Referencia para citar este artículo: 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), pp. 1469-1477.

Percepción de la claridad de los ítems: Comparación del juicio de estudiantes y jueces-expertos*

César Merino-Soto**
Investigador Universidad de San Martín de Porres, Perú.

Artículo recibido en junio 12 de 2015; artículo aceptado en julio 14 de 2015 (eds.)

• ***Resumen (analítico):*** *El presente reporte de investigación preliminar tiene por objetivo evaluar la percepción de la dificultad de los ítems de un cuestionario de personalidad, en dos grupos de evaluadores: estudiantes universitarios y profesores (jueces expertos). La muestra fueron dos grupos: estudiantes y profesores identificados como jueces expertos. Independientemente, para ambos grupos se administró un formulario de opinión sobre la dificultad de lectura o claridad de los ítems del EPQR (Eysenk Personality Questionnaire-Revised). Se hallaron diferencias en los porcentajes de ítems identificados como poco claros, y el acuerdo entre ambos grupos fue mayormente bajo. Estos resultados ponen en relevancia la inclusión del examinado en la evaluación de la dificultad de los ítems usando formularios estandarizados.*

Palabras clave: Validez, pruebas, metodología, acuerdo, Perú (Tesauro de Ciencias Sociales de la Unesco).

Perception of item clarity: Comparison of judgments between students and expert-judges

• ***Abstract (analytic):*** *The aim of this preliminary study is to assess the perceived difficulty of the items of a personality questionnaire in two groups of participants: university students and teachers (expert-judges). As part of content validity, it examines the degree of readability of the items, usually evaluated by expert judges; however, contributions from the potential examinees have not been previously examined. The sample consisted of two groups: students and teachers identified as expert judges. Independently, both groups were asked their opinions on the readability or clarity of the items in an EPQR test (Eysenck Personality Questionnaire-Revised). The results demonstrated a significant difference in the percentage of items identified as unreadable and agreement between the two groups was mostly low. These results point to the inclusion of examinees in assessing the readability of items contained in standardized forms, and the potential increase in the construction of irrelevant variance if the process does not include the active participation of those filling out the form or taking the test.*

Key words: Validity, tests, methodology, agreement, Peru (Unesco Social Sciences Thesaurus).

---

* Este artículo corto (área de ciencias sociales, subárea psicología) se ha realizado durante el periodo de enero del 2012 a enero de 2014, como parte de la línea de investigación psicométrica del Instituto de Investigación de Psicología de la Universidad de San Martín de Porres (Lima, Perú). El nombre de la investigación es: "Metodologías de validez de contenido", avalada por el Instituto de Investigación de Psicología, Universidad de San Martín de Porres, sin referencia numérica.

** Psicólogo, Master en Psicología educativa, investigador en el Instituto de Investigación de Psicología de la Universidad de San Martín de Porres. Dirección de correspondencia: Av. Tomás Marsano 242 (5to piso)-Lima 34, Perú. Correos electrónicos: cmerinos@usmp.pe; sikayax@yahoo.com.ar

Percepção de clareza de itens: Comparação do julgamento de estudantes e juízes-especialistas

• ***Resumo (analítico):*** *Esta pesquisa preliminar consiste em avaliar a percepção da dificuldade dos itens de um questionário de personalidade, em dois grupos de avaliadores: estudantes universitários e professores (juízes especialistas tematicamente). A amostragem é composta de dois grupos: estudantes e professores identificados como juízes especialistas. Independentemente disso, para ambos os grupos foi aplicado um formulário de avaliação sobre a dificuldade de leitura e clareza dos itens de EPQR (Eysenk Personality Questionnaire-Revised). Foram encontradas diferenças nos percentuais de itens identificados como incertos, e o acordo entre os dois grupos foi baixo, em sua maioria. Estes resultados colocam em relevância a inclusão da pessoa examinada na avaliação da dificuldade de compreensão dos itens utilizados em formulários padronizados.*

Palavras-chave: Validade, teste, metodologia, acordo, Peru (Thesaurus de Ciências Sociais da Unesco).

