Avances en Medición, 6, 27–36
2008

VALIDEZ DE CONTENIDO Y JUICIO DE EXPERTOS: UNA APROXIMACIÓN A SU UTILIZACIÓN

Jazmine Escobar-Pérez[*]
Universidad El Bosque, Colombia

Ángela Cuervo-Martínez[*]*
Institución Universitaria Iberoamericana, Colombia

Resumen

El presente artículo se centra en la validez de contenido, específicamente, en la utilización del juicio de expertos como parte del proceso para su estimación. Se presenta una conceptualización de la validez de contenido, seguida de la definición y caracterización del juicio de expertos. Finalmente se propone una guía para la realización del juicio que incluye una planilla de calificación con indicadores y la aplicación e interpretación de los estadísticos Kappa y Kendall como índices de concordancia.

Palabras clave: Juicio de expertos, validez de contenido, estadística Kappa, Coeficiente de Kendall.

Abstract

This paper focuses on the content validity, specifically, on the use of experts judgments as part of the process to estimate it. First, a content validity conceptualization is introduced, followed by the definition and characterization of the experts judgment. Finally, a guide to carry out the judgment is proposed including a grade chart with indicators and the application and interpretation of the Kappa and Kendall statistics as concordance indexes.

Key words: Experts judgment, content validity, Kappa's statistics Kendall's coefficient.

Introducción

Una pregunta que surge cuando se intenta medir el comportamiento es qué tan válida y confiable es la medición. El presente escrito se centra en la validez de contenido, específicamente, en la utilización del juicio de expertos como parte del proceso para su estimación. Esta técnica debe realizarse de manera adecuada, ya que muchas veces constituye el único indicador de la validez de contenido. Actualmente el juicio de expertos es una práctica generalizada que requiere interpretar y aplicar sus resultados de manera acertada, eficiente y con toda la rigurosidad metodológica y estadística, para permitir que la evaluación basada en la información obtenida de la prueba pueda ser utilizada con los propósitos para la cual fue diseñada.

La validez de contenido se establece en diferentes situaciones, siendo dos de las más frecuentes: (a) el diseño de una prueba, y (b) la validación de un instrumento que fue construido para una población diferente, pero que se adaptó mediante un procedimiento de traducción (equivalencia semántica). Hyrkäs, Appelqvist--Schmidlechner y Oksa (2003) plantean que es muy frecuente que instrumentos que ya han sido estandarizados en países de habla inglesa sean utilizados en países de habla no inglesa, por lo cual se debe realizar el proceso de traducción, adaptación y estandarización del instrumento para dichos países. Estos procesos presentan dificultades importantes, ya que la sola adaptación del instrumento no necesariamente genera una equivalencia cultural debido a las barreras del idioma, a significados culturales

[*] Facultad de Psicología Universidad El Bosque, Bogotá-Colombia. E-mail: escobarjazmine@unbosque.edu.co

[*] * Facultad de Psicología Institución Universitaria Iberoamericana, Bogotá-Colombia. E-mail: cuervomla@hotmail.com

diferentes de un constructo particular o a una variedad de interpretaciones de un comportamiento observado basado en normas culturales. Es por esto que se hace necesario validar dichos instrumentos en términos de su contenido, y es allí donde la evaluación realizada por expertos cobra especial relevancia, pues son ellos quienes deben eliminar los ítems irrelevantes y modificar los ítems que lo requieran, como en el caso de expresiones idiomáticas.

En este contexto surgen los objetivos del presente artículo: (a) Realizar una aproximación a la conceptualización de la validez de contenido y del juicio de expertos, (b) proponer un procedimiento para su realización, y (c) brindar algunas alternativas estadísticas para el análisis de los datos obtenidos del proceso que permitan tomar decisiones.

Validez de contenido

La validez de contenido consiste en qué tan adecuado es el muestreo que hace una prueba del universo de posibles conductas, de acuerdo con lo que se pretende medir (Cohen & Swerdik, 2001); los miembros de dicho universo U pueden denominarse reactivos o ítems. Para autores como Ding y Hershberger (2002), la validez de contenido es un componente importante de la estimación de la validez de inferencias derivadas de los puntajes de las pruebas, ya que brinda evidencia acerca de la validez de constructo y provee una base para la construcción de formas paralelas de una prueba en la evaluación a gran escala.

