CUANTIFICACION DE LA VALIDEZ DE CONTENIDO POR CRITERIO DE JUECES

Luis Miguel Escurra M.*

El presente estudio evalúa tres formas de cuantificar la validez de Contenido por criterio de Jueces: el Indice de Acuerdo (IA), la Prueba Binomial (PB) y el coeficiente V de Aiken (V); computados en base a todas las respuestas posibles asignadas por 10 jueces a un ítem. Los resultados permiten concluir que el coeficiente V de Aiken es el más adecuado para determinar este tipo de validez, ya que permite obtener valores factibles de ser contrastados estadísticamente según el tamaño de la muestra de jueces seleccionada.

Three ways of assessing content validity are examined: inter-rater agreement (IA), binomial test (BT) and Aiken V coeficiente (V). Aiken V coeficient proves to be more adequate for assessing content validity.

* Docente de la Pontificia Universidad Católica del Perú.

Un aspecto relevante en el proceso de construcción de instrumentos psicológicos es el garantizar su validez, la que ha sido definida como: el grado en que un test mide lo que se propone medir (Anastasi, 1968); lo cual equivaldría a responder a la pregunta ¿qué mide el test?. Las formulaciones propuestas para solucionar esta inquietud, fue en sus inicios algo caótico, pues los autores tendían a definir desde su propio punto de vista el concepto de validez; así por ejemplo tenemos que Anastasi se refiere a validez empírica y aparente, Gullisken a validez intrínseca, Mosier a validez de definición y otros se refieren a validez curricular, factorial, etc. (Cortada de Kohan, 1968; Ebel, 1977). Hasta 1954 en que la Asociación Psicológica Americana se propuso uniformizar la terminología y formular las reglas básicas para la estandarización de los tests (Anastasi, 1986) adoptando la clasificación tripartita que aún subsiste y asumió como vigentes los siguientes tipos de validez: de contenido, de construcción y la de criterio, que se subdividió en predictiva y concurrente.

Con relación a la validez de contenido vemos que esta ha sido definida como el grado en que los ítem que constituyen la prueba son una muestra representativa del dominio de contenido que se mide (Nunnally, 1973; Mehrens y Lehmann, 1982). Usualmente se ha recomendado que este tipo de validez sea asignado a las pruebas de rendimiento escolar (Cronbach, 1972; Wood, 1975; Magnusson, 1976; Gronlund, 1980; Thorndike, 1986) y en algunos casos para las pruebas de adaptación basadas en observaciones (Karmel, 1984); aunque también se ha sugerido su utilización en escalas de actitudes y otras mediciones de rasgos (Bohmsted, 1978).

Por lo general esta forma de validez se ha determinado mediante la comparación sistemática de los ítem de la prueba con el dominio de contenido estudiado, este análisis es factible de ser llevado a cabo de dos formas, la primera que consiste en estudiar de manera lógica y racional los ítem explicitando el porque se incluye en la prueba; y la segunda en la cual con ayuda de un grupo de jueces competentes y calificados se evalúa el grado en que los reactivos concuerdan con los planteamientos del construct[illegible] instrumento, siendo denominada esta técnica como el criterio de jueces (Andreani, 1975);

y que en muchos casos ha sido la estrategia usada por excelencia para evaluar la validez de contenido (Aiken, 1980).

La modalidad mas común para realizar la validez de contenido por criterio de los jueces, consiste en solicitar la aprobación o desaprobación de la inclusión de un ítem en la prueba por parte de varios jueces, cuyo número puede variar según los requerimientos del autor del instrumento.

Quizás el problema más importante derivado del uso que esta técnica ha generado, es el referido a la dificultad para la cuantificación de sus resultados (Aiken, 1980; Brown, 1980). Un intento de solución a este problema ha sido calcular un índice de acuerdo entre los jueces al evaluar el ítem, siendo en este caso el grado de concordancia el que indicaría la confiabilidad de los juicios (Tinsley y Weiss, 1975), y por ende la validez del mismo, ya que evaluaría el concenso que existe para la inclusión del ítem de la prueba.

Matemáticamente este índice de acuerdo ha sido definido como la proporción que existe entre los juicios que coinciden con la definición propuesta por el autor (acuerdo A) y el total de juicios emitidos (acuerdos A y desacuerdos D); siendo su fórmula IA= A/(A+D), tomándose como válidos los reactivos cuyos valores sean iguales o mayores que 0.80 (Guilford, 1954).

