además, se incluyen únicamente los instrumentos que no han sido previamente publicados y de los que se posee la autorización escrita del autor, o que sean de autoría propia.

¹² Una descripción detallada de las mismas se encuentra en el capítulo 6 del Manual de publicaciones de la American Psychological Association (APA, 2010).

¹³ Para especificaciones adicionales, se debe revisar el capítulo 2 del Manual de publicaciones de la American Psychological Association (APA, 2010).

Consideraciones éticas

Las investigaciones realizadas deben respetar los principios éticos planteados por el Código de Ética Profesional del Colegio de Psicólogos y la Comisión de Ética para la Investigación con Seres Humanos y Animales. Asimismo, deben tener en cuenta las siguientes consideraciones éticas respecto al consentimiento informado:

- Se debe solicitar consentimiento informado a todos los participantes¹⁴.
- En caso de que el participante tenga 12 años o menos, el consentimiento informado debe ser firmado por los padres o apoderado.

Si no es posible que los padres brinden su consentimiento y la institución tiene autorización para brindarla por ellos, se debe redactar un documento en el que se especifique que la institución está brindando el consentimiento para la aplicación y se responsabiliza por ella. Adicionalmente, en estos casos, el investigador debe enviar una carta que informe a los padres de los niños acerca de la investigación.

Si el participante es mayor de 12 años y menor de 18, se solicita autorización a la institución y el consentimiento informado al participante. Igualmente, el investigador debe enviar una carta que informe a los padres de los niños acerca de la investigación.

En caso de que el participante no se encuentre en pleno uso de sus facultades, el consentimiento informado debe ser firmado por los padres o apoderado. No obstante, si el investigador se encuentra estudiando un tema delicado o sensible para la población, debe enviar un consentimiento informado a los padres.

- En el consentimiento informado, se debe especificar que el investigador es un alumno de pregrado y el compromiso de devolución de datos o accesibilidad de la información, tanto para la institución como para los participantes.

En caso de que, al realizar el análisis de la data, el investigador encuentre casos que presenten niveles altos en ciertas variables que representen una alta probabilidad de riesgo para el participante, estos se deben reportar al asesor para una mayor discusión de las medidas a seguir. Cabe resaltar que, si la muestra seleccionada por el investigador

¹⁴ Para ver el modelo propuesto por la Comisión de Ética para la Investigación con Seres Humanos y Animales, se puede revisar el siguiente vínculo: <http://textos.pucp.edu.pe/pdf/3614.pdf>

pertenece a una población en riesgo o vulnerable, debe presentar el documento planteado por el Comité de Ética para la Investigación con Seres Humanos y Animales¹⁵.

¹⁵ Disponible en el siguiente vínculo: <http://textos.pucp.edu.pe/pdf/3612.pdf>

Aspectos formales¹⁶

A continuación, se presenta una lista de los aspectos principales a tener en cuenta, la cual debe ser complementada con la lectura del manual referido:

- La impresión del documento es por ambas caras.
- Se usa letra Times New Roman, tamaño 12.
- El interlineado es doble; sin embargo, en el documento de tesis, se permite utilizar espacio y medio, y espacio doble luego del título, encabezamiento y citas.
- Se usa sangría para comenzar un párrafo y en la primera línea de un pie de página.
- Para el encabezado, se usa un título abreviado, de un máximo de cincuenta caracteres incluidos los espacios, en cada página del texto. Se imprime en la parte superior de las páginas, justificado al margen izquierdo, con mayúscula en la primera letra.
- La numeración de páginas se coloca en la esquina superior derecha.
- El resumen se presenta en una página separada del texto, al inicio del documento.
- Las referencias y el apéndice comienzan en una página separada.
- Cada sección se inicia en página impar.

¹⁶ Para especificaciones adicionales, se debe revisar el capítulo 2 del Manual de publicaciones de la American Psychological Association (APA, 2010).