-1. Introducción. -2. Método. -3. Resultados. -4. Discusión. -Lista de referencias.

1. Introducción

Desde hace años, la participación del examinado como fuente de valoración e impacto en las características psicométricas de los instrumentos ha sido establecida como una fuente importante e interactiva con los resultados de los análisis psicométricos subsecuentes (Fiske, 1967, Morse & Morse, 2002, Nevo, 1995, Schinka & Borum, 1994, Secolsky, 1983), pero parece poco frecuente que sea incluido en los planes de validación como un agente proactivo, menos pasivo y discerniente de las características del procedimiento de medición de atributos psicológicos. Un área en que los examinados han hecho valoraciones sobre las pruebas que les administran ha sido sobre sus actitudes hacia las pruebas de rendimiento (McCarthy & Goffin, 2003), pero podría requerirse su participación respecto a sus percepciones sobre algunos aspectos relevantes de la validez del instrumento (Nevo, 1995). En este marco, el examinado puede usar su experiencia acumulativa y directa sobre el material leído, y presumiblemente relacionada con el constructo. Dentro de las propiedades de los ítems evaluados durante el proceso de construcción de instrumentos, la claridad de los mismos se refiere al grado en que el ítem puede ser entendido por el examinado, asumiendo un nivel constante de comprensión lectora. Este aspecto puede estar involucrado en la validez *aparente* o *de facie* (DeVon et al. 2007, Haladyna, 2004), un aspecto que puede ser útil para tomar decisiones sobre la construcción inicial de los ítems, y que ha sido abordado en algunas medidas de personalidad (McCrae, Costa & Martin, 2005).

Efectivamente, la percepción de los examinados sobre la dificultad de los ítems interacciona con algunos aspectos de la validez de los ítems (Morse & Morse, 2002), y debería ser considerada como una fuente activa para tomar decisiones en el proceso de adaptación o construcción de pruebas. Esta percepción de los examinados pueden ser identificados como juicios sobre la utilidad de los ítems o de sus propiedades aparentes, pues señalarían potenciales problemas en los ítems (Downing, 2005) y una amenaza a la validez de constructo (Downing, 2002). Esto puede considerarse como parte de la dificultad intrínseca de una tarea, que exige al evaluado un monto de carga cognitiva para realizarlo (Sweller, 2010); este aspecto podría deducirse de un análisis lógico de las demandas cognitivas de la tarea, pero también de la propia percepción de examinado (Secolsky, 1983, Secolsky, Kossar, Magaram & Fuentes, 2011) que potencialmente representa la muestra hacia la cual se dirige el instrumento. Aunque parece más razonable que los investigadores con experiencia y conocimientos en construcción de pruebas o en el constructo de interés, sean elegidos en el proceso de identificar el nivel de

Rev.latinoam.cienc.soc.niñez juv 14 (2): 1469-1477, 2016
http://revistalatinoamericanaumanizales.cinde.org.co
*DOI:*10.11600/1692715x.14239120615

claridad de los ítems, ellos únicamente pueden hacer estimaciones sobre la dificultad que los examinados encontrarían al responder los ítems. En esta situación, nos podemos preguntar si sería similarmente efectivo indagar sobre la claridad de los ítems en los mismos sujetos de estudio; por lo tanto, el cuestionamiento que plantea el presente manuscrito es ¿por qué no indagar en los mismos examinados la percepción de la legibilidad o claridad de los ítems, indicadores que finalmente se utilizarán en ellos?

El propósito de este breve reporte es explorar los efectos en la percepción de claridad de los ítems de un cuestionario de personalidad, comparando los juicios de estudiantes y expertos en evaluación psicológica. El enfoque exploratorio de los resultados permitiría descubrir si la evaluación de la claridad de los ítems puede ser diferente entre ambos grupos, y lo cual llevaría a sugerir que la valoración de la claridad de los elementos que componen un instrumento contiene un monto de variabilidad dependiente de los grupos evaluados. Si esta variabilidad se confirma, respecto a los jueces expertos y propios examinados, se debería sugerir que la inteligibilidad de los ítems debe ser evaluada por diferentes grupos y no únicamente por el grupo de expertos.