Si bien este intento permitió solucionar parcialmente el problema de la cuantificación de la validez de contenido, también ha dado orígen a otras controversias ya que no se ha indicado cual puede ser el número adecuado de jueces, pues el valor limite de 0.80, puede ser obtenido tanto para grupos de 5 como para grupos de más de 10 jueces, asimismo no se conoce la significación estadística de los resultados, lo cual podría hasta cierto punto indicarnos que sería algo arbitrario y subjetivo trabajar bajo esta forma de validez.

Es con miras a solucionar esta situación que nos proponemos cuantificar la validez de contenido por criterio de jueces aplicando como análisis estadísticos, la prueba Binomial y el coeficiente V de Aiken.

La Prueba Binomial

Es un análisis estadístico que estudia la probabilidad de obtener x objetos en una categoría y n-x objetos en la otra (Hoel, 1976).

La fórmula de cálculo es la siguiente:

$$P_{(x)} = \binom{n}{x} p^x \, q^{(n-x)}$$

siendo:

p = Proporción de casos esperados en una de las categorías

q = 1 - p proporción de casos esperado en la otra categoría

Para el caso de la validez de contenido, las categorías son p (acuerdos) y q (desacuerdos) y se asume que p = q = 0.50. Se elige esta prueba porque los datos son dicotómicos y se tiene un solo grupo de sujetos (Siegel, 1980). El cálculo realizado nos da la probabilidad de ocurrencia de manera directa, de manera que si es menor de .05 ó .01, se asume que el ítem posee validez de contenido.

El Coeficiente de Validez V (Aiken, 1980; 1985)

Es un coeficiente que se computa como la razón de un dato obtenido sobre la suma máxima de la diferencia de los valores posibles. Puede ser calculado sobre las valoraciones de un conjunto de jueces con relación a un ítem o como las valoraciones de un juez respecto a un grupo de ítem. Asimismo las valoraciones asignadas pueden ser dicotomicas (recibir valores de 0 ó 1) ó politomicas (recibir valores de 0 a 5). Para nuestro caso se calculará para respuestas dicotomicas y el análisis de un ítem por un grupo de jueces, haciendo para ello uso de la siguiente fórmula:

$$V = \frac{S}{(n(c-1))}$$

siendo:

S = la sumatoria de si

s_i= Valor asignado por el juez i,

n = Número de jueces

c = Número de valores de la escala de valoración (2. en este caso)

Este coeficiente puede obtener valores entre 0 y 1, a medida que sea más elevado el valor computado, el ítem tendrá una mayor validez de contenido. El resultado puede evaluarse estadísticamente haciendo uso de la tabla de probabilidades asociadas de cola derecha, tabuladas por el autor.

Es precisamente esta posibilidad de evaluar su significación estadística lo que hace a este coeficiente uno de los más apropiados para estudiar este tipo de validez.

Metodología

Este estudio es de tipo metodológico (Kerlinger, 1975), y consiste en la aplicación de las fórmulas del Indice de Acuerdo (IA), Prueba Binomial y el Coeficiente de Validez de Aiken (V), computándose para el caso hipotético del análisis de la validez de contenido de un ítem por un grupo de jueces entre 5 y 10 personas. Igualmente se ha considerado sólo los tres valores más altos de acuerdo, con la finalidad de determinar cual es el valor óptimo de selección, así como el tamaño mínimo del grupo de jueces necesario para considerar los resultados como estadísticamente significativos; de tal manera que sea posible determinar objetivamente la validez de contenido del ítem.

Resultados y Discusión

En la Tabla 1, encontramos en primer lugar que los valores computados para el IA y el V de Aiken, arrojan resultados similares, con lo cual comprobamos que para el caso de los ítem dicotómicos ambas fórmulas son equivalentes; y en segundo lugar, que las probabilidades asociadas a la PB y al V son parecidas, de tal modo que podemos concluir que estos resultados nos confirman la pertinencia del coeficiente V ya que tiene la facilidad del cómputo del IA y la posibilidad de la contrastación estadística de la PB.

En lo que respecta a la proporción de acuerdos que debe existir por cada grupo de jueces para evaluar la validez de contenido, encontramos:

- En grupos de 5, 6 y 7 jueces, se necesita un completo acuerdo entre ellos para que el ítem sea válido.
- En un grupo de 8 jueces, se requiere que deben estar por lo menos 7 jueces en concordancia para que el ítem sea válido a un nivel de significación estadística de $p < .05$.
- En un grupo de 9 jueces, por lo menos 8 de ellos deben estar de acuerdo en la evaluación del ítem para que tenga validez de contenido, asumiendo un nivel de significación estadística de $p < .05$.
- Para el caso de contar con 10 jueces, se necesita el acuerdo de por lo menos 8 de ellos para que a un nivel de $p < .05$ el ítem sea considerado como válido.