GLOSARIO

Análisis factorial: Es una técnica estadística que reproduce las matrices de covarianzas o correlaciones entre variables observadas. En el caso de las pruebas psicológicas, el análisis factorial permite conocer si los ítems o las preguntas de la prueba se agrupan de alguna manera al compartir un significado común. El propósito de la técnica es, en última instancia, extraer el mínimo número de dimensiones que permitan explicar la mayor cantidad de información posible contenida en los datos. Existen dos tipos de análisis factorial: exploratorio y confirmatorio. En el primero de ellos, no se cuenta con una hipótesis, ni se conocen los factores sino que estos son identificados por el método factorial. En cambio, en el análisis confirmatorio se cuenta con un modelo con diversas variables y se contrastan hipótesis para confirmar la existencia de dichos factores.

- **Índice de medida de adecuación de la muestra KMO:** Es uno de los supuestos del análisis factorial. Es una medida de adecuación muestral basada en los coeficientes de correlación observados de cada par de variables y en sus coeficientes de correlación parcial. En el modelo de análisis factorial, los coeficientes de correlación teóricos calculados entre cada par de factores únicos son nulos por hipótesis; por lo tanto, los coeficientes de correlación parcial de las variables deben ser próximos a cero. En el caso de que exista adecuación de la data en un modelo de análisis factorial, la medida KMO será próxima a la unidad; por lo tanto, los valores por debajo de .5 no serán aceptables. Mientras más cercano sea el valor a 1, mejor es la adecuación de los datos a un modelo factorial.
- **Test de esfericidad de Barlett:** Es otro de los supuestos del análisis factorial. Contrasta la hipótesis nula de que las variables no se encuentran correlacionadas entre sí, lo cual indicaría que la matriz de correlaciones es una matriz de identidad y no existen factores comunes; por lo tanto, no sería adecuado aplicar un análisis factorial. En este análisis, se esperaría obtener valores de $p < .05$ para continuar con el análisis factorial.
- **Método de extracción:** Es el método por el cual las variables latentes de interés son generadas. Existen varios métodos de extracción y cada uno representa un algoritmo matemático distinto. Los factores obtenidos en el análisis factorial son el producto entre el método de extracción y la matriz de asociaciones. Durante la extracción, los factores son generados uno por uno; el primer factor es el que da cuenta de la mayor

cantidad de varianza, mientras que los siguientes factores dan cuenta de menos varianza. Por ello, es importante retener la cantidad adecuada de factores, ya que retener demasiados causa que no expliquen un porcentaje de varianza significativa.

- **Análisis de componentes principales:** Se dispone de una muestra de un tamaño determinado acerca de las variables medidas, las cuales inicialmente se encuentran correlacionadas, para luego obtener a partir de ellas un número de variables incorrelacionadas que sean la combinación lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad. Establece qué componentes lineales existen en la data y cómo una variable en particular puede contribuir a cada componente. Este análisis no es un verdadero método de análisis factorial, sino más bien de reducción de datos. Algunos autores critican este método debido a que indican que toma en cuenta toda la varianza, y no discrimina entre la varianza compartida y la única.
- **Máxima verosimilitud:** Proporciona estimaciones de los parámetros que con mayor probabilidad ha producido la matriz de correlaciones observadas si la muestra procede de una distribución normal multivariada. Algunos autores señalan que si la distribución es relativamente normal esta es la mejor opción.
- **Factorización de ejes principales:** Parte de la matriz de correlaciones original con los cuadrados de los coeficientes de correlación múltiple insertados en la diagonal principal como estimaciones iniciales de las comunalidades. Las saturaciones factoriales resultantes se utilizan para estimar de nuevo las comunalidades que reemplazan a las estimaciones previas de comunalidad en la diagonal. Las iteraciones continúan hasta que el cambio en las comunalidades, de una iteración a la siguiente, satisfaga el criterio de convergencia para la extracción. Es recomendado cuando los datos no se ajustan a una distribución normal.
- **Rotación:** La meta de la rotación es simplificar y clarificar la estructura de los datos; esta no puede mejorar los aspectos principales del análisis, pero sí ayuda a mejorar la interpretación de las variables latentes. Varimax es la más reportada y se prefieren las rotaciones ortogonales dado que son más fáciles de interpretar. En los estudios, se debe especificar y justificar el tipo de rotación usada.
 - **Rotación ortogonal:** Se asume que la correlación entre las variables latentes o factores es cero; por ello, supone que los factores son independientes. Los ejes