2. Método

Participantes

Los participantes fueron dos grupos de valoración o "jueces", estudiantes y profesores. En los estudiantes, el contexto del estudio fue la escuela de psicología de una universidad pública y otra privada, ambos en Lima Metropolitana. Fueron 36 participantes de estudios regulares, mayoritariamente mujeres (n=23, 63%), todos entre 18 y 24 años de edad (promedio 21.3 años), cursando entre el tercero y sexto ciclo de estudios. Respecto a los profesores elegidos que sirvieron como "jueces expertos", fueron 6 de cursos de pregrado y 1 de postgrado de la universidad privada (edad promedio 39 años), que enseñaban por lo menos 2 cursos sobre pruebas psicológicas; tres de ellos con maestría y uno con doctorado; algunos con publicaciones y con experiencia en construcción de pruebas; además, declararon tener experiencia en investigación y en asesoría de tesis.

Instrumento

Cuestionario de Personalidad de Eysenck, versión revisada (*EPQR*; Eysenck & Eysenck, 2001). Este instrumento evalúa la personalidad en tres dimensiones básicas (Extraversión, Neuroticismo, Psicoticismo), y es ampliamente usado en adultos. Son 83 ítems respondidos como Sí o No. Hay una extensa investigación de validez y confiabilidad internacional sobre esta versión revisada.

Procedimiento

Los estudiantes llenaron el EQPR en sus aulas de clase o pequeños grupos, aplicándose las instrucciones estandarizadas del instrumento. Inmediatamente luego de responder al instrumento, se les dio un formulario de opinión sobre el grado en que los ítems fueron comprensibles y claros durante la experiencia de llenado. Cada ítem fue valorado usando una escala del 1 (*Nada entendible*) al 5 (*Completamente entendible: Ninguna duda para entender el ítem*). Por otro lado, los profesores-jueces llenaron independientemente el mismo formulario, pero con las instrucciones modificadas para valorar el grado de dificultad o claridad que creían que los examinados encontrarían en cada ítem; en estas instrucciones para los profesores-jueces, se identificó que los potenciales examinados serían adultos egresados, universitarios y personas de la comunidad en general. Ambos grupos primero llenaron el consentimiento informado.

El análisis aplicó el método descrito por Merino-Soto y Livia (2009) para representar la estimación la validez de cada ítem respecto a su claridad, con el cual se obtiene una estimación puntual y por intervalos de confianza del coeficiente V de validez de contenido (Aiken, 1980, 1985). Estos intervalos de confianza se basaron en la metodología de Penfield y

Gioccobi (2004), que usaron la técnica score (Wilson, 1927) para desarrollar intervalos de confianza asimétricos para variables binomiales. Se usará el nivel de confianza del 95%. Respecto al coeficiente V, se examinará la dirección de las diferencias (resta simple entre V de estudiantes y V de profesores-jueces) y el acuerdo en la detección de ítems que superan o están por debajo del criterio de aceptación de la validez.

3. Resultados

Uno de los aspectos que se analizó fue la dirección de las diferencias, es decir, si los profesores o estudiantes tendieron a mostrar coeficientes relativamente elevados (Tablas 1 y 2). En la escala Extroversión, el 68% de los ítems tuvieron coeficientes V que fueron mayores en los profesores; en la escala Sinceridad, la orientación de las diferencias fue del 50%, mientras que en Psicoticismo y Neuroticismo, el porcentaje de coeficientes mayores en profesores fue 52% y 41%, respectivamente. Considerando todos los ítems, el 51% tuvo coeficientes V mayores en los profesores respecto a los estudiantes. Esto sugiere que hubo una tendencia similar entre estudiantes y profesores-jueces respecto a la magnitud de los coeficientes V; justamente, la diferencia estandarizada (Cohen, 1992) entre los coeficientes V promedio de estudiantes y profesores-jueces fue trivial: -0.22 (Extroversión), 0.04 (Sinceridad), 0.07 (Psicoticismo) y 0.27 (Neuroticismo). Estos resultados parecen señalar que la percepción de la dificultad de los ítems es similar entre ambos grupos, sin embargo, los siguientes análisis mostraron resultados disímiles.