Para establecer un posible universo de reactivos se requiere tener una adecuada conceptualización y operacionalización del constructo, es decir, el investigador debe especificar previamente las dimensiones a medir y sus indicadores, a partir de los cuales se realizarán los ítems. Los ítems deben capturar las dimensiones que la prueba pretende medir, por ejemplo, en la prueba de procesos lectores (PROLEC) de Cuetos, Rodríguez y Ruano (2001) el constructo *procesos lectores* está evaluado en las dimensiones de procesos sintácticos, semánticos y pragmáticos. Los ítems seleccionados deben por tanto medir las dimensiones del constructo: Un error de validez de contenido sería que la dimensión semántica no tuviera ningún ítem que la evaluara, o que los ítems de la dimensión sintáctica sólo evaluaran una parte de ésta, al contrastar con lo que se pretende evaluar en dicha dimensión.

El constructo medido por el instrumento y el uso que se les dará a las puntuaciones obtenidas son aspectos fundamentales tanto para la estimación como para la conceptualización de la validez de contenido. En efecto, en la evaluación de un instrumento debe tenerse en cuenta su función, es decir, si será utilizado para el diagnóstico, la medición de habilidades o la medición de desempeño, entre otros; los índices de validez para una función de un instrumento no son necesariamente generalizables a otras funciones del mismo instrumento (Ding & Hershberger, 2002). A su vez, la validez de contenido no sólo puede variar de acuerdo con las poblaciones en las cuales será utilizado el instrumento, sino que puede estar condicionada por un dominio particular del constructo; diferentes autores pueden asignarle el mismo nombre a un constructo, pero poseer diferentes dimensiones y conceptualizaciones, por lo tanto, un instrumento puede tener una validez de contenido satisfactoria para una definición de un constructo pero no para otras. En síntesis, el concepto esencial de validez de contenido es que los ítems de un instrumento de medición deben ser relevantes y representativos del constructo para un propósito evaluativo particular (Mitchell, 1986, citado en Ding & Hershberger, 2002)

Juicio de expertos

La validez de contenido generalmente se evalúa a través de un panel o un juicio de expertos, y en muy raras ocasiones la evaluación está basada en datos empíricos (Ding & Hershberger, 2002). En concordancia con esto, Utkin (2005) plantea que el juicio de expertos en muchas áreas es una parte importante de la información cuando las observaciones experimentales están limitadas. Esta aseveración

es particularmente cierta en el caso de la psicología, donde dicho juicio se ha convertido en la estrategia principal para la estimación de la validez de contenido.

El juicio de expertos se define como una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en éste, y que pueden dar información, evidencia, juicios y valoraciones. La identificación de las personas que formarán parte del juicio de expertos es una parte crítica en este proceso, frente a lo cual Skjong y Wentworht (2000) proponen los siguientes criterios de selección: (a) Experiencia en la realización de juicios y toma de decisiones basada en evidencia o experticia (grados, investigaciones, publicaciones, posición, experiencia y premios entre otras), (b) reputación en la comunidad, (c) disponibilidad y motivación para participar, y (d) imparcialidad y cualidades inherentes como confianza en sí mismo y adaptabilidad. También plantean que los expertos pueden estar relacionados por educación similar, entrenamiento, experiencia, entre otros; y en este caso la ganancia de tener muchos expertos disminuye. Otros autores como McGartland, Berg, Tebb, Lee y Rauch (2003), proponen como criterio básico de selección únicamente el número de publicaciones o la experiencia. Para una discusión sobre educación vs. experiencia en los jueces, ver Summers, Williamson y Read (2004).