se rotan de forma que persevere la incorrelación de los factores. En las rotaciones ortogonales, se interpreta la matriz de factores rotados.

- **Varimax**: Este método obtiene los ejes de los factores maximizando la suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor. Es la rotación que mejor se adapta a la estructura simple.
- **Quartimax**: Este método maximiza la difusión de las cargas factoriales para una variable a través de todos los factores. Se recomienda cuando el número de factores es elevado.
- **Equamax**: Combina el método Varimax con el Quartimax. Minimiza tanto el número de variables como el número de factores necesario para explicar la variable.
- **Rotación oblicua**: Permite que los factores se correlacionen y detecta factores generales; por ello, supone cierto grado de relación entre los factores. El traslape entre los factores puede significar que hay más áreas de generalización que los factores primarios, por lo que se recomienda en estos extraer factores de orden superior. En las rotaciones oblicuas, se utiliza la matriz de patrones para examinar las cargas factoriales y la matriz de correlaciones de factores para revelar alguna correlación entre factores.
 - **Oblimin**: Usa valor Delta, que determina la oblicuidad que pueden llegar a alcanzar los factores. El valor por defecto es 0 (mayor oblicuidad) y este puede ser cambiado por un número menor o igual a .8.
 - **Promax**: Permite que los factores estén correlacionados. Usa valor Kappa para controlar el cálculo de la rotación, el valor por defecto es 4, el cual es adecuado para la mayoría de análisis. Suele ser útil para grandes conjuntos de datos.

Confiabilidad: Se refiere al grado de precisión y consistencia de la medida; por ello, no es una característica del instrumento, sino de las puntuaciones obtenidas en una muestra determinada. El coeficiente de confiabilidad es un índice de confianza, una proporción que indica la razón entre la varianza de la puntuación verdadera en una prueba y la varianza total. Supone que los puntajes de un instrumento son capaces de diferenciar adecuadamente a los sujetos en el constructo por medir, ya que el instrumento da cuenta de la variabilidad de los sujetos y la varianza de las puntuaciones. Indica hasta qué punto los resultados que se han obtenido a partir de la aplicación del instrumento han sido

afectados por el error que se ha podido cometer al momento de medir. Si un instrumento de medida es preciso, obtendrá un error pequeño y las medidas repetidas serán similares.

- **Método test-retest:** Se conoce también como coeficiente de estabilidad y es una estimación de la confiabilidad obtenida al correlacionar pares de puntuaciones de las mismas personas en dos aplicaciones diferentes de la misma prueba. Indica el grado en que se pueden ver afectadas las puntuaciones de una prueba por las fluctuaciones que se pueden producir en el intervalo temporal. Este intervalo siempre debe especificarse en el estudio. Este tipo de confiabilidad se suele emplear con constructos que son estables, ya que factores como la maduración y el aprendizaje podrían interferir en las mediciones. El efecto del aprendizaje puede ser paliado si, entre ambas aplicaciones de la prueba, se deja transcurrir un tiempo suficiente. Se considera que las puntuaciones de un instrumento son confiables si en las dos mediciones se obtienen resultados similares.
- **Método de formas alternas o paralelas:** También llamado coeficiente de equivalencia y mide la consistencia de las respuestas a diferentes muestras de reactivos utilizando dos versiones del mismo instrumento que son similares entre sí. Estas dos versiones son aplicadas a los sujetos en el mismo momento. Si se obtiene un coeficiente alto, se puede interpretar que existe equivalencia entre las puntuaciones de las formas de la prueba y que estas pueden ser intercambiables. No obstante, es importante tomar en consideración si las respuestas de la primera forma influyen en la segunda.
- **Consistencia interna:** Los coeficientes de consistencia interna son los más empleados. Se refieren a la consistencia en las respuestas de los ítems y expresan el grado en que cada ítem de una prueba muestra una equivalencia adecuada con el resto de ítems; es decir, mide con el mismo grado el constructo medido. Si todos los ítems se encuentran midiendo lo mismo, se puede concluir que son sumables en una puntuación total única que representa un rasgo.
 - o **Método de dos mitades:** Es el grado de equivalencia y paralelismo de dos partes de una prueba. Se puede evaluar calculando la correlación entre las puntuaciones de los sujetos en estas dos mitades aplicadas una sola vez, como si fueran dos formas paralelas. Se puede dividir una prueba asignando al azar los ítems a una u otra mitad de la prueba o colocando los ítems pares en un grupo y los impares en otro; no se recomienda dividir la prueba simplemente por la mitad.

