Revista Argentina de Ciencias del Comportamiento
ISSN: 1852-4206
paulaabate@gmail.com
Universidad Nacional de Córdoba
Argentina

Merino-Soto, César; Calderón-De la Cruz, Gustavo A.; Gil-Monte, Pedro; Juárez-García, Arturo
Validez sustantiva en el marco de la validez de contenido: Aplicación en la escala de Carga de Trabajo.
Revista Argentina de Ciencias del Comportamiento, vol. 13, núm. 1, 2021, Enero-, pp. 81-92
Universidad Nacional de Córdoba
Córdoba, Argentina

Revista Argentina de Ciencias del Comportamiento

ISSN 1852-4206

Abril 2021, Vol. 13, N°1, 81-92

revistas.unc.edu.ar/index.php/racc

Validez sustantiva en el marco de la validez de contenido: Aplicación en la escala de Carga de Trabajo.

Merino-Soto, César[a] *, Calderón-De la Cruz, Gustavo A.[a], Gil-Monte, Pedro[b] , Juárez-García Arturo[c]

Artículo Original

Resumen

La validez de contenido requiere evaluar la claridad y relevancia de los ítems, pero el contenido sustancial del mismo también. El objetivo del estudio fue utilizar la estrategia de validez sustantiva para evaluar el contenido de una Escala de Carga de Trabajo. Participaron 30 trabajadores (Lima, Perú), de instrucción superior y con experiencia laboral mayor a un año. Los participantes juzgaron la correspondencia de los ítems con su constructo, y fue también contrastado con otros constructos (conflicto de rol, afrontamiento y autoeficacia laboral); adicionalmente, se evaluó la claridad de los ítems. Se halló niveles elevados de claridad de los ítems; y la validez sustantiva fue satisfactoria para evaluar la divergencia con otros constructos en todos los ítems. Se discute la inclusión de los participantes no-expertos en los métodos de validez de contenido, para la diferenciación conceptual de los ítems, y en constructos asociados a los factores psicosociales.

Palabras clave:
Validez sustantiva, validez de contenido, carga de trabajo, factores psicosociales, trabajo.

Abstract

The content validity requires to assess not only the clarity and relevance of the items, but also the substantial content. The aim of the study was to use the substantive validity strategy to evaluate the content of the UNIPSICO's workload scale. 30 workers from Lima (Perú) with undergraduate or graduate studies and with one year of work experience participated. Participants judged the correspondence of the workload items with their construct, and it was also contrasted with other constructs (role conflict, coping and labor self-efficacy); additionally, the clarity of the items was evaluated. High levels of item's clarity was found; and substantive validity of all items was satisfactory to evaluate the divergence with other constructs. We discuss the importance of including non-expert participants in content validity methods, conceptual differentiation of items, and in constructs associated with psychosocial factors.

Keywords:
substantive validity, content validity, workload, psychosocial factors, work.

Tabla de Contenido

Recibido el ejemplo:22 de Diciembre de 2018; Aceptado el 14 de Abril de 2020
Editaron este artículo: Cecilia Reyna, Paula Abate, Gabriela Raynaudo y Florencia Dadam

Introducción

La participación del examinado en su función de juez, como fuente de valoración de algunas características psicométricas de los instrumentos, ha sido reconocida y establecida desde hace años (Anderson & Gerbing, 1991; Fiske, 1967; Hinkin & Tracey, 1999; Morse, D. T. & Morse, 2002; Nevo, 1995). Es habitual incluirlos en la etapa de generación de ítems, así como en la evaluación de la relevancia y claridad de mediciones asociadas a diversos constructos (Magasi et al., 2012; van Kooten, Terwee, Kaspers, Raphaële, & van Litsenburg, 2016). Generalmente forma parte de la estrategia recomendada para asegurar la validez de contenido desde múltiples perspectivas y para verificar la replicabilidad de los resultados (Rothman et al., 2009). La multiplicidad de perspectivas para evaluar la validez de los ítems es una práctica recomendable pues pueden existir discrepancias entre las valoraciones de grupos de sujetos elegidos como expertos. Una situación así ha sido demostrada recientemente (e.g., Merino-Soto, 2016), en que se resaltan a) la existencia de

[a] Universidad de San Martín de Porres, Escuela de Psicología, Instituto de Investigación de Psicología, Lima, Perú.
[b] Universitat de València, Unidad de Investigación Psicosocial de la Conducta Organizacional, Valencia, España.
[c] Universidad Autónoma del Estado de Morelos, Centro de Investigación Transdisciplinar en Psicología, Cuernavaca, México.
*Enviar correspondencia a: Merino-Soto, C.; E-mail: sikayax@yahoo.com.ar

Merino-Soto, C., Calderón-De la Cruz, G., Gil-Monte, P., Juárez-García, A. (2021). Validez sustantiva en el marco de la validez de contenido: Aplicación en la escala de Carga de Trabajo. *Revista Argentina de Ciencias del Comportamiento, 13*(1), 81-92

discrepancias, y b) lo apropiado de incluir varios grupos de evaluadores de ítems que sean intrínsecamente diferentes.

Desde este marco argumentativo, se deduce que la credibilidad del contenido de los ítems de una medida no solo requiere del juicio de expertos cuya posición frente al constructo es realmente indirecta y moderadora. Por el contrario, o complementariamente, se requiere verificar la significancia de tales contenidos para la vida en los examinados o de los participantes-objetivo, pero desde su propio juicio (Anderson & Gerbing, 1991; Lasch et al., 2010; Magasi et al., 2012; Patrick et al., 2011). Un procedimiento que incluya a los examinados como "jueces" es altamente recomendado para valorar la eficacia de una escala o instrumento generado en un contexto cultural diferente al que se aplica, asegurando así la total equivalencia conceptual, semántica, idiomática y experiencial del constructo evaluado (Beaton, Bombardier, Guillemin, & Ferraz, 2000). Esto también crea el contexto para usar una metodología que evalúe la convergencia del juicio de jueces expertos y de los usuarios participantes, siendo estos últimos los que podrían decidir finalmente el valor práctico o conceptual de los ítems.

El juicio de expertos tiende a ser el método más frecuente en la evaluación de la validez de contenido (Urrutia, Barrios, Gutiérrez, & Mayorga, 2014), y generalmente los expertos son seleccionados con base a su experticia conceptual (conocimientos especializados) más que por su experiencia (vivencias concretas en relación al constructo evaluado). Por otro lado, en algunas áreas de investigación y desarrollo de instrumentos es usual incluir a los mismos sujetos o examinados como fuentes de valoración de los ítems (por ejemplo, DeWalt, Rothrock, Yount, & Stone, 2007; Magasi et al., 2012), mientras en otras ésta no parece ser una práctica habitual. Por ejemplo, en la investigación organizacional publicada pueden hallarse la predominante participación de jueces expertos, pero no de participantes examinados. Por tanto, la falta de evidencias de la valoración del examinado en contenidos críticos del constructo puede ser una amenaza para la validez de contenido (Rothman et al., 2009).