Para las siguientes comparaciones, se decidió que el criterio mínimo para aceptar un consenso en la claridad del ítem con el coeficiente V fuera 0.75 (nivel mínimo ligeramente superior al propuesto por Merino-Soto & Livia, 2009). Por lo tanto se usó su límite inferior del I.C. 95% para identificar los ítems percibidos como poco claros, de tal modo que si el límite inferior del coeficiente V para un ítem era igual o mayor a 0.75, éste se considerada como válido; en caso contrario, el ítem se identificada como válido. En la Tabla 1 y 2 aparecen las estimaciones de V y sus intervalos de confianza. La proporción de ítems identificados fue diferente entre profesores y estudiantes: Extraversión (0.36 vs 0.10), Mentira (0.44 vs 0.05), Psicoticismo (0.30 vs 0.17) y Neuroticismo (0.52 vs 0.13); la magnitud de las diferencias de estas proporciones, evaluado por el coeficiente *w* (Sheskin, 2007) fue, 0.46, 0.71, 0.21 y 0.65, respectivamente, y sugieren discrepancias que van desde pequeñas a grandes. Por otro lado, el coeficiente *kappa* calculado para examinar el acuerdo obtenido entre ambos grupos fue, respectivamente, 0.33 ($p > 0.05$), 0.13 ($p > 0.05$), .41 ($p < 0.05$) y .24 ($p > 0.05$). El ajuste por sesgo y prevalencia del coeficiente kappa, (*Pabak*: Byrt, Bishop & Carlin, 1993), fue 0.47 (aceptable), 0.22 (pobre), 0.56 (aceptable) y 0.21 (pobre). Los niveles cualitativos se refieren a la propuesta de Cicchetti (2001).

Rev.latinoam.cienc.soc.niñez juv 14 (2): 1469-1477, 2016
http://revistalatinoamericanaumanizales.cinde.org.co
DOI:10.11600/1692715x.14239120615

Tabla 1. *V de Aiken y sus Intervalos de Confianza para Extroversión y Sinceridad*[a]