De otra parte, el número de jueces que se debe emplear en un juicio depende del nivel de experticia y de la diversidad del conocimiento; sin embargo, la decisión sobre qué cantidad de expertos es la adecuada varía entre autores. Así, mientras Gable y Wolf (1993), Grant y Davis (1997), y Lynn (1986) (citados en McGartland et al. 2003) sugieren un rango de dos hasta 20 expertos, Hyrkäs et al. (2003) manifiestan que diez brindarían una estimación confiable de la validez de contenido de un instrumento. Si un 80 % de los expertos han estado de acuerdo con la validez de un ítem éste puede ser incorporado al instrumento (Voutilainen & Liukkonen, 1995, citados en Hyrkäs et al. (2003).

El juicio de expertos se usa en múltiples ámbitos de la evaluación psicológica, desde la medición de la confiabilidad de los profesionales en salud mental para evaluar la competencia de pacientes psiquiátricos respecto al consentimiento informado (Kitamura & Kitamura, 2000), hasta la validación de contenido de pruebas estandarizadas de altas especificaciones. Existen muchos ejemplos de la utilización específica del juicio de expertos, entre ellos se encuentran Olea, Abad y Ponsoda (2002), quienes diseñaron y calibraron un banco de ítems (635) de conocimiento en gramática inglesa; y Lobo et al. (2003), quienes realizaron la primera validación en español del método INTERMED (sistema de detección temprana de problemas biopsicosociales) y del uso de servicios en pacientes médico- quirúrgicos, asimismo se encuentran aplicaciones del juicio de expertos en el área de detección de riesgos y fiabilidad de procesos.

Pasos para realizar un juicio de expertos

Varios autores como Skjong y Wentworht (2000), y de Arquer (1995) han propuesto diversos pasos para la realización del juicio de expertos: (a) Preparar instrucciones y planillas, (b) seleccionar los expertos y entrenarlos, (c) explicar el contexto, (d) posibilitar la discusión, y (e) establecer el acuerdo entre los expertos por medio del cálculo de consistencia. Además de estos pasos comunes a los diferentes autores, se debe instruir claramente al juez en la dimensión y el indicador que mide cada ítem o un grupo de ellos. Es de especial relevancia brindar información sobre el uso que tendrán los resultados de la prueba, ya que como se mencionó en un apartado anterior, estos están estrechamente relacionados con la validez de contenido. En efecto, utilizaciones diferentes de las puntuaciones harán que varíe la pertinencia y suficiencia de los ítems.

Si tomamos el caso de una prueba de autoestima para deportistas de alto rendimiento, por ejemplo, es diferente el valor que se le atribuye a los ítems si la prueba se va a usar para seleccionar a los deportistas que asistirán a competencias de alta exigencia por primera vez, que si se van a usar para describir un perfil de los diferentes aspectos psicológicos que pueden tener efecto en el desempeño del deportista. En el primer caso los ítems de autoeficacia (una dimensión de la autoestima) deben tener una ponderación más

alta o ser mayor en número frente a otras dimensiones como el autoconcepto y la autoimagen; en el segundo caso, la prueba de autoestima debe medir equilibradamente todas las dimensiones.

Existen varios métodos para la obtención de juicios de expertos, que pueden clasificarse según si la evaluación se realiza de manera individual o grupal. En el primer grupo se encuentran algunos como el método de agregados individuales y el método Delphi, en ambos métodos cada juez realiza la evaluación individualmente, pero en el Delphi, luego de analizar las respuestas se le envía a cada juez la mediana obtenida y se le pide que reconsidere su juicio hasta que se logre un consenso (de Arquer, 1995). Según Van Der Fels-Klerx, Gossens, Saaticamp y Horst (2002) esta técnica ofrece un alto nivel de interacción entre los expertos, evitando las desventajas de la dinámica grupal.