Comúnmente, la relevancia y la claridad son componentes formales en la evaluación del contenido de los ítems (DeWalt et al., 2007; Mokkink et al., 2006, 2010); otros sugieren agregar la integridad semántica y la aceptabilidad (Manson, 1997; Mora-Ríos, Bautista-Aguilar, Natera, & Duncan, 2013). Pero no es menos importante otro componente evaluativo del contenido, identificado como la *sustancialidad* o *validez sustantiva* (Anderson & Gerbing, 1991). La validez sustantiva es definida como "el grado en que una medida se juzga como teóricamente vinculada, o refleja, algún constructo de interés" (Anderson & Gerbing, 1991, pp. 732), y está directamente vinculada con el planteamiento de Loevinger (1957) sobre grado de representatividad del constructo mediante la examinación de los ítems que lo exploran, la cual es igualmente importante que otros componentes de la validez. Esta representatividad es la característica esencial de la sustancialidad del contenido de los ítems, y es la primera fuente de respalda de la validez de constructo (Loevinger, 1957). En ese sentido, el método de verificación de la validez sustantiva responde a la exploración de la estructura interna de un instrumento, que es habitualmente aplicable de manera preliminar de la creación o adaptación de un instrumento de medición (Hanley et al., 2015; Zabkar, 2000). Por otro lado, el concepto utilizado aquí, no es precisamente igual, pero sí vinculado, con el concepto de Messick (1995) de la validez sustantiva, referida a los procesos latentes que se activan en el examinado frente a los estímulos del instrumento. En ambas definiciones, la de Anderson y Gerbing (1991) y Messick (1995), existe una comunalidad clara: que la fuente para evaluar la sustancialidad de los ítems proviene de los participantes, no de los jueces expertos. En un sentido particular, el participante provee de la información central para evaluar la validez sustantiva, y en un sentido general, este proceso está subsumido en la validez de contenido.

La validez sustantiva parece una alternativa recomendable para minimizar los potenciales problemas de ambigüedad y sesgo en la especificación de modelos fuertes evaluados por el análisis factorial confirmatorio (Anderson & Gerbing, 1991). El motivo de esto es que permite probar la estructura interna preliminarmente, antes de la aplicación de métodos estadísticos en una muestra grande. Los estudios que han obtenido evidencias de validez sustantiva, mediante la diferenciación del constructo de interés frente a otros constructos, han logrado buenos resultados

al respecto (autoeficacia, Chen, Gully, & Eden, 2001; liderazgo de servicio, Farrel, Souchon, & Durden, 2003; regulación de la ansiedad, Hanley et al., 2015; reputación, Helm, Eggert, & Garnefeld, 2010). La operacionalización de la validez sustantiva, de acuerdo al enfoque de Anderson & Gerbing (1991), consiste en solicitar a los mismos examinados que determinen la correspondencia comparativa de los ítems respecto al constructo de interés y frente a otros constructos presentados. Fundamentalmente, esta actividad requiere discriminar el significado de los ítems frente a varios constructos teóricamente relacionados.

En el presente escrito, ejemplificaremos la aplicación del método de la validez sustantiva como primera fase indispensable de la validación del constructo de carga de trabajo para el contexto peruano, pero con modificaciones que permitirán fortalecer esta estrategia de validez. Esta aproximación no demerita la importancia de explorar otros indicadores psicométricos y de validez en general, sin embargo, en este estudio se decidió dar prioridad en aplicar un procedimiento novedoso, con potencial impacto en la práctica de estudios psicométricos futuros, enfocado más en la metodología que en el instrumento, y sin riesgo de dispersar la atención y el espacio a evidencias de estructura o de confiabilidad de la escala de carga de trabajo a utilizar. Estas evidencias previas ya se han explorado en estudios publicados (Calderón-De la Cruz, Merino-Soto, Juárez-García, & Jiménez-Clavijo, 2018), y el presente estudie añade otra evidencia más.

La carga de trabajo está referida al nivel de exigencia de la tarea en relación a las capacidades que dispone el trabajador para llevarlas a cabo (Tovalín & Rodríguez, 2013). Desde esta concepción, la carga de trabajo puede implicar dos aspectos (Gil-Monte, 2014): la carga de trabajo cuantitativa (la cantidad de tareas laborales que el trabajador debe realizar dentro de un periodo de tiempo delimitado), y la carga de trabajo cualitativa (relacionada con la complejidad en el contenido de la tarea y la falta de recursos para su realización. La carga de trabajo es entendida como un factor psicosocial en el trabajo, pues es producto de la interacción entre las condiciones laborales y las experiencias y percepción de los empleados (Juárez-García, 2007), que, dependiendo de sus sinergias, trae como resultante consecuencias favorables o desfavorables en la calidad vida y la salud de los trabajadores, y en la productividad de la organización laboral (Gil-Monte, 2012, 2014; Meliá et al., 2006).

El constructo de carga de trabajo es particularmente importante para el presente estudio debido principalmente a la susceptibilidad de su presencia en toda ocupación laboral. Dicho constructo ha sido explorado desde la perspectiva de riesgo psicosocial y en específico, de la evaluación de las demandas laborales (Gil-Monte, 2016), recibiendo también la denominación de sobrecarga de trabajo, referido a la percepción del trabajador sobre el nivel de exceso y la complejidad de la tarea que debe realizar en su trabajo, y la carencia de recursos para afrontarlo satisfactoriamente dentro de un tiempo previsto (Kirch, 2008; Veloutsou & Panigyrakis, 2004).

Entre los grupos ocupacionales afectados por la sobrecarga de trabajo se reporta el personal de enfermería (Gil-Monte, García-Juesas, & Caro, 2008), médicos (Patlán, 2013), personal administrativo (Gil-Monte, López-Vílchez, Llorca-Rubio, & Sánchez, 2016) y docentes escolares (Byrne, 1994) y docentes universitarios (Unda et al., 2016). Igualmente estudios indican que su impacto implica una serie de perjuicios en los trabajadores y en la organización laboral desencadenando una baja autoeficacia laboral (Gil-Monte & Peiró, 1997), deterioro emocional (Gil-Monte & García-Juesas, 2008), el desarrollo del síndrome de burnout (Greenglass, Burke, & Moore, 2003), suicidios y abandono del trabajo (Organización Internacional del Trabajo, 2016), conflictos en la vida laboral (Geurts & Demerouti, 2003; Skinner & Pocock, 2008), conflictos familiares (Frone, Yardley, & Markel, 1997), problemas de salud (Gil-Monte et al., 2016), disminución de la productividad e incremento del absentismo y la alta rotación (Bacharach, Bamberger, & Conley, 1990) y adicción al trabajo (Kanai & Wakabayashi, 2001).