Extroversion					Sinceridad				
Ítem	Estudiantes		Profesores		Ítem	Estudiantes		Profesores	
	V_{Aiken}	IC 95%	V_{Aiken}	IC 95%		V_{Aiken}	IC 95%	V_{Aiken}	IC 95%
3	0.945	0.895 - 0.972	1.000	0.879 – 1.000	5	0.833	0.763- 0.885	0.928	0.772 – 0.980
6	**0.805**	**0.733- 0.861**	**0.715**	**0.530 – 0.848**	7	0.895	0.834- 0.935	**0.858**	**0.686 – 0.943**
12	**0.653**	**0.572- 0.725**	**0.785**	**0.604 – 0.897**	10	0.938	0.885- 0.967	0.965	0.824 – 0.994
16	0.918	0.861- 0.952	0.965	0.824 – 0.994	11	0.86	**0.794- 0.907**	**0.858**	**0.686 – 0.943**
22	0.93	0.876- 0.961	0.893	0.728 – 0.963	14	0.855	0.788- 0.903	**0.823**	**0.645 – 0.922**
25	0.93	0.876- 0.961	0.965	0.824 – 0.994	21	0.938	0.885- 0.967	0.928	0.772 – 0.980
27	0.945	0.895- 0.972	0.928	0.772 – 0.980	30	**0.785**	**0.711- 0.844**	**0.643**	**0.458 – 0.793**
28	0.965	0.921- 0.985	1.000	0.879 – 1.000	33	0.93	0.876- 0.961	0.965	0.824 – 0.994
31	0.848	0.780- 0.897	**0.893**	**0.728 – 0.963**	36	0.895	0.834- 0.935	**0.858**	**0.686 – 0.943**
39	0.923	0.867- 0.956	**0.858**	**0.686 – 0.943**	38	0.903	0.843- 0.941	**0.858**	**0.686 – 0.943**
46	0.867	0.802- 0.913	1.000	0.879 – 1.000	43	0.895	0.834- 0.935	**0.823**	**0.645 – 0.922**
47	0.918	0.861- 0.952	0.965	0.824 – 0.994	45	0.965	0.921- 0.985	1.000	0.879 – 1.000
49	0.910	0.852- 0.947	0.965	0.824 – 0.994	56	0.895	0.834- 0.935	0.965	0.824 – 0.994
53	0.93	0.876- 0.961	0.965	0.824 – 0.994	60	0.938	0.885- 0.967	**0.823**	**0.645 – 0.922**
57	0.93	0.876- 0.961	0.965	0.824 – 0.994	65	0.93	0.876- 0.961	1.000	0.879 – 1.000
58	0.953	0.904- 0.977	1.000	0.879 – 1.000	68	0.953	0.904- 0.977	1.000	0.879 – 1.000
69	0.828	0.757- 0.880	**0.643**	**0.458 – 0.793**	79	0.923	0.867- 0.956	0.928	0.772 – 0.980
70	0.883	0.820- 0.925	0.965	0.824 – 0.994	82	0.945	0.895- 0.972	1.000	0.879 – 1.000
77	0.86	0.794- 0.907	**0.858**	**0.686 – 0.943**					

[a]: en negrita, los coeficientes (V Aiken) y sus intervalos que están por debajo del criterio V = .75

Calculado para el grupo total de ítems, los estudiantes y profesores identificaron 10 (0.12) y 34 (0.41) ítems respectivamente que no superaban el criterio *V* = .75. Es decir, los profesores detectaron 5 veces más ítems (Odds Ratio = 5.07) que los estudiantes. La diferencia entre estas proporciones fue moderada (*w* = 0.49).

Tabla 2. *V de Aiken y sus Intervalos de Confianza para Psicoticismo y Neuroticismo*[a]