Entre las técnicas grupales se encuentra la nominal y el consenso, en ambas se requiere reunir a los expertos, pero en la última se exige mayor nivel de acuerdo. Estas técnicas pueden tener problemas si se generan discusiones tensas o si existen variables individuales como la personalidad y las habilidades sociales de los jueces que generen sesgos. Otro procedimiento utilizado para el juicio de expertos se basa en el emparejamiento de los ítems con el dominio. En este caso se entrega a los jueces una lista de objetivos (categorías) y se les presenta cada ítem en una ficha separada. El juez compara cada ítem con la lista y registra el resultado en una hoja de respuestas, indicando al lado de cada ítem el número del objetivo. (Martínez, 1995). La precisión de los juicios según Stewart, Roebber y Bosart, (1997) depende tanto de las características de los jueces y de su experiencia, como de las características de la tarea. Dentro de esta última, la teoría cognitiva sugiere tres categorías importantes: (a) La complejidad de la estructura de la tarea, (b) la ambigüedad en el contenido de la tarea, y (c) la forma de la presentación de la tarea.

Guía para la realización de un juicio de expertos

El juicio de expertos es un procedimiento que nace de la necesidad de estimar la validez de contenido de una prueba. Para realizarlo se debe recabar información de manera sistemática. A continuación se proponen una serie de pasos que permiten organizar la información, de manera que el proceso de juicio de expertos sea más eficiente.

1. *Definir el objetivo del juicio de expertos*. En este apartado los investigadores deben tener clara la finalidad del juicio, ya que puede utilizarse con diferentes objetivos: (a) Establecer la equivalencia semántica de una prueba que se encuentra validada en otro idioma, (b) evaluar la adaptación cultural, es decir, el objetivo de los jueces es evaluar si los ítems de la prueba miden el mismo constructo en una cultura distinta; así por ejemplo, los ítems que midan agresividad en una prueba validada en el Tibet, pueden no estar midiendo lo mismo en Alemania, y (c) validar contenido en una prueba diseñada por un grupo de investigadores.

2. *Selección de los jueces*. Para ello han de tomarse en cuenta los criterios especificados anteriormente para la selección, considerando la formación académica de los expertos, su experiencia y reconocimiento en la comunidad. Se propone un mínimo de cinco jueces, dos de los cuales deben ser expertos en medición y evaluación, y para el caso de traducciones y adaptaciones de pruebas, se requiere por lo menos un experto en lingüística.

3. *Explicitar tanto las dimensiones como los indicadores que está midiendo cada uno de los ítems de la prueba*. Esto le permitirá al juez evaluar la relevancia, la suficiencia y la pertinencia del ítem. No hay que dar por sentado que el juez únicamente con la descripción del constructo a medir pueda identificarlo claramente, ya que como se mencionó anteriormente, es posible que existan diferentes definiciones de un mismo constructo.

4. *Especificar el objetivo de la prueba*. El autor debe proporcionar a los jueces la información relacionada con el uso de la prueba, es decir, para qué van a ser utilizados los puntajes obtenidos a partir

de ésta. Esto aumenta la contextualización del juez respecto a la prueba, incrementando a su vez el nivel de especificidad de la evaluación; ya que la validez de los ítems está directamente relacionada con su utilización, por ejemplo, para hacer un diagnóstico o un tamizaje, o evaluar desempeño, entre otros.

5. *Establecer los pesos diferenciales de las dimensiones de la prueba*. Esto sólo se hace cuando algunas de las dimensiones tienen pesos diferentes. Por ejemplo, si una prueba va a ser utilizada para el diagnóstico y asignación a un programa de rehabilitación de una adicción, se debe dar mayor peso a las dimensiones que midan la calidad de vida que a las que evalúen personalidad adictiva.

6. *Diseño de planillas*. La planilla se debe diseñar de acuerdo con los objetivos de la evaluación. No obstante, en el Anexo 1 proponemos una planilla que puede ser utilizada en la gran mayoría de juicios de expertos, con sus respectivos indicadores para la calificación.

7. *Calcular la concordancia entre jueces*. Para esto se utilizan los estadísticos Kappa y Kendall que se describirán a continuación. La información sobre cada estadístico, las hipótesis de trabajo y los criterios de interpretación, se muestran en la tabla 1.