Puesto que son los propios trabajadores quienes están inmersos en las vivencias vinculadas al factor psicosocial de carga de trabajo, la obtención de sus valoraciones sobre el constructo será relevante pues sus experiencias o percepciones directas son información de primera mano. El investigador puede inspeccionar estas vivencias respecto al constructo específico o verificar sus relaciones con constructos

teóricamente vinculables como aquellos que se incorporan dentro de los factores psicosociales (e. g., conflicto de rol) para ayudarse a decidir sobre los límites conceptuales, hacer distinciones entre constructos, y verificar que los indicadores elegidos son representativos. Más aún, si los trabajadores pueden expresar sus respuestas a estas indagaciones mediante una estrategia estructurada incorporadas en métodos cuantitativos, se puede alcanzar una más precisa comprensión del constructo. Por lo tanto, el propósito del presente estudio es evaluar la representatividad del contenido de una escala de carga de trabajo mediante las distinciones conceptuales que pueden realizar los trabajadores (en su carácter de jueces), a través del procedimiento de validez sustantiva.

Método

Participantes

Fueron 30 adultos trabajadores entre 24 y 40 de edad (M = 29.20; DE = 4.55), todos de varias carreras, como psicología (n = 40%), medicina (n = 10 %), docencia (n = 10%), abogacía, (n = 6.7%), entre otros. El género se distribuyó similarmente entre mujeres (n = 50%) y varones. El último grado de instrucción alcanzado se distribuyó como bachilleres (n = 56.7%), magister (n = 23,3%), doctorado (n = 6.7%) y técnico (n = 13.3%). Los criterios de selección fueron que aceptaran participar voluntariamente, sean trabajadores con un año mínimo de experiencia, que posean estudios universitarios o técnicos de la ciudad de Lima Metropolitana, y que la distribución final de la muestra sea homogénea respecto al género. La heterogénea distribución de la muestra respecto a la edad y carreras se debió para obtener un amplio rango de experiencias y percepciones no asociadas a una sola área de trabajo; de esta manera, se puede representar mejor las variadas experiencias en el mundo real (May & Warren, 2001; Shepard, Jensen, Schmoll, Hack, & Gwyer, 1993) y se ajustada a la universalidad de la variable en el contexto del trabajo. El muestreo fue no probabilístico, consistiendo en la selección de los participantes a través de la disponibilidad y acceso directo por parte de los investigadores.

Instrumento

Escala de carga de trabajo (Gil-Monte, 2016). Es una medida unidimensional incluida en la Batería UNIPSICO para la evaluación de los factores psicosociales de demandas (Gil-Monte, 2016). La escala de carga trabajo está compuesta por 6 ítems cuyos contenidos exploran la cantidad de tareas laborales a efectuar en un tiempo determinado (carga de trabajo cuantitativa; e. g., *¿Ha tenido que hacer más de una cosa a la vez?)* y cómo el trabajador procesa la información respecto a la dificultad de la tarea laboral (carga de trabajo cualitativa; e. g., *¿Piensa que tiene que hacer un trabajo demasiado difícil para usted?),* las respuestas están cuantificadas en una escala ordinal de cinco puntos *(entre nunca, raramente: algunas veces al año, a veces: algunas veces al mes, frecuentemente: algunas veces por semana, y muy frecuentemente: todos los días*) sobre la frecuencia en que ocurren las situaciones listadas. El instrumento de carga de trabajo reporta evidencias de validez de estructura interna y referido al criterio predictivo siendo sus niveles elevados vinculados a problemas de salud, igualmente, se reportan valores altos de confiabilidad a través de la consistencia interna para cada uno de los ítems (Gil-Monte, 2016). En trabajadores peruanos, una reciente validación halló resultados satisfactorios de su estructura interna, confiabilidad (α = .80, ω = .80) y relaciones con otros constructos (Calderón-De la Cruz et al., 2018)

Procedimiento

El material de aplicación fue elaborado en un formato electrónico, que contenía un apartado de presentación de la investigación el cual se incluía el objetivo del estudio, el consentimiento informado y el formulario para la Validez de Contenido (FCV). Los participantes-jueces fueron invitados a un ambiente equipado con computadoras que contenía el formato electrónico de la actividad, y facilitado por una universidad privada de Lima Metropolitana. Para el procedimiento de obtención de respuestas de los participantes en el FVC se tuvo dos partes; en la primera se presentó la lista de ítems, y se solicitó al sujeto que calificara cada ítem respecto a su claridad o grado en que son entendibles; la escala de respuesta fue de siete puntos *(desde nada claro hasta completamente claro).* En la segunda y siguiente parte se solicitó que el sujeto evaluara la validez sustantiva presentando nuevamente la lista de ítems, esta vez con las definiciones conceptuales del constructo de interés (carga de

trabajo) y de otros tres constructos que permitieran valorar su discriminación; esos últimos fueron elegidos particularmente para maximizar el ejercicio de contraste de significado del sujeto. Los sujetos tuvieron que asignar el ítem en su constructo, de acuerdo a las instrucciones específicas descritas en los siguientes párrafos.

La definición de carga de trabajo fue: "*Interacción entre el nivel de exigencia de la tarea y el grado de movilización de las capacidades del sujeto que debe realizarse para llevar a cabo la tarea"* (Tovalín & Rodríguez, 2013, p. 99). Sobre los otros constructos, uno de ellos fue conflicto de rol, seleccionado para maximizar el contraste realizado por el sujeto frente al constructo de interés, debido que ambos deberían vincularse dentro de las experiencias negativas del trabajador, se incluyen dentro de los factores psicosociales de demanda, y pueden covariar (Gil-Monte, 2012, 2014, 2016). Se la definió como la experiencia que *"se produce cuando hay demandas, exigencias en el trabajo que son entre sí incongruentes o incompatibles"* (Arquer, Daza, & Nogareda, 1995, p. 3). Otros constructos fueron afrontamiento *("Esfuerzos cognitivos y conductuales constantemente cambiantes, que se desarrollan para manejar las demandas externas y/o internas que son evaluadas como excedentes o desbordantes de los recursos del individuo")* y autoeficacia laboral *("Creencias que poseen los trabajadores para ejecutar exitosamente las actividades asociadas a su profesión")*; estos fueron elegidos para maximizar el contraste frente a constructos con plausible correlación negativa, pues se identifican como aspectos positivos vinculados a los recursos del trabajador para hacer frente a los factores psicosociales de demanda. Estas definiciones se ajustaron las conceptualizaciones propuestas por Lazarus & Folkman (1986, p. 164) y Maffei, Spontón C., Spontón, Castellano, y Medrano (2012, p. 54) respectivamente. Si el sujeto lograba diferenciar la correspondencia de los ítems de carga de trabajo frente a los otros constructos, el respaldo a la validez sustantiva estaría presente.