- **Método de covariación de ítems:** Se basa en la magnitud de las covariaciones entre los ítems de la prueba en relación con la varianza total; es decir, mide el paralelismo de cada ítem respecto a la prueba en conjunto. Este índice de consistencia entre los reactivos es útil para evaluar la homogeneidad de la prueba. Cuando en una prueba las opciones de respuesta son politómicas, se emplea el Alfa de Cronbach, mientras que, cuando son dicotómicas, se utiliza el coeficiente Kuder-Richardson (20 o 21).
- **Confiabilidad intercalificadores:** Conocido también como coeficiente de concordancia, es el grado de acuerdo o consistencia que existe entre dos o más evaluadores o calificadores. Se suele emplear cuando en la medición existe un componente subjetivo. Si el coeficiente de confiabilidad es muy alto, se puede interpretar que las puntuaciones de la prueba pueden ser derivadas de manera consistente y sistemática por varios evaluadores que hayan tenido la capacitación requerida para la calificación de la prueba. El método más común de obtener este tipo de confiabilidad es registrando el porcentaje de veces en que dos o más evaluadores han presentado acuerdo; sin embargo, este porcentaje no considera el nivel de acuerdo que se puede conseguir debido al azar. Por ello, se recomienda el empleo del coeficiente de correlación intraclase o el estadístico kappa, los cuales superan esta limitación.

Validez: Consiste en el grado en el que la evidencia y la teoría apoyan las interpretaciones de los tests de acuerdo con el uso que se va a hacer de esos tests. El proceso de validación es la recolección de evidencias que apoyen el tipo de inferencias que se harán a partir de los puntajes obtenidos con el instrumento. Por lo tanto, la validación es un proceso; no es un cálculo, y no hace referencia a la prueba psicológica, sino a las interpretaciones y decisiones que se toman a partir de las puntuaciones de esta. Por ello, es específica al uso que se le brinde al instrumento.

- **Evidencias externas de validez del instrumento:** Es el tipo de evidencia más empleado en el proceso de validación y se define como el grado en que las relaciones de la medida obtenida por el instrumento con variables externas reflejan las relaciones esperadas en la teoría del constructo, en la cual se deberán especificar estas relaciones. Este aspecto de la validez integra la evidencia convergente y divergente, las relaciones test-criterio y los estudios de generalización de la validez.