Psicoticismo					Neuroticismo				
Ítem	Estudiantes		Profesores		Ítem	Estudiantes		Profesores	
	V_{Aiken}	IC 95%	V_{Aiken}	IC 95%		V_{Aiken}	IC 95%	V_{Aiken}	IC 95%
1	0.903	0.843-0.941	0.715	0.530 – 0.848	2	0.918	0.861 – 0.952	0.928	0.772 – 0.980
9	0.93	0.876-0.961	0.858	0.686 – 0.943	4	0.883	0.820 – 0.925	**0.715**	**0.530 – 0.848**
15	0.778	**0.703-0.838**	0.715	**0.530 – 0.848**	8	0.875	0.811 – 0.919	**0.858**	**0.686 – 0.943**
17	0.828	0.757-0.880	1.000	0.879 – 1.000	13	**0.73**	**0.652 – 0.796**	**0.678**	**0.492 – 0.820**
23	0.84	0.771-0.891	0.965	0.824 – 0.994	18	0.945	0.895 – 0.972	**0.893**	**0.728 – 0.963**
26	0.89	0.828-0.931	0.678	**0.492 – 0.820**	19	0.855	0.788 – 0.903	**0.823**	**0.645 – 0.922**
29	0.84	0.771-0.891	0.823	**0.645 – 0.922**	20	**0.813**	**0.741 – 0.868**	**0.573**	**0.392 – 0.736**
34	0.93	0.876-0.961	0.858	**0.686 – 0.943**	24	0.89	0.828 – 0.931	**0.893**	**0.728 – 0.963**
37	0.918	0.861-0.952	0.928	0.772 – 0.980	32	0.903	0.843 – 0.941	**0.823**	**0.645 – 0.922**
40	0.958	0.911-0.980	0.928	0.772 – 0.980	35	0.875	0.811 – 0.919	**0.785**	**0.604 – 0.897**
44	0.723	**0.644-0.789**	0.573	**0.392 – 0.736**	41	0.855	0.788 – 0.903	0.928	0.772 – 0.980
48	0.89	0.828-0.931	0.785	**0.604 – 0.897**	42	0.965	0.921 – 0.985	0.965	0.824 – 0.994
50	0.958	0.911-0.980	0.965	0.824 – 0.994	52	0.91	0.852 – 0.947	0.965	0.824 – 0.994
51	0.938	0.885-0.967	0.965	0.824 – 0.994	54	0.86	0.794 – 0.907	**0.823**	**0.645 – 0.922**
55	0.93	0.876-0.961	0.965	0.824 – 0.994	62	0.89	0.828 – 0.931	0.928	0.772 – 0.980
59	0.918	0.861-0.952	0.965	0.824 – 0.994	64	0.895	0.834 – 0.935	0.928	0.772 – 0.980
61	0.93	0.876-0.961	0.928	0.772 – 0.980	72	0.945	0.895 – 0.972	1.000	0.879 – 1.000
63	0.645	**0.564-0.718**	0.643	**0.458 – 0.793**	73	0.953	0.904 – 0.977	1.000	0.879 – 1.000
66	0.93	0.876-0.961	1.000	0.879 – 1.000	75	0.938	0.885 – 0.967	0.858	0.686 – 0.943
67	0.923	0.867-0.956	0.965	0.824 – 0.994	76	0.938	0.885 – 0.967	0.928	0.772 – 0.980
71	0.945	0.895-0.972	0.965	0.824 – 0.994	78	0.903	0.843 – 0.941	**0.785**	**0.604 – 0.897**
74	0.958	0.911-0.980	0.965	0.824 – 0.994	81	**0.583**	**0.501 – 0.660**	**0.428**	**0.264 – 0.608**
80	0.793	**0.719-0.851**	0.965	0.824 – 0.994	83	0.91	0.852 – 0.947	1.000	0.879 – 1.000

[a]*En negrita, los coeficientes (V Aiken) y sus intervalos que están por debajo del criterio V = .75*

Rev.latinoam.cienc.soc.niñez juv 14 (2): 1469-1477, 2016
http://revistalatinoamericanaumanizales.cinde.org.co
*DOI:*10.11600/1692715x.14239120615

4. Discusión

El presente estudio expone la comparación de los juicios de jueces expertos (profesores) y examinados (estudiantes) respecto a la dificultad percibida de los ítems del EPQR. Considerando los índices de acuerdo obtenidos, especialmente los corregidos por acuerdo aleatorio, ambos los grupos coinciden pocas veces en identificar los ítems que consideran difíciles de entender por los potenciales examinados. Esto significa que la discrepancia en percibir el grado de inteligibilidad no puede considerarse equivalente entre jueces expertos y los mismos examinados, pues habrá diferencias en los ítems que identificarán como problemáticos. Esto debe alertar al investigador o constructor de pruebas, sobre la efectividad de la valoración obtenida en un solo grupo de evaluados, la cual es rutinariamente obtenida desde profesionales o investigadores con experiencia temática o experiencia en la interacción con la muestra objetivo.

En el papel de "juez", el evaluador puede expresar percepciones sobre las características de los ítems, y desde la propia experiencia como receptor del cuestionario. La percepción de los examinados sobre la dificultad de los ítems se ha relacionado con varias características estadísticas de los ítems y con el desempeño esperado en una prueba (Morse & Morse, 2002), los cuales pueden ser mejores aproximaciones de lo que se puede hallar durante el piloteo o aplicación de los instrumentos a la muestra de estudio. Entonces, el juicio de jueces expertos puede no ser un suficiente criterio para valorar el grado de inteligibilidad de un ítem; más bien, el juicio de los examinados puede dar una señal de alerta sobre los ítems que podrían generar varianza no relacionada al constructo (Downing, 2002), y tomar decisiones al respecto. Obtener la percepción de jueces expertos puede parecer, en consecuencia, menos eficaz en este proceso, pues las decisiones que se tomarían sobre la modificación, remoción o permanencia de los ítems serán diferentes de acuerdo al tipo de "juez" que de utilice.