En el documento presentado a los participantes, en un primer apartado aparecen las definiciones y luego los ítems; la primera definición presentada fue la de carga de trabajo, seguida de conflicto de rol, afrontamiento y autoeficacia laboral. Se indicó a la persona que lea la siguiente instrucción: *"En este apartado, inicialmente leerá las definiciones de cuatro conceptos psicológicos (carga de trabajo, conflicto de rol, afrontamiento y autoeficacia laboral); luego, se le presentarán 6 ítems que deberá analizar y decidir a qué concepto psicológico se encuentra más vinculado empleando la siguiente calificación: Colocará el puntaje de 0 si el ítem no es nada relevante al concepto, 1 si el ítem es Medianamente relevante al concepto y 2 si el ítem es Totalmente relevante al concepto. Recuerde todos los casilleros deben estar completos con alguna calificación y elegirá el puntaje de 2 una sola vez por cada ítem."* Esta instrucción es diferente al propuesto por Anderson y Gerbing (1991) y los estudios precedentes, pero se consideró más apropiado para representar la continuidad del juicio del examinado, y evitar decisiones categóricas (Sí/No) derivadas de la falta de experiencia de los participantes en este tipo de tareas. También, esta propuesta es una modificación del método de Hinkin y Tracey (1999), en que ellos solicitaron que cada constructo sea calificado ordinalmente.

Por otra parte, el análisis de los datos consistió en evaluar dos aspectos; el primero fue la claridad de los ítems, y segundo la validez sustantiva. Para evaluar la claridad de los ítems, se utilizó el coeficiente V (Aiken, 1980, 1985), el cual re-escala la calificación promedio de los participantes hacia un número que varía entre cero y uno; los coeficientes V que se acerquen a 1, indican el grado de la validez de contenido; en este trabajo, este coeficiente cuantificó el grado de claridad. Se decidió que los coeficientes V deberían superar el valor .60 para que un ítem sea evaluado como suficientemente claro, y de esta manera balancear el efecto de la relativa pequeña muestra y reducir el falso el rechazo los ítems que no cumplían con el criterio (i.e., error tipo I). Con este mismo argumento, utilizamos el nivel de confianza en el 90%. Para determinar la significancia poblacional de este coeficiente, se construyeron intervalos de confianza asimétricos (Penfield & Giacobbi, 2004), mediante el programa ICAiken (Merino-Soto & Livia, 2009). Por otra parte, la validez sustantiva fue cuantificada por dos elementos (Anderson & Gerbing, 1991): la proporción de sujetos que asignaron el ítem al constructo principal (*proporción de acuerdo sustantivo,* p_{as}) y el grado en que el ítem se identifica con su constructo comparado con los otros constructos no relevantes *(coeficiente de validez sustantiva,* c_{vs}).

Los puntos de corte para definir la validez usando p_{as} y c_{vs} de cada ítem fue .25, considerando la distribución aleatoria posible en relación al número de constructos comparados, es decir, cuatro (Hoehle, Aljafari, & Venkatesh, 2016; Yao, Wu, & Yang, 2007). Se esperó que el p_{as} fuera sustancialmente grande y diferente en el constructo principal, y cerca de cero en los constructos secundarios. Como una medida adicional, en esta parte de la evaluación de la validez sustantiva, también se utilizó el coeficiente V para cada ítem en cada constructo; se esperó que la magnitud de este coeficiente sería sustancialmente mayor en el constructo de interés, comparado con los otros constructos. En los cálculos con el coeficiente V, también se obtuvo el V total, como una medida sumaria del conjunto de ítem evaluado como similarmente se hace con otras medidas cuantitativas de validez de contenido (Polit & Beck, 2006; Polit, Beck, & Owen, 2007).

Aspectos Éticos

Los procedimientos de la presente investigación se aplicaron conforme a la Declaración de Helsinki.

Resultados

Claridad de los Ítems

En la Tabla 1 se presentan los estadísticos descriptivos obtenidos del grado de claridad de los ítems, observándose que la mayoría de las calificaciones se direccionan hacia las respuestas extremas en la calificación más alta posible; la dispersión es relativamente homogénea.

Tabla 1.
V de Aiken con intervalos de confianza para la claridad de los ítems de carga de trabajo

Ítem	*M*	*DE*	*V*	*IC* 90%
Situaciones duras	5.03	1.326	.67	.61, .72
Hacer varias cosas	6.27	.907	.88	.83, .91
Complicarse trabajo	5.57	1.478	.76	.71, .81
Relajado	6.13	1.074	.85	.81, .89
Tiempo Suficiente	6.23	1.194	.87	.82, .91
Trabajo difícil	5.17	1.859	.69	.63, .74

Nota. IOV: Índice de variación ordinal, *V: V* de Aiken, *IC* 90%: Intervalos de confianza para la *V* de Aiken.

Respecto al coeficiente de V, las estimaciones muestrales superan el valor .66; y de acuerdo a los intervalos de confianza, las estimaciones para población superaron el criterio de .60, por lo tanto, todos los coeficientes son estadísticamente significativos respecto a tal criterio. Comparativamente, el ítem 1 y 6 obtuvieron los coeficientes más bajos, pero aun así superaron el criterio elegido.

Validez Sustantiva

Los resultados sobre p_{as} y c_{vs} (Tabla 2) muestran las fuertes discrepancias entre los valores de los ítems en carga de trabajo y en el resto de constructos secundarios. Los p_{as} en el constructo de interés (i.e., Carga de Trabajo) se caracterizaron por superar el nivel mínimo de acuerdo aleatorio (.25), mientras que los constructos secundarios estuvieron por debajo de este valor. Dos ítems (3 y 6) observaron relación con otros constructos (Afrontamiento y Autoeficacia Laboral, respectivamente), pero estos aun fueron bajos comparados con el pas del constructo principal. Esto también se reflejó en c_{vs}, pues en estos dos ítems fueron también bajos comparados con el resto de los ítems. En resumen, excepto los dos ítems (3 y 6), todos superaron el criterio de aleatoriedad (.25), pero no fueron muy altos como los ejemplos desarrollados en Anderson y Gerbing (1991). Por otro lado, la baja validez sustantiva hallada en el ítem 6 parece replicar lo reportado estudios de estructura interna en España (Gil-Monte, 2016) y Perú (Calderón-De la Cruz et al., 2018), respecto a su comparativa baja carga factorial.