8. Elaboración de las conclusiones del juicio que serán utilizadas para la descripción psicométrica de la prueba.

Estadísticos para análisis

Para estimar la confiabilidad de un juicio de expertos, es necesario conocer el grado de acuerdo entre ellos, ya que un juicio incluye elementos subjetivos (Aiken, 2003). Cuando la medida de acuerdo obtenida es alta indica que hay consenso en el proceso de clasificación o asignación de puntajes entre los evaluadores, igualmente da cuenta de la intercambiabilidad de los instrumentos de medición y reproducibilidad de la medida. (Ato, Benavente & López, 2006).

Para determinar el grado de acuerdo entre los jueces se han utilizado diferentes procedimientos, una aproximación inicial fue calcular el porcentaje de acuerdo, medida que resulta insuficiente ya que no incluye el acuerdo esperado por el azar (Jakobsson & Westergren, 2005). Luego se incluyeron medidas de correlación que eran interpretadas como índices de acuerdo; sin embargo un alto índice de correlación no necesariamente implica que el acuerdo sea alto también. Artstein y Poesio (2005) adaptaron un ejemplo de Barko y Carpenter (1976) (citados en Artstein & Poesio, 2005) que refleja esta situación: En dos evaluaciones, dos codificadores asignaban a cada ítem una puntuación entre uno y diez, en la primera evaluación los codificadores A y B están completamente de acuerdo; en la segunda evaluación los codificadores C y D están en desacuerdo en todos los ítems, pero les asignan valores que están linealmente correlacionados. En los dos casos se obtiene el mismo índice, con lo que queda claramente expresada la inconveniencia de medidas únicamente de correlación para la estimación del acuerdo.

Posteriormente se propuso el coeficiente Kappa, que se convirtió rápidamente en el índice de acuerdo más utilizado en ciencias biológicas y sociales. Inicialmente el coeficiente se utilizaba únicamente en datos nominales, después se hizo una generalización para incluir datos ordinales a este nuevo coeficiente al que se le denominó *weighted k-coefficient.* Kendall también propuso un coeficiente de acuerdo para datos ordinales, basado en el grado de varianza de la suma de los rangos obtenidos de los diferentes jueces. Actualmente se vienen investigando otros procedimientos para estimar el acuerdo, se están aplicando los modelos log-lineales y los mixtos (mezcla de distribuciones). En el primero se analizan tanto la estructura del acuerdo como la del desacuerdo que se presentan en los datos, con este enfoque se puede conocer el ajuste del modelo y se puede aplicar a datos ordinales; mientras que en el segundo se incluyen variables latentes (Ato et al., 2006).

Estadístico Kappa. Este estadístico genera una medida de acuerdo entre evaluadores y se utiliza cuando las variables están dadas en una escala nominal, es decir únicamente clasifican. Por ejemplo, un juez clasifica los ítems de una prueba de conocimientos en contestables o no contestables por una persona que tenga un nivel adecuado de conocimiento en el área, o el caso de psicólogos clínicos que tienen que clasificar a pacientes entre los que requieren seguimiento permanente y los que no.

El estadístico tiene un rango entre -1 y 1, pero generalmente se ubica entre 0 y 1. Si el coeficiente es 1 indica acuerdo perfecto entre los evaluadores, si es 0 indica que el acuerdo no es mayor que el esperado por el azar, y si el valor del coeficiente es negativo el nivel se acuerdo es inferior al esperado por el azar (Sim & Wright, 2005). No obstante, obtener estos valores extremos es improbable, lo común es obtener un amplio espectro de valores intermedios que se interpretan teniendo como referencia la complejidad de la evaluación y el número de categorías a evaluar, es decir, la interpretación es relativa al fenómeno medido. En el caso de los psicólogos que deciden cuáles pacientes requieren supervisión y cuáles no, como la complejidad de la evaluación es moderada (con sólo dos categorías de clasificación), se espera un alto acuerdo entre ellos. Un acuerdo de 0.55 sería considerado bajo, y se podría inferir que hay dificultad en la clasificación, o que incluso, pueden tener ambigüedad en los indicadores que les permiten decidir en uno u otro sentido. En otro caso, si en un colegio el objetivo es clasificar los alumnos con trastornos de aprendizaje y discapacidad, para identificar el número de casos de dislexia, discalculia, disgrafía, discapacidad cognoscitiva, y discapacidad sensorial; obtener 0.55 se interpretaría como un índice de acuerdo moderado, atendiendo a la mayor complejidad de la evaluación. Sin embargo si dicha clasificación se va a realizar con el objetivo de enviarlos a terapia o a aulas de apoyo se requiere un acuerdo mayor, al igual que si se trata de ítems para validación de una prueba.