Los resultados conducen a plantear varias cuestiones sobre la valoración de la dificultad de los ítems, efectuada únicamente por un equipo de expertos o incluyendo a los potenciales examinados del instrumento. Primero, que la evaluación de la claridad de los ítems podría ser efectuada por varios grupos de potenciales examinados, más aún si se consideran una amplia cobertura de aplicación de un instrumento (estudiantes no universitarios, trabajadores, etc.), lo que garantizaría mejor la generalizabilidad de estos juicios y aportaría evidencias de la replicación de los resultados. En segundo lugar, que los juicios de profesores-jueces pueden tomarse con precaución cuando evalúan la inteligibilidad de los ítems. Y tercero, que una muestra combinada de participantes-objetivo y jueces expertos pueden dar una mejor contribución.

Una cuestión más crítica es si los jueces expertos pueden desempeñarse mejor respecto al constructo en lugar que calificar el grado de comprensión de los ítems, pero se puede requerir un mejor muestreo de jueces expertos, entre los que se priorice la mayor experiencia posible con el tipo de sujetos a los que el instrumento se dirige.

Algunos aspectos metodológicos pueden resaltarse para la presente investigación. Primero, la valoración obtenida fue mediante el mismo método (cuestionario), y logró capturar las diferentes percepciones en los dos grupos. Esto ayudó a evitar el posible sesgo del método y lograr identificar las diferencias entre los grupos. Por lo tanto, hay una ventaja de aplicar el mismo método de obtención de juicios en evaluados y jueces para propósitos de comparación y cuantificación. Sin embargo, esto no excluye alguna metodología cualitativa que pueda responder al mismo problema. Segundo, se debe señalar que, aunque varios de los coeficientes de acuerdo kappa no fueron estadísticamente significativos, la significancia práctica de los resultados fue de mayor importancia, pues es conocido que el efecto del tamaño muestra regularmente ocasiona una pérdida del poder estadístico en todas las pruebas inferenciales de evaluación del acuerdo (Cicchetti, 2007). Finalmente, debe anotarse que la metodología de intervalos de confianza para el coeficiente V permite una mejor descripción cuantitativa de su valor poblacional (Penfield & Gioacobbi, 2004) y observar las diferencias

entre grupos de manera heurística, tal como se lo aplicó aquí. Sin embargo, se requiere una estimación cuantitativa y de las diferencias, y posiblemente, obtener intervalos de confianza para las diferencias entre coeficientes V sea una opción razonable e interesante. El intervalo de confianza para las diferencias puede tener un valor descriptivo y de prueba de hipótesis (Newcombe, 2012), y ambos aspectos son útiles para producir decisiones mejor informadas sobre la validez de contenido.

El presente estudio concluye que, mediante un método sencillo en su implementación, sus resultados pueden crear una línea base específica para guiar el diseño de estudios de validez de contenido, en el que se incluya la perspectiva del participante mediante la evaluación auto-informada de su compresión de los ítems. Esto también es pertinencia para contribuir a la validez aparente o de fascie, y extender su aplicación hacia otros aspectos de la validez de contenido habitualmente evaluadas, como la relevancia y la representatividad de los ítems respecto a su constructo.

Una limitación importante de la investigación es que el nivel de habilidad lectora del examinado pudo haber puesto un necesario límite a su percepción de claridad de los ítems, pero en el presente estudio esta habilidad se asumió constante o al menos distribuida homogéneamente. Una futura investigación requerirá controlar esta variable mediante la evaluación directa de la capacidad lectura o una aproximación psicométricamente válida de la misma (por ejemplo, Giménez, Luque, López-Zamora & Fernández-Navas, 2015, Lefly & Pennington, 2000, Olson, Smyth, Wang & Pearson, 2011).