Mediante los coeficientes V (Tabla 3) se detectaron varios resultados que fortalecieron la validez de los ítems de carga de trabajo. Primero, que los coeficientes V para este constructo fueron elevados (> .75) para cada ítem, y fueron estadísticamente significativos respecto al criterio .60; para el puntaje total, también fue elevado ($V_{\bar{X}}$ = .80).

En el resto de constructos secundarios, los coeficientes fueron definitivamente bajos, y ninguno fue estadísticamente significativo respecto al criterio .60. El coeficiente V promedio para conflicto de rol ($V_{\bar{X}}$ = .24), afrontamiento ($V_{\bar{X}}$ = .80).

En el resto de constructos secundarios, los coeficientes fueron definitivamente bajos, y ninguno fue estadísticamente significativo respecto al criterio .60. El coeficiente V promedio

para conflicto de rol (= .24), afrontamiento (= .41) y autoeficacia laboral (= .36) también estuvieron en la línea de la divergencia de los ítems con las definiciones de estos constructos.

Tabla 2.
Proporción de acuerdo sustantivo (p_{as}) y coeficiente de validez sustantiva (c_{vs}) para los ítems de carga de trabajo

	p_{as}				c_{vs}
Ítems	Carga de Trabajo	Conflicto de rol	Afrontamiento	Autoeficacia laboral	
1. Situaciones duras	.63	.03	.23	.10	.40
2. Hacer varias cosas	.60	.23	.13	.07	.40
3. Complicarse trabajo	.63	.03	.30	.00	.37
4. Relajado	.70	.03	.17	.07	.50
5. Tiempo Suficiente	.80	.00	.17	.03	.60
6. Trabajo difícil	.60	.00	.13	.30	.27

Nota. p_{as}: proporción de acuerdo sustantivo. c_{vs}: coeficiente de validez sustantiva

Tabla 3.
Coeficientes V aplicados a las calificaciones de validez sustantiva

Ítems	Carga de trabajo			Conflicto de rol			Afrontamiento			Autoeficacia Laboral		
	M	Coeficiente *V*		*M*	Coeficiente *V*		*M*	Coeficiente *V*		*M*	Coeficiente *V*	
		V	*IC*90%		*V*	*IC*90%		*V*	*IC*90%		*V*	*IC*90%
1. Situaciones duras	1.53	.76	.66, .84	.37	.18	.12, .28	.87	.43	.33, .54	.87	.43	.33, .54
2. Hacer varias cosas	1.6	.80	.70, .87	.97	.48	.38, .59	.67	.33	.24, .44	.63	.31	.22, .42
3. Complicarse trabajo	1.63	.81	.73, .88	.37	.18	.12, .28	1.13	.56	.46, .66	.53	.26	.18, .38
4. Relajado	1.7	.85	.76, .91	.33	.16	.10, .26	.87	.43	.33, .54	.73	.36	.27, .47
5. Tiempo Suficiente	1.73	.86	.78, .92	.27	.13	.08, .22	.87	.40	.33, .54	.63	.31	.22, .42
6. Trabajo difícil	1.53	.76	.66, .84	.5	.25	.17, .35	.7	.35	.26, .45	1.00	.51	.41, .61

Notas. V: V de Aiken, IC90%: Intervalos de confianza para la V de Aiken. *M*: media

Discusión

El presente estudio tuvo por objetivo evaluar el grado de representatividad de los ítems de una escala de carga de trabajo mediante la valoración de participantes en su carácter de jueces no expertos por medio del enfoque de la validez sustantiva. Los resultados fueron satisfactorios, evidenciando un alto grado de acuerdo entre los jueces no expertos (participantes) respecto al contenido y sustancialidad del constructo. En primer lugar, los ítems mostraron ser adecuadamente claros para la muestra examinada; esto es un avance en la adaptación de los ítems en la nueva muestra del presente estudio, dado que permite generalizar el contenido desde su grado de claridad. En este sentido, sus contenidos no deberían ser problemáticos para los futuros examinados en Perú que dominen el habla hispana, respecto a la legibilidad. Al realizar esta evaluación directamente en participantes y no en jueces especialistas o expertos, se tiene garantía

de que sus juicios pueden ser más justificados, ya que son una fuente directa para proveer de información relevante al grado de comprensión de los ítems. La importancia de incluir a los mismos participantes ha sido enfatizada en estudios anteriores (Beaton et al., 2000; Downing, 2005; Fiske, 1967; Merino-Soto, 2016; Morse, D. T. & Morse, 2002), dado que logran representar una válida fuente de expresión de evaluaciones en torno a experiencias concretas del constructo objetivo, y a las características semánticas y culturales de los ítems y el instrumento. Se puede argumentar que, dado que el fraseo de los ítems no parece dependiente de alguna expresión local, entonces los presentes resultados no solo podrían ser generalizables al sector de profesionales participantes de donde proviene la muestra del presente estudio, sino también al trabajador promedio peruano. Sin embargo, replicar este estudio, aumentar tamaño muestral, y obtener mejor representatividad serán condiciones para asegurar este argumento.

Los resultados de la claridad deben ser evaluados en el contexto estadístico, pues lo obtenido no fue en términos absolutos (ítem claro vs ítem no claro), sino más bien alienados a su magnitud. En este marco, los ítems no fueron percibidos igualmente claros; por ejemplo, los ítems 1 y 6 mostraron coeficientes V entre debajo de 70, y el límite inferior de sus intervalos de confianza asimétricos estuvo cerca de .60. Esto converge con la dispersión de las calificaciones, en que también mostraron la variabilidad más larga comparada con el resto de los ítems, e indicando que algunos examinados reportaron bajas calificaciones de claridad en estos ítems. En estos ítems, se podría hallar sujetos que no los perciben tan claros, y por lo tanto requerirán aclaraciones situacionalmente disponibles.