- **Evidencia convergente:** Es la relación entre la prueba y otras medidas del mismo constructo. Se espera encontrar correlaciones elevadas si el test es una buena medida del constructo.
- **Evidencia divergente:** Es la relación entre la prueba y otras medidas de constructos diferentes. Se espera encontrar correlaciones inferiores con otros constructos a las que muestra con las medidas del propio constructo si el test es una buena medida del constructo. Además, estas correlaciones serán bajas o nulas con las medidas con las que se hipotetiza que no está correlacionado.
- **Relaciones test-criterio:** Es el proceso que busca evaluar si los puntajes en la prueba pueden ser empleados para predecir una conducta específica, denominada criterio. Adquiere gran relevancia en contextos en los que es fundamental la precisión con que se efectúa este tipo de predicciones. El criterio empleado —que pueden ser otras pruebas— suele ser conductas que son relevantes y de interés en situaciones aplicadas. Cabe resaltar que las relaciones entre estas conductas y el constructo también estarán previstas en la red nomológica del constructo. Los diseños utilizados para la obtención de los índices de validez dependen del tiempo transcurrido entre la recogida de datos en el test y el criterio, la cual puede ser concurrente, predictiva o retrospectiva.
 - Concurrente: Se aplica el instrumento al mismo tiempo en que se evalúa el criterio.
 - Predictiva: Se aplica el instrumento en un momento y el criterio se evalúa después de un tiempo determinado.
 - Retrospectiva: El criterio ha sido registrado anteriormente a la aplicación del instrumento.
- **Evidencia de decisión (sensibilidad y especificidad):** Se emplea cuando el instrumento tiene como objetivo el diagnóstico y la clasificación de los sujetos en dos grupos, con diagnóstico positivo y diagnóstico negativo. Se compara la prueba con un método de diagnóstico suficientemente contrastado y se distinguen dos procesos: sensibilidad y especificidad. La sensibilidad de la prueba hace referencia a la capacidad de esta en la detección de verdaderos positivos, mientras que la especificidad es la capacidad de determinar diagnósticos negativos verdaderos.
- **Generalización de la validez:** Su objetivo es la generalización de los resultados sin la necesidad de realizar nuevos estudios de validación. El método de

generalización de la validez más empleado es el meta-análisis, el cual unifica los resultados de diversas aplicaciones de un mismo instrumento mediante técnicas específicas. De esta manera, intenta estimar un promedio de validez corrigiendo los efectos de los diversos factores que han podido influenciar en los resultados de cada estudio revisado. Entre estos factores se pueden encontrar: los errores de medida, la dicotomización, la varianza debida a factores extraños, el error muestral, etc.

- **Evidencias internas de validez del instrumento:**

- **Validez de contenido:** Es el proceso por el cual se obtienen evidencias de validez de la relación entre el contenido de la prueba psicológica y el constructo que intenta medir; es decir, es el grado en el cual los ítems que componen un instrumento son una muestra adecuada y representativa de un constructo. Cabe resaltar que el contenido de un test no solo está formado por los enunciados, sino también por el formato de los ítems o tareas que lo conforman y el peso asignado a cada uno de estos. Este tipo de evidencia genera un impacto en la precisión con que se pueden realizar inferencias acerca de la puntuación de la persona a partir de la puntuación en la muestra de ítems que constituyen la prueba. Entre las evidencias de validez de contenido, se pueden incluir los análisis lógicos o empíricos de la adecuación con la que el contenido de la prueba representa el universo de posibles manifestaciones del constructo. Para que este proceso se realice adecuadamente, se deben especificar los límites del dominio del constructo por evaluar y determinar los atributos que debe poner de relieve la tarea de evaluación. Este tipo de evidencias se suelen expresar a través de juicios de expertos. Los jueces seleccionados deben contar con un amplio conocimiento del constructo medido en el instrumento, y deben tener conocimientos sobre la construcción de pruebas psicológicas y experiencia con la población a la cual está destinada la prueba.
- **Dimensionalidad:** Evalúa el grado en que las relaciones entre los ítems y los componentes de la prueba conforman el constructo sobre el cual se basarán las interpretaciones. Es el estudio del número de factores o dimensiones subyacentes a un conjunto determinado de variables y su objetivo es determinar el mínimo número de estructuras necesario para explicar la máxima varianza observada. Para obtener este tipo de evidencia de validez, se suele emplear el análisis factorial.