El coeficiente de Kappa tiene como ventaja que corrige el porcentaje de acuerdo debido al azar y es muy sencillo de calcular. Sin embargo, se han realizado varias críticas principalmente relacionadas con que el índice de acuerdo se ve afectado por el número de categorías y por la forma en la que están distribuidas las observaciones.

Coeficiente de concordancia W de Kendall: Este coeficiente se utiliza cuando se quiere conocer el grado de asociación entre *k* conjuntos de rangos (Siegel & Castellan, 1995), por lo cual es particularmente útil cuando se les solicita a los expertos asignarle rangos a los ítems, por ejemplo de 1 a 4. El mínimo valor asumido por el coeficiente es 0 y el máximo 1, y su interpretación es la misma que para el coeficiente de Kappa. Sin embargo, hay que hacer la salvedad que hay que revisar la calificación dada a cada ítem, ya que puede haber una alta concordancia en los aspectos, un ejemplo de ello es que el ítem no sea adecuado. Obviamente en este caso se debe eliminar o modificar el ítem completamente hasta que ajuste a los objetivos de la medición de forma acertada.

Según Siegel y Castellan (1995), un valor alto de la *w* puede interpretarse como un reflejo de que los *k* observadores o jueces están aplicando los mismos estándares al asignar rangos a los ítems. Esto no garantiza que los ordenamientos observados sean correctos, ya que todos los jueces pueden concordar si todos están utilizando un criterio incorrecto para clasificar. Es debido a esto último que el criterio de selección de jueces cobra especial relevancia al igual que la independencia entre los mismos.

Para estimar en SPSS 14 el coeficiente de Kappa siga estos pasos: a) Haga clic en Analizar y seleccione Estadísticos descriptivos, b) Haga clic en Tablas de contingencia, allí encontrará un cuadro de diálogo y c) Haga clic en Estadísticos y seleccione Kappa.

Para estimar en SPSS 14 el coeficiente de Kendall siga estos pasos: a) Haga clic en Analizar y seleccione Pruebas no paramétricas, b) Haga clic en *k* muestras relacionadas y seleccione W de Kendall y c) seleccione Kendal (ver tabla 1).

Tabla 1.
Resumen de estad´siticos para el análisis de los datos

COEFICIENTES	ESCALA DE LOS DATOS	INFORMACIÓN QUE PROVEE	HIPÓTESIS	RECHAZO DE H_0 E INTERPRETACIÓN
Coeficiente de concordancia *W* de Kendall	Escala ordinal.	El grado de concordancia entre varios rangos de *n* objetos o individuos. Aplicable a estudios interjuicio o confiabilidad interprueba.	H_0: Los rangos son independientes, no concuerdan. H_1: Hay concordancia significativa entre los rangos.	Se rechaza H_0 cuando el valor observado excede al valor crítico (con un α de 0.05). El SPSS indica el nivel de significancia, y cuando es inferior al 0.05, se rechaza la H_0 y se concluye que hay concordancia significativa entre los rangos asignados por los jueces. Además se interpreta la fuerza de la concordancia, que aumenta cuando *W* se acerca a 1.
Estadístico Kappa (*K*) para datos en escalas nominales.	Escala nominal	El grado de acuerdo entre evaluadores	H_0: El grado de acuerdo es 0, es decir no hay acuerdo. H_1: Existe un acuerdo significativo entre evaluadores, es decir K >0	Al igual que en el caso anterior se rechaza H_0 cuando el valor observado excede al valor crítico (con un α de 0.05). El SPSS indica el nivel de significancia, y cuando es inferior al 0.05, se rechaza la H_0 y se concluye que hay acuerdo entre los evaluadores, el valor de *k* brinda la proporción de acuerdo quitándole el acuerdo que puede darse por azar.