En cuanto a la validez sustantiva y con base en la aplicación del método propuesto, se observaron resultados esperados. Los jueces (i. e., participantes no expertos), luego de la lectura de la definición de los cuatro constructos (carga de trabajo, conflicto de rol, afrontamiento y autoeficacia laboral) y posterior análisis de cada uno de los ítems, seleccionaron a la carga de trabajo como constructo que hacía referencia en los contenidos de los ítems. Si bien la carga de trabajo fue la variable de mayor preferencia por parte de los jueces no expertos en cada uno de los ítems, en menor proporción otros constructos presentaron niveles elevados como el afrontamiento en el ítem 3 (p_{as} = .30; $V_{\bar{X}}$ = .56) y la autoeficacia laboral en el caso del ítem 6 (p_{as} = .30; $V_{\bar{X}}$ = .51). Esto puede ser debido en parte están vinculados a dichos constructos, y esto es esperable siguiendo la propuesta teórica de la validez de la sustantiva en cuanto a los contrastes teóricos que se vinculan a la variable (Anderson & Gerbing, 1991).

En relación al ítem 6 (referido a la carga de trabajo cualitativa), resaltamos que, aunque es sustantivamente asociado a su constructo, con autoeficacia laboral puede emerger un monto sustantividad no planificado en su construcción original (ver Tabla 2 y Tabla 3). En los estudios previos de validación español (Gil-Monte, 2016) y peruano (Calderón-De la Cruz et al., 2018), se observó una relativa baja representatividad de constructo de este ítem. Este resultado puede ser coherente porque es el único ítem que evalúa la carga de trabajo cualitativa. Específicamente, el ítem 6 explora grado de dificultad y el volumen de la tarea a realizar, lo que supone en el trabajador una previa evaluación de sus recursos personales para su afronte, y entre ellos podría estar involucrado la autoeficacia laboral. Sin embargo, esto no es conclusivo y más bien es un señalamiento para plantear una futura hipótesis de validez de constructo.

La aplicación de ambos procedimientos (claridad y sustancialidad) también aportó información útil para tomar decisiones con mejor respaldo sobre la viabilidad de los indicadores como buenos representantes del constructo de carga de trabajo. En este sentido, si los resultados alcanzan elevado consenso, se podría avanzar a una siguiente fase en la validación de los ítems (e. g.: análisis factorial confirmatorio), concluyendo que los ítems elegidos demuestran ser un muestreo de contenido suficiente y adecuado. Por el contrario, resultados discrepantes entre ambos métodos suponen un desafío para el investigador en el rubro del mundo laboral, ya que debe ponderar los resultados para tomar una decisiones eficaces, sobre todo para la evaluación de constructos que generan una afección en las personas como son los factores psicosociales dado que su precisión métrica debe ser correctamente delimitada pues cuando su afección es negativa orienta a una serie de

decisiones no sólo de nivel personal sino en el plano del contexto organizacional.

Precisamente la carga de trabajo desde su perspectiva de demanda es un fenómeno psicosocial universal de gran prevalencia y de gran riesgo a la salud para diversas poblaciones laborales en el ámbito internacional (Schnall, Dobson, & Rosskam, 2009), por ello su efectiva medición psicométrica, a través de la claridad y validez sustantiva resultan una necesidad esencial y emergente. Existen diversos problemas de lenguaje y comprensibilidad de escalas estandarizadas para la medición de este tipo de constructos psicosociales que han sido subestimados, sobre todo cuando se trata de escalas que fueron desarrolladas en otros contextos culturales diferentes al que se aplican, lo que ineludiblemente está conectado con el avance universal en la investigación, la difusión y la prevención de estos problemas (Choi & Juárez-García, 2017), a manera de conclusión, este estudio demuestra que la escala de carga de trabajo de la batería UNIPSICO (Gil Monte, 2016), posee la claridad y validez sustantiva suficiente para recomendar su uso en la investigación y vigilancia psicosocial de la población laboral peruana, y potencialmente, latinoamericana.

Antes de finalizar, declaramos las limitaciones más relevantes: primero, la elección del nivel de confianza (i.e., 90%) y el nivel mínimo de V (i.e., .60) en nuestro estudio, pueden modificar las decisiones de retención o aceptación de la validez sustantiva de los ítems. Sin embargo, en la práctica de investigación, los criterios o sugerencias pueden ser convenciones arbitrarias (Tan & Tan, 2010, p. 276; Wood, 2019, p. 3), incluidos las aplicables al coeficiente V (e.g., Merino-Soto & Livia, 2009), y necesariamente no pueden generalizarse a todas las situaciones de investigación. Segundo, el tamaño muestral utilizado (n = 30) puede parecer insuficientemente; sin embargo, aunque supera el tamaño muestra del estudio germinal (n = 20; Anderson & Gerbing, 1991) y la tendencia metodológica general (n > 10; Anderson & Gerbing, 1991; Aravamudhan & Krishnaveni, 2015; Grant & Davis, 1997), puede ser perfectamente adecuado para situaciones de pre-testing, esto es, de muestra pequeña (Anderson & Gerbing, 1991). Tercero, los trabajadores elegidos en el estudio pueden tener características laborales diferentes otros grupos profesionales, y no se garantiza la generalización de los resultados. En contraste, consideramos que esta aparente limitación no redujo la replicabilidad de la validez de los ítems, porque un mismo ítem fue identificado como problemático en este y otros estudios (e.g., Calderón-De la Cruz et al., 2018; Gil-Monte, 2016). Texto correspondiente a la discusión. Respetar las negritas y cursivas en el original.

Referencias

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 40(2), 955–959. doi: 10.1177/001316448004000419

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(2), 131-142. doi: 10.1177/0013164485451012

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. Journal of Applied Psychology, 76(3), 732-740. doi: 10.1037/0021-9010.76.5.732

Aravamudhan, N. R., & Krishnaveni, R. (2015). Establishing and reporting content validity evidence of new Training and Development Capacity Building Scale (TDCBS). Management, 20(1), 131–58

Arquer, I., Daza, F., & Nogareda., C. (1995). NTP 338: Ambigüedad y conflicto de rol. Barcelona: INSHT.