Referencias

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington D.C., Estados Unidos: AERA
- American Psychological Association (2010). *Manual de publicaciones de la American Psychological Association* (3ra ed.). México D.F., México: Manual Moderno.
- Clark, L., & Watson, D. (2003). Constructing validity: Basic issues in objective scale development. En A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (3rd ed.) (207-231). Washington, D. C., Estados Unidos: APA.
- De Winter, J. C. F., Dodou, D., & Wieringa, P.A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*, 147-181.
- European Federation of Psychologists Associations (2013). EFPA Review model for the description and evaluation of psychological and educational tests. Test review form and notes for reviewers. Version 4.2.6. *EFPA Board of Assessment Document*, *110c*, 1-72.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2004). *Análisis Multivariante*. Madrid, España: Pearson.
- Hogarty, K., Hines, C., Kromrey, J., Ferron, J., & Mumford, K. (2005). The Quality of Factor Solutions in Exploratory Factor Analysis: The Influence of Sample Size, Communalities, and Overdetermination. *Educational and Psychological Measurement*, *65*(2), 202-226 DOI: 10.1177/0013164404267287
- Keiffer, K., & MacDonald, G. (2011). Exploring factors that affect score reliability and variability in the ways of Coping Questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Individual Differences*, *32*(1), 26-38. DOI: 10.1027/1614-0001/a000031
- Martínez-Arias, R. (1995). *Psicometría: teoría de los test psicológicos y educativos*. Madrid, España: Síntesis.
- Nicol, A., & Pexman, P. (2010). *Presenting your findings: A practical guide for creating tables*. Washington, D. C., Estados Unidos: American Psychological Association.
- Osterling, S. J. (1989). *Constructing Test Items*. Londres, Inglaterra: Kluwer Academic Publishers.
- Stevens, J. (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2007). Multivariate analysis of variance and covariance. *Using multivariate statistics*, 3, 402-407.

Referencias de investigaciones usadas para los ejemplos de este documento

- Alfaro, J., Castellá, J., Bedin, L., Abs, D., Casas, F., Valdenegro, B., & Oyarzún, D. (2014). Adaptación del Índice de Bienestar Personal para adolescentes en lengua española y portuguesa. *Universitas Psychologica*, 13(1), 1-23. doi: 10.11144/Javeriana.UPSY13-1.aibp
- Blomquist, K., Roberto, C., Barnes, R., White, M., Masheb, R., & Grilo, C. (2014). Development and validation of the Eating Loss of Control Scale. *Psychological Assessment*, 26(1), 77-89. doi: 10.1037/a0034729
- Boyes, M., & Cluver, L. (2013). Performance of the Revised Children's Manifest Anxiety Scale in a sample of children and adolescents from poor urban communities in Cape Town. *European Journal of Psychological Assessment*, 29(2), 113-120. doi: 10.1027/1015-5759/a000134.
- Durak, M., & Senol-Durak, E. (2013). The development and psychometric properties of the Turkish Version of the Stress Appraisal Measure. *European Journal of Psychological Assessment*, 29(1), 64-71. doi: 10.1027/1015-5759/a000079.
- Fernández, T., Medina, S., Herrera, I., Rueda, S., & Fernández, A. (2011). Construcción y validación de una escala de autoeficacia para la actividad física. *Revista Española de Salud Pública*, 85(4), 405-417.
- Gámez, W., Chmielewski, M., Kotov, R., Ruggero, C., Suzuki, N., & Watson, D. (2014). The Brief Experiential Avoidance Questionnaire: Development and Initial Validation. *Psychological Assessment*, 26(1), 35-45. doi: 10.1037/a0034473
- Moreno-Jiménez, B., Rodríguez-Muñoz, A., Garrosa, E., & Blanco, L. (2014). Development and validation of the Occupational Hardiness Questionnaire. *Psicothema*, 26(2), 207-214. doi: 10.7334/psicothema2013.49
- Peña-Suárez, E., Muñiz, J., Campillo-Álvarez, A., Fonseca-Pedrero, E., & García-Cueto, E. (2013). Assessing organizational climate: Psychometric properties of the CLIOR scale. *Psicothema*, 25(1), 137-144. doi: 10.7334/psicothema2012.260.
- Olivares-Faúndez, V. E., Mena-Miranda, L., Jélvez-Wilke, C., & Marcía-Sepúlveda, F. (2014). Validez factorial del Maslach Burnout Inventory Human Services (MBI-