Recomendaciones finales

Hay aspectos dentro del juicio de expertos que no pueden ser controlados por el investigador, como por ejemplo la complejidad o el nivel de dificultad de la tarea; sin embargo, los factores de ambigüedad del contenido de la tarea y su forma de presentación deben manejarse en el procedimiento de juicio de expertos de manera que no aumenten el error ni disminuyan la confiabilidad. Otro aspecto a considerar es que el investigador debe propiciar el contexto adecuado para obtener la mayor cantidad de información posible de los jueces expertos y solicitar opiniones adicionales sobre la prueba que pueden dar información sobre aspectos que no se evaluaron en el juicio. Finalmente, se debe recordar que aunque una prueba obtenga una muy buena evaluación de los jueces y un alto índice de concordancia, debe estar en continua revisión y mejoramiento.

Referencias

Aiken, Lewis (2003). *Test psicológicos y evaluación*. México: Pearson Education.

Artstein, R. & Poesio, M. (2005). *Kappa3 = Alpha (or Beta)*. (Technical Report CSM-437). Department of Computer Science: University of Essex.

Ato, M., Benavente, A., & López, J. J. (2006). Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores. Psicothema, 18(3), 638 – 645.

Cohen, R. & Swerdlik, M. (2001). Pruebas y evaluación psicológicas: Introducción a las pruebas y a la medición. (4ª ed.). México: Mc Graw Hill.

Cuetos, F., Rodríguez, B & Ruano, E (2001). PROLEC, Batería de evaluación de los procesos lectores de los niños de educación primaria. Madrid: TEA Ediciones.

de Arquer, M. (1995). *Fiabilidad Humana: métodos de cuantificación, juicio de expertos*. Centro Nacional de Condiciones de Trabajo. Recuperado el 3 de Junio de 2006, de http://www.mtas.es/insht/ntp/ntp_401.htm

Ding, C. & Hershberger, S. (2002). Assessing content validity and content equivalence using structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal,* 9 (2), 283-297.

Hyrkäs, K., Appelqvist-Schmidlechner, K & Oksa, L. (2003). Validating an instrument for clinical supervision using an expert panel. *International Journal of nursing studies*, 40 (6), 619 -625.

Jakobsson, U. &y Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of sCaring Science*, 19(4), 427-431.

Kitamura, T. & Kitamura, F. (2000). Reliability of clinical judgment of patients' competency to give informed consent: A case vignette study. *Psychiatry and Clinical Neurosciences*, 54(2), 245-247.

Lobo, E. Bellido, M. Campos, R., Saz, P., Huyse, F., De Jonge, P. & Lobo, A. (2003). Primera validación en español del método INTERMED: Un sistema de temprana detección de problemas biopsicosociales y de consumo de servicios en pacientes médico-quirúrgicos. *Cuadernos de Medicina Psicosomática y Psiquiatría de Enlace*, 67/68, 89- 97.

Martínez, R. (1995). Psicometría: teoría de los test psicológicos y educativos. Madrid: Editorial Síntesis.

McGartland, D. Berg, M., Tebb, S. S., Lee, E. S. & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research,* 27 (2), 94-104.

Olea, J, Abad, F. J. & Ponsoda, V. (2002). Elaboración de un banco de ítems, predicción de la dificultad y diseño de anclaje. *Metodología de las ciencias del comportamiento, Vol. Especial,* 427-430.

Siegel, S. & Castellan, N. J. (1995) Estadística no paramétrica aplicada a las ciencias de la conducta. México: Trillas.

Sim, J. & Wright, C. (2005) The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85 (3), 257-268.

Skjong, R. & Wentworth, B. (2000). *Expert Judgement and risk perception.* Recuperado el 15 de Enero de 2006, de http://research.dnv.com/skj/Papers/SkjWen.pdf

Stewart, T., Roebber, P. & Bosart, L. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision processes*, 69 (3), 205-219.