Bacharach, S. B., Bamberger, P., & Conley, S. C. (1990). Work processes, role conflict, and role overload: The case of nurses and engineers in the public sector. Work and Occupations, 17(2), 199-228. doi: 10.1177/0730888490017002004

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. Spine, 25(24), 3186-3191. doi: 10.1097/00007632-200012150-00014

Byrne, B. (1994). Burnout: testing for the validity replication and invariance of causal structure across elementary, intermediate and secondary teachers. American Educational Research Journal, 31(3), 654-673. doi: 10.3102/00028312031003645

Calderón-De la Cruz, G., Merino-Soto, C., Juárez-García, A., & Jimenez-Clavijo, M. (2018). Validación de la escala de carga de trabajo en trabajadores peruanos. Archivos de Prevención de Riesgos Laborales, 21(3), 123-127. doi: 10.12961/aprl.2018.21.03.2

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. Organizational Research Methods 4(1), 62-83. doi: 10.1177/109442810141004

Choi, B., & Juárez-García, A. (2017). Language issues in standard questionnaires for assessing psychosocial working conditions: the case of the JCQ and the ERIQ. En S. Cassilde, & A. Gilson (Eds.), Psychosocial Health at Work and Language. International Perspectives toward their Categorizations at Work (pp. 3-18). Cham, Switzerland: Springer. doi: 10.1007/978-3-319-50545-9_1

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. Medical Care, 45(5), 12-21. doi: 10.1097/01.mlr.0000254567.79743.e2

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Advances in Health Sciences Education: Theory and Practice, 10(2), 133-143. doi: 10.1007/s10459-004-4019-5

Farrel, A. M., Souchon, A. L., & Durden, G. R. (diciembre, 2003). The service leadership scale: a substantive validity test. Paper presentado en the Australian and New Zealand Marketing Academy Conference (ANZMAC), Melbourne, Australia.

Fiske, D. W. (1967). The subjects react to test. American Psychologist, 22(2), 287-296. doi: 10.1037/h0024523

Frone, M. R., Yardley, J. K., & Markel, K. (1997). Developing and testing an integrative model of the work-family interface. Journal of Vocational Behavior, 50(2), 145-167. doi: 10.1006/jvbe.1996.1577

Geurts, S. A. E., & Demerouti, E. (2003). Work/non-work interface: A review of theories and findings. En M. J. Schabracq, J. A. M. Winnubst, & C. L. Cooper (Eds.), The handbook of work and health psychology (pp. 279-312). Chichester, UK: John Wiley & Sons.

Gil-Monte, P. (2012). Riesgos psicosociales en el trabajo y salud ocupacional. Revista Peruana de Medicina Experimental y Salud Pública, 29(2), 237-241. doi: 10.1590/S1726-46342012000200012

Gil-Monte, P. R. (2014). Manual de psicosociología aplicada al trabajo y a la prevención de los riesgos laborales. Madrid: Pirámide.

Gil-Monte, P. R. (2016). La Batería UNIPSICO: propiedades psicométricas de las escalas que evalúan los factores psicosociales de demanda. Archivos de Prevención de Riesgos Laborales, 19(2), 86-94. doi: 10.12961/aprl.2016.19.02.2

Gil-Monte, P. R., & Peiró, J. M. (1997). Desgaste psíquico en el trabajo: el síndrome de quemarse. Madrid: Síntesis.

Gil-Monte, P., & García-Juesas, J. (2008). Efectos de la sobrecarga laboral y la autoeficacia sobre el síndrome de quemarse por el trabajo (burnout). Un estudio longitudinal en enfermería. Revista Mexicana de Psicología, 25(2), 329-337.

Gil-Monte, P., García-Juesas, J., & Caro, M. (2008). Influencia de la sobrecarga laboral y la autoeficacia sobre el síndrome de quemarse por el trabajo (burnout) en profesionales de enfermería. Revista Interamericana de Psicología, 42(2), 113-118.

Gil-Monte, P., López-Vílchez, J., Llorca-Rubio, J., & Sánchez, J. (2016). Prevalencia de riesgos psicosociales en personal de la administración de justicia de la comunidad valenciana (España). Liberabit, 22(1), 7-19.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. Research in Nursing and Health, 20(3), 269–274. doi: 10.1002/(SICI)1098-240X(199706)20:3<269::AID-NUR9>3.0.CO;2-G

Greenglass, E. R., Burke, R. J., & Moore, K. A. (2003). Reactions to increased workload: effects on professional efficacy of nurses. Applied Psychology: An International Review, 52(2), 580-597. doi: 10.1111/1464-0597.00152

Hanley, K., Howard, M. C., Zhong, B., Soto, J. A., Pérez, C. R., Lee, E. A., … Minnick, M. R. (2015). The communication anxiety regulation scale: development and initial validation. Communication Quarterly, 63(1), 23-43. doi: 10.1080/01463373.2014.965836

Helm, S., Eggert, A., & Garnefeld, I. (2010). Modeling the impact of corporate reputation on customer satisfaction and loyalty using partial least squares. En V. Esposito, W. W. Chin, J. Henseler, & H. Wang (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications (pp. 515-534). Berlin: Springer.

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. Organizational Research Methods, 2(2), 175-186. doi: 10.1177/109442819922004

Hoehle, H., Aljafari, R., & Venkatesh, V. (2016). Leveraging Microsoft's mobile usability guidelines: Conceptualizing and developing scales for mobile application usability. International Human-Computer Studies, 89(2), 35–53. doi: 10.1016/j.ijhcs.2016.02.001

Juárez-García, A. (2007). Factores psicosociales laborales relacionados con la tensión arterial y síntomas cardiovasculares en personal de enfermería en México. Salud Pública de México, 49(2), 109-117.

Kanai, A., & Wakabayashi, M. (2001). Workaholism among Japanese blue-collar employees. International Journal of Stress Management, 8(2), 129-145. doi: 10.1023/A:1009529314121

Kirch, W. (2008). Encyclopedia of Public Health. New York: Springer.

Lasch, K. E., Marquis, P., Vigneux, M., Abetz, L., Arnould, B., Bayliss, M., … Rosa, K. (2010). PRO development: rigorous qualitative research as the crucial foundation. Quality of Life Research, 19(8), 1087-1096. doi: 10.1007/s11136-010-9677-6

Lazarus, R. S., & Folkman, S. (1986). Estrés y procesos cognitivos. Barcelona: Martínez Roca.

Loevinger, J. (1957). Objective test as instruments of psychological theory. Psychology Reports, 3(2), 635-694. doi: 10.2466/pr0.1957.3.3.635

Maffei, L., Spontón, C., Spontón, M., Castellano, E., & Medrano, L. A. (2012). Adaptación del Cuestionario de Autoeficacia Profesional (AU-10) a la población de trabajadores cordobeses. Pensamiento Psicológico, 10(1), 51-62.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., … Cella, D. (2012). Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Quality of Life Research, 21(5), 739-746. doi: 10.1007/s11136-011-9990-8

Manson, S. (1997). Cross-cultural and multi-ethnic assessment of trauma. En J. P. Wilson, & T. M. Keane (Eds.), Assessing psychological trauma and PTSD: A handbook for practitioners (pp. 239-266). New York: Guilford.