HSS) en profesionales chilenos. *Universitas Psychologica*, 13(1), 145-159. doi: 10.11144/Javeriana.UPSY13-1.vfmb

Ruch, W., Weber, M., Park, N., & Peterson, C. (2014). Character Strengths in Children and Adolescents: Reliability and Initial Validity of the German Values in Action Inventory of Strengths for Youth (German VIA-Youth). *European Journal of Psychological Assessment*, 30(1), 57-64. doi: 10.1027/1015-5759/a000169

Referencias empleadas para el Glosario

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

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111–150. doi: 10.1207/S15327906MBR3601_05

Cohen, R., & Swerdlik, M. (2006). *Pruebas y evaluación psicológica. Introducción a las pruebas y a la medición*. Madrid, España: Mc Graw Hill.

Costello, A., & Osborne, J. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment Research & Evaluation*, 10(7), 1-9.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema*, 15(2), 315-321.

Field, A. (2000). *Discovering Statistics using SPSS for Windows*. Londres, Inglaterra: Sage Publications.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, 286–299.

Frias-Navarro, D., & Soler, M. P. (2012). Prácticas del Análisis Factorial Exploratorio (AFE) en la investigación sobre conducta del consumidor y marketing. *Suma Psicológica*, 19(1), 47-58.

Hair, J. F., Anderson, R. L., Tatham, R. L., & Black, W. C. (1999). *Análisis Multivariante*. Madrid, España: Prentice Hall Iberia.

Hunter, J. E., & Schmidt, F. L. (1991). Meta-analysis. En R. K. Hambleton & J. N. Zaal (Eds.), *Advances in Educational and Psychological Testing: Theory and*

- Applications* (pp. 157-184). Boston, MA, Estados Unidos: Kluwer Academic Publishers.
- Kaplan, R., & Saccuzzo, D. (2005). *Psychological Testing: Principles, Applications, and Issues*. Belmont, CA, Estados Unidos: Thomson Wadsworth.
- Litwin, M. (2003). *How to assess and interpret survey psychometrics*. Thousand Oaks, CA, Estados Unidos: Sage Publications.
- Martínez, M. R., Hernández, M. J., & Hernández, M.V. (2006). *Psicometría*. Madrid, España: Alianza Editorial.
- Martínez-Arias, R. (1995). *Psicometría: teoría de los test psicológicos y educativos*. Madrid, España: Síntesis.
- Martínez, R. (2008). *El análisis multivariante en la investigación científica*. Madrid, España: La Muralla.
- Meneses, J., Barrios, M., Bonillo, A., Cosculluela, A., Lozano, L. M., Turbany, J., & Valero, S. (2013). *Psicometría*. Barcelona, España: Editorial UOC.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256.
- Paz, M. D. (1996). Validez. En J. Muñiz (Ed.) *Psicometría* (pp. 49-103). Madrid, España: Universitas.
- Pérez, C. (2004). *Técnicas de análisis multivariante de datos: Aplicaciones con SPSS*. Madrid, España: Pearson Educación.
- Popham, W. J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders* (3rd ed). Boston, MA, Estados Unidos: Allyn and Bacon.
- Santesteban, C. (2009). *Principios de Psicometría*. Madrid, España: Síntesis.
- Steger, M. F. (2006). An illustration of issues in factor extraction and identification of dimensionality in psychological assessment data. *Journal of Personality Assessment*, 86, 263-272. doi: 10.1207/s15327752jpa8603_03
- Thompson, B. (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA, Estados Unidos: Sage Publications.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington D.C., Estados Unidos: American Psychological Association.