Summers, B., Williamson, T. & Read, D. (2004). Does method of acquisition affect the quality of expert judgment? A comparison of education with on-the-job learning. *Journal of Occupational and Organizational Psychology*. 77(2), 237-258.

Utkin, L. V. (2006). A method for processing the unreliable expert judgments about parameters of probability distributions. [Versión Electrónica]. *European Journal of Operational Research*. 175(1), 385-398.

Van Der Fels-Klerx, I. Gossens, L. Saaticamp, H & Horst, S. (2002). Elicitation of quantitative data from a heterogeneous Expert Panel: Formal process and application in animal health. *Risk Analisis*, 22 (1), 67-81.

Manuscrito recibido en Septiembre de 2006
Aceptado para publicación en Noviembre de 2007

Anexo 1: Planillas Juicio de Expertos

Respetado juez: Usted ha sido seleccionado para evaluar el instrumento_____ que hace parte de la investigación _____. La evaluación de los instrumentos es de gran relevancia para lograr que sean válidos y que los resultados obtenidos a partir de éstos sean utilizados eficientemente; aportando tanto al área investigativa de la psicología como a sus aplicaciones. Agradecemos su valiosa colaboración.

NOMBRES Y APELLIDOS DEL JUEZ: _____

FORMACIÓN ACADÉMICA _____

AREAS DE EXPERIENCIA PROFESIONAL_____

TIEMPO_____ CARGO ACTUAL _____

INSTITUCIÓN_____

Objetivo de la investigación: _____

Objetivo del juicio de expertos: _____

Objetivo de la prueba: _____

De acuerdo con los siguientes indicadores califique cada uno de los ítems según corresponda.

CATEGORIA	CALIFICACIÓN	INDICADOR
SUFICIENCIA	1 No cumple con el criterio	Los ítems no son suficientes para medir la dimensión
	2. Bajo Nivel	Los ítems miden algún aspecto de la dimensión pero no corresponden con la dimensión total
Los ítems que pertenecen a una misma dimensión bastan para obtener la medición de ésta.	3. Moderado nivel	Se deben incrementar algunos ítems para poder evaluar la dimensión completamente.
	4. Alto nivel	Los ítems son suficientes
CLARIDAD	1 No cumple con el criterio	El ítem no es claro
	2. Bajo Nivel	El ítem requiere bastantes modificaciones o una modificación muy grande en el uso de las palabras de acuerdo con su significado o por la ordenación de las mismas.
El ítem se comprende fácilmente, es decir, su sintáctica y semántica son adecuadas.	3. Moderado nivel	Se requiere una modificación muy específica de algunos de los términos del ítem.
	4. Alto nivel	El ítem es claro, tiene semántica y sintaxis adecuada.
COHERENCIA	1 No cumple con el criterio	El ítem no tiene relación lógica con la dimensión
	2. Bajo Nivel	El ítem tiene una relación tangencial con la dimensión.
El ítem tiene relación lógica con la dimensión o indicador que está midiendo.	3. Moderado nivel	El ítem tiene una relación moderada con la dimensión que esta midiendo.
	4. Alto nivel	El ítem se encuentra completamente relacionado con la dimensión que está midiendo.
RELEVANCIA	1 No cumple con el criterio	El ítem puede ser eliminado sin que se vea afectada la medición de la dimensión
	2. Bajo Nivel	El ítem tiene alguna relevancia, pero otro ítem puede estar incluyendo lo que mide éste.

El ítem es esencial o importante, es decir debe ser incluido.	3. Moderado nivel 4. Alto nivel	El ítem es relativamente importante. El ítem es muy relevante y debe ser incluido.

DIMENSIÓN	ITEM	SUFICIENCIA*	COHERENCIA	RELEVANCIA	CLARIDAD	OBSERVACIONES
X1						
X2						
X3						

¿Hay alguna dimensión que hace parte del constructo y no fue evaluada? ¿Cuál?_____

_____

*Para los casos de equivalencia semántica se deja una casilla por ítem, ya que se evaluará si la traducción o el cambio en vocabulario son suficientes.