De estos resultados, podemos concluir que a medida que se tengan grupos de jueces más numerosos, se requiere que la concordancia sea algo menor, sin por ello dejar de ser válido el ítem evaluado. El asumir como adecuado el valor del índice de acuerdo como mayor de 0.80, es solo relativo y depende del tamaño de la muestra de jueces que se estudia, por lo que se

Tabla No. 1

Jueces	Acuerdos	IA	PB	V	p
5	3	0.60	.312	0.60	
	4	0.80	.156	0.80	
	5	1.00	.031	1.00	.032
6	4	0.67	.234	0.67	
	5	0.83	.094	0.83	
	6	1.00	.016	1.00	.016
7	5	0.71	.164	0.71	
	6	0.86	.054	0.86	
	7	1.00	.008	1.00	.008
8	6	0.75	.109	0.75	
	7	0.88	.031	0.88	.035
	8	1.00	.004	1.00	.004
9	7	0.77	.070	0.77	
	8	0.89	.018	0.89	.020
	9	1.00	.002	1.00	.002
10	8	0.80	.043	0.80	.049
	9	0.90	.009	0.90	.001
	10	1.00	.000	1.00	.001

recomienda se tome en cuenta los resultados encontrados y aceptar como válido solo los ítem que sean estadísticamente significativos a .05.

Finalmente podemos concluir que para evaluar la validez de contenido por criterio de jueces, es preferible hacer uso del coeficiente V de Aiken, que combina la facilidad del cálculo y la evaluación de los resultados con la correspondiente docimacia estadística, con lo cual garantizamos la objetividad del procedimiento, a la vez que solucionamos el problema que plantea la cuantificación de la validez de contenido, de tal forma que sea factible impulsar el desarrollo de la construcción de instrumentos psicológicos adecuados a nuestro medio tanto con fines de investigación como de trabajo profesional en específico.

BIBLIOGRAFIA

Aiken, L. (1980). Content Validity and Reliability of Single Items or Questionnaire. *Educational and Psychological Measurement 40*, 955-959,

Aiken, L. (1985). Three Coeficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement 45*, 131-142.

Anastasi, A. (1968). Tests Psicológicos. Madrid: Aguilar.

Anastasi, A. (1986). Evolving Concepts of Test Validation. *Annual Review Psychology, 37*, 1-15.

Andreani, O. (1975). *Aptitud Mental y Rendimiento Escolar*. Barcelona: Herder.

Bohrnstedt, G. (1980). Evaluación de la Confiabilidad y Validez en la medición de actitudes. En: Sumers G. (Ed.) *Medición de Actitudes*. México: Trillas

Brown, F. (1980). *Principios de la Medición en Psicología y Educación*. México: El Manual Moderno.

Cortada de Kohan, N. (1968). *Manual para la Construcción de Tests Objetivos*. Buenos Aires; Paidos.

Cronbach, L.J. (1972). *Fundamentos de la Exploración Psicológica*. Madrid: Biblioteca Nueva.

Ebel, R.L. (1977). *Fundamentos de la Medición Educacional*. Buenos Aires: Guadalupe.

Gronlund, N. (1980). *Elaboración de Tests de Aprovechamiento*. México: Trillas.

Guilford, J.P. (1954). *Psychometrics Methods*. New York; McGraw-Hill.

Hoel, Paul (1976). *Introducción a la Estadística Matemática*. Barcelona: Ariel.

Karmel, L.J. (1974). *Medición y Evaluación Escolar*. México: Trillas.

Kerlinger, F. (1975). *Investigación del Comportamiento: Técnicas y Metodología*. México: Interamericana.

Magnusson, D. (1976). *Teoría de los Tests*. México: Trillas.

Mehrens, A., William & Lehmann I.J. (1982). *Medición y Evaluación en la Educación y en la Psicología*. México: C.E.C.S.A.

Nunnally, J. (1973). *Introducción a la Medición Psicológica*. Buenos Aires: Paidos.

Siegel, S. (1980). *Estadísticas no Paramétricas Aplicadas a las Ciencias de la Conducta. México:* Trillas.

Thorndike, R. & Hagen, E. (1986). *Tests y Técnicas de Medición en Psicología y Educación*. México: Trillas.

Tinsley H.E.A. & Weiss D.J. (1975). Interrater Reliability and Agreement of Subjective Judgements. *Journal of Counseling Psychology, 22,* 358-376.
Wood, D.A. (1975). *Elaboración de Tests: Desarrollo e Interpretación de los Tests de Aprovechamiento.* México: Trillas.