May, L. A., & Warren, S. (2001). Measuring quality of life of persons with spinal cord injury: Substantive and structural validation. Quality of Life Research, 10(2), 503-515. doi: 10.1023/A:1013027520429

Meliá, J. L., Nogareda, C., Lahera, M., Duro, A., Peiró, J. M., Salanova, M., & Gracia, D. (2006). Principios comunes para la evaluación de riesgos psicosociales en la empresa. En J. L. Meliá, C. Nogareda, M. Lahera, A. Duro, J. M. Peiró, R. Pou, … F. Martínez-Losa (Eds.), Perspectivas de Intervención en Riesgos Psicosociales. Evaluación de Riesgos (pp. 13-36). Barcelona: Foment del Treball Nacional.

Merino-Soto, C. (2016). Percepción de la claridad de los ítems: Comparación del juicio de estudiantes y jueces-expertos. Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 14(2), 1469-1477. doi: 10.11600/1692715x.14239120615

Merino-Soto, C., & Livia, C. (2009). Intervalos de confianza asimétricos para el índice la validez de contenido: Un programa Visual Basic para la V de Aiken. Anales en Psicología, 25(1), 169-171.

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

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., ... De Vet, H. C. (2006). Protocol of the COSMIN study: Consensus-based Standards for the selection of health Measurement Instruments. BMC Medical Research Methodology, 6(2), 2-10. doi: 10.1186/1471-2288-6-2

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., ... De Vet, H. C. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Medical Research Methodology, 10(2), 22. doi: 10.1186/1471-2288-10-22

Mora-Ríos, J., Bautista-Aguilar, N., Natera, G., & Duncan, P. (2013). Adaptación cultural de instrumentos de medida sobre estigma y enfermedad mental en la Ciudad de México. Salud Mental, 36(1), 9-18. doi: 10.17711/SM.0185-3325.2013.002

Morse, D. T., & Morse, L. W. (2002). Are undergraduate examiner's perceptions of item difficulty related to item characteristics? Perceptual and Motor Skills, 95(3-2), 1281-1286. doi: 10.2466/pms.2002.95.3f.1281

Nevo, B. (1995). Examinee Feedback Questionnaire: Reliability and validity measures. Educational and Psychological Measurement, 55(3), 499-504. doi: 10.1177/0013164495055003017

Organización Internacional del Trabajo. (2016). Estrés en el Trabajo: un reto colectivo. (Informe No. 1). Ginebra: Organización Internacional del Trabajo.

Patlán, J. (2013). Efecto del burnout y la sobrecarga en la calidad de vida en el trabajo. Estudios Gerenciales, 29(129), 445-455. doi: 10.1016/j.estger.2013.11.010

Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011). Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2--assessing respondent understanding. Value Health, 14(8), 978-988. doi: 10.1016/j.jval.2011.06.013

Penfield, R. D., & Giacobbi, P. R. Jr. (2004) Applying a score confidence interval to Aiken's item content-relevance index. Measurement in Physical Education and Exercise Science, 8(4), 213-225. doi: 10.1207/s15327841mpee0804_3

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(2), 489-497. doi: 10.1002/nur.20147

Polit, D. F., Beck, C. T., & Owen, S. (2007). Is the CVI an acceptable indicator of content validity? Research in Nursing & Health, 30(2), 459-467. doi: 10.1002/nur.20199

Rothman, M., Burke, L., Erickson, P., Leidy, N. K., Patrick, D. L., & Petrie, C. D. (2009). Use of

existing patient-reported outcome (PRO) instruments and their modification: the ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO Task Force Report. Value Health, 12(8), 1075-1083. doi: 10.1111/j.1524-4733.2009.00603.x

Schnall, P. L., Dobson, M., & Rosskam, E. (2009). Unhealthy work: Causes, consequences, cures. London: Baywood Publishing Company.

Shepard, K. F., Jensen, G. M., Schmoll, B. J., Hack, L. M., & Gwyer, J. (1993). Alternative approaches to research in physical therapy: Positivism and phenomenology. Physical Therapy, 73(2), 88-97. doi: 10.1093/ptj/73.2.88. PMID: 8421722.

Skinner, N., & Pocock, B. (2008). Work-life conflict: is work time or work overload more important? Asia Pacific Journal of Human Resources, 43(3), 303-315. doi: 10.1177/1038411108095761

Tan, S. H., & Tan, S. B. (2010). The correct interpretation of confidence intervals. Proceedings of Singapore Healthcare, 19(3), 276–278. doi: 10.1177/201010581001900316

Tovalín, H., & Rodríguez, M. (2013). Conceptos básicos en la evaluación del riesgo psicosocial en los centros de trabajo. En A. J. García, & A. Camacho Ávila (Eds.), Reflexiones teórico-conceptuales de lo psicosocial en el trabajo (pp. 95-112). Cuernavaca, México: Universidad Autónoma del Estado de Morelos, Juan Pablos Editor.

Unda, S., Uribe, F., Jurado, S., García, M., Tovalín, H., & Juárez, A. (2016). Elaboración de una escala para valorar los factores de riesgo psicosocial en el trabajo de profesores universitarios. Journal of Work and Organizational Psychology, 32(2), 67-74. doi: 10.1016/j.rpto.2016.04.004

Urrutia, M., Barrios, S., Gutiérrez, M., & Mayorga, M. (2014). Métodos óptimos para determinar la validez de contenido. Educación Médica Superior, 28(3), 547-558.

van Kooten, J. A. M. C., Terwee, C. B., Kaspers, G. J. L., Raphaële, R. L., & van Litsenburg, R. R. L. (2016). Content validity of the Patient-Reported Outcomes Measurement Information System Sleep Disturbance and Sleep Related Impairment item banks in adolescents. Health and Quality of Life Outcomes, 14(2), 92-101. doi: 10.1186/s12955-016-0496-5

Veloutsou, C. A., & Panigyrakis, G. G. (2004). Consumer brands managers´ job stress, job satisfaction, perceived performance and intention to leave. Journal of Marketing Management, 20(2), 105-131. doi: 10.1362/026725704773041140

Wood, M. (2019). Simple methods for estimating confidence levels, or tentative probabilities, for hypotheses instead of p values. Methodological Innovations, 1(2), 1-9. doi: 0.1177/2059799119826518

Yao, G., Wu, C. H., & Yang, C. T. (2007). Examining the content validity of the WHOQOL- BREF from respondents' perspective by quantitative methods. Social Indicator Research, 85(3), 483-498. doi: 10.1007/s11205-007-9112-8

Zabkar, V. (2000). Some methodological issues with structural equation modeling application in relationship quality context. New Approaches in Applied Statistics, 16(2), 211-224.