Lista de referencias

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40, pp. 955-959.*

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45, pp. 131-142.*

Byrt, T., Bishop, J. & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46, pp. 423-429.*

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23, pp. 695-700.*

Cicchetti, D. V. (2007). Assessing the reliability of blind wine tasting: Differentiating levels of clinical and statistical meaningfulness. *Journal of Wine Economics*, *2 (2), pp. 196-202.*

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112 (1), pp. 155-159. Doi. org/10.1037/0033-2909.112.1.155.*

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J. (…) Kostas-Polson, E. (2007). A psychometric Toolbox for testing Validity and Reliability. *Journal of Nursing Scholarship, 39 (2), pp. 155-164.*

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, *77 (10), pp. 103-104.*

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), pp. 133-143.*

Eysenck, H. J. & Eysenck, S. G. (2001). *Cuestionario de Personalidad de Eysenck, versión revisada (EPQ-R)*. Madrid: TEA.

Fiske, D. W. (1967). The subjects react to test. *American Psychologist*, *22, pp. 287-296.*

Giménez, A., Luque, A. L., López-Zamora, M. & Fernández-Navas, M. (2015). Autoinforme de Trastornos Lectores para Adultos (Atlas). *Anales de Psicología, 3 (1), pp. 109-119. Doi.org/10.6018/ analesps.31.1.166671.*

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah: Lawrence Erlbaum Associates.

Rev.latinoam.cienc.soc.niñez juv 14 (2): 1469-1477, 2016
http://revistalatinoamericanaumanizales.cinde.org.co
*DOI:*10.11600/1692715x.14239120615

Lefly, D. L. & Pennington, B. F. (2000). Reliability and Validity of the Adult Reading History Questionnaire. *Journal of Learning Disability, 33, pp. 286-296. Doi: 10.1177/002221940003300306.*

McCarthy, J. M. & Goffin, R. D. (2003). Is the Test Attitude Survey psychometrically sound? *Educational and Psychological Measurement, 63 (3), pp. 446-464. Doi:10. 1177/0013164403063003007.*

McCrae, R. R., Costa, P. T. Jr. & Martin, T. A. (2005). The NEO-PI-3: A more readable Revised NEO Personality Inventory. *Journal of Personality Assessment, 84, pp. 261-270.*

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), pp. 169-171.*

Morse, D. T. & Morse, L. W. (2002). Are undergraduate examinee's perceptions of item difficulty related to item characteristics? *Perceptual and Motor Skills, 95 (3-2), pp. 1281-1286.*

Nevo, B. (1995). Examinee Feedback Questionnaire: Reliability and validity measures. *Educational and Psychological Measurement*, *55 (3), pp. 499-504. Doi:10.1177/0013164406299132.*

Newcombe, R. G. (2012). *Confidence intervals for proportions and related measures of effect size*. Boca Raton: CRC Press.

Olson, K., Smyth, J. D., Wang, Y. & Pearson, J. E. (2011). The self-assessed literacy index: Reliability and validity. *Social Science Research, 40 (5), pp. 1465-1476. Doi:10.1016/j.ssresearch.2011.05.002.*

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), pp. 213-225.*

Schinka, J. A. & Borum, R. (1994). Readability of normal personality inventories. *Journal of Personality Assessment, 62, pp. 95-101. Doi: http://dx.doi.org/10.1207.*

Secolsky, C. (1983). Using examinee judgments for detecting invalid items on teacher-made criterion-referenced tests. *Journal of Educational Measurement, 20, pp. 51-63.*

Secolsky, C., Kossar, B., Magaram, E. & Fuentes, V. (2011, october). *Estimating examinee intrinsic difficulty for providing greater specificity of feedback for instruction*. Paper presented at the annual meeting of the International Association of Educational Assessment, Manila, Philippines.

Sheskin, J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Ratón: Chapman and Hall/CRC.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22, pp. 123-138. Doi:10.1007/ s10648-010-9128-5.*

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22, pp. 209-212.*