EQUIVALENCIA PSICOMÉTRICA DE TRES FORMATOS DE PREGUNTA

Elsy Urdaneta Durán*

RESUMEN

En esta investigación se comparó la validez y la confiabilidad de tres distintos formatos de pregunta: elección múltiple, respuesta corta y desarrollo. Para el estudio de la validez se reunió evidencia en relación a la equivalencia del rasgo, la dimensionalidad y a la validación convergente y discriminante. Para el estudio de la confiabilidad se calculó el coeficiente alfa y se obtuvieron las funciones de información para subtests con tiempos de ejecución aproximadamente iguales. Los resultados permiten concluir que los formatos estudiados tienen características psicométricas distintas. Las evidencias en relación a la validez denotan que cada formato mide una dimensión distinta del constructo y en relación a la confiabilidad se concluye que se obtienen puntuaciones de mayor precisión con el formato de elección múltiple.

Palabras Clave: *formato de la pregunta, validez y confiabilidad.*

* Doctora en Metodología de las Ciencias del Comportamiento (Universidad Autónoma de Madrid), Diplomado de Estudios Avanzados en Metodología de las Ciencias del Comportamiento (UAM). Profesora Asociada en el Área de Estadística en el Núcleo "Rafael Rangel" de la Universidad de Los Andes Trujillo–Venezuela, E-mail: elsyurdaneta@ula.ve

Recibido:29/06/2012 **Aprobado:** 13/11/2012

EQUIVALENCE PSYCHOMETRIC OF THREE QUESTION FORMATS

ABSTRACT

The psychometric equivalence of multiple-choice, free response and performance assessment items was studied by comparing the validity and reliability of test scores based on items with those formats. In order to study the validity, evidence was gathered across different lines of research: data were collected on the equivalence of the construct measured, the dimensionality of tests was examined, and convergent and discriminant evidence was obtained. To study the reliability, the Cronbach's alpha coefficient was calculated, and the information functions were obtained. The results showed that the formats studied have different psychometric characteristics and that each format may be measuring a different dimension of the construct with different levels of reliability.

Key words: question formats, validity and reliability.

INTRODUCCIÓN

La toma de decisiones basadas en las puntuaciones obtenidas en pruebas o exámenes son un tema de gran impacto social, puesto que afectan directamente la vida de las personas. Razón suficiente para exigir que las pruebas aplicadas tengan un adecuado grado de calidad psicométrica. Esto quiere decir, que midan lo que dicen medir y que lo hagan de la manera más precisa posible. El primer paso para lograr este propósito es escribir las preguntas apropiadas para evaluar correctamente los contenidos de interés.

Las pruebas con preguntas de elección múltiple han sido la opción preferida durante años, cuando el número de sujetos a aplicar la prueba es grande y se requiere además evaluar un buen número de contenidos en la misma prueba. Tal es el caso de las prueba de acceso a la universidad en nuestro país y algunas pruebas de oposición para optar a cargos públicos en un conjunto importante de países del mundo en las que el formato de pregunta que prevalece es el de elección múltiple. Además,

es el formato usual en las encuestas de opinión y en la mayoría de investigaciones sociales a gran escala, tales como encuestas electorales y macroencuestas económicas. El formato de elección exige la selección de una respuesta entre varias que son dadas junto con el enunciado de la pregunta. Los argumentos a favor de este tipo de preguntas son su bajo costo operativo, su facilidad de calificar objetivamente, su mayor precisión y las evidencias de validez basadas en el contenido (Mumford, Baughman, Supinski y Anderson, 1998; Osterlind, 1998; Ryan y Greguras, 1998). Pero pese a estas características también se ha señalado un conjunto de aspectos negativos de este tipo de formato, tales como el hecho de que no son capaces de capturar habilidades complejas, pues su nivel de exigencia de procesamiento cognitivo es pobre; inhiben la expresión de la creatividad y la demostración de un pensamiento innovador y original por parte del examinado, al no enfrentarlos con tareas realmente similares a las que encontrará en la vida real y son limitados para proporcionar información diagnóstica, ya que no permiten conocer los procesos y estrategias cognitivas que usan los examinados para producir una respuesta (Martínez, 1999; Osterlind, 1998).

Como respuesta a esta demanda de otra forma de evaluar conocimientos y habilidades cognitivas se considera el uso de pruebas con preguntas elaboradas en formatos de respuesta construida, ya sean de respuesta abierta u otras formas de evaluación en las que el sujeto demuestre habilidades y destrezas mediante la elaboración de tareas más cercanas a las situaciones de la vida real, conocidas como preguntas de evaluación de la actuación o con el término anglosajón de *performance assessment*.

Las preguntas con formato de respuesta abierta exigen la elaboración de la respuesta, ya sea esta corta o larga, a través de la interpretación, análisis y utilización de la información suministrada por el ítem, la cual debe ser presentada a continuación de la pregunta o estímulo. Aun cuando se dice que la demanda en cuanto a procesos cognitivos involucrados para llegar a la respuesta es superior que la que implica la selección de una respuesta entre varias, adolecen de prácticamente las mismas críticas que los ítems de elección múltiple por cuanto también se les considera diferentes de lo que podría ser un problema real.

Las preguntas de evaluación de la actuación o *performance assessment*, por su parte, se refieren a la ejecución de tareas prácticas que involucran instrumentos y equipos, o a la creación de un producto, todo como medio para evaluar contenidos, comprensión de procedimientos y habilidad del examinado para usar esos conocimientos para razonar y enfrentarse a problemas reales (Harmon et al, 1997).Este formato de preguntas resulta atractivo no sólo porque parece capturar habilidades complejas sino también porque permite ver las diferentes estrategias usadas por un sujeto para lograr una respuesta, lo cual posibilita establecer diferencias cualitativas entre los evaluados y, por otro lado, ayuda a obtener información para fines de diagnóstico. No obstante, también tiene sus detractores, que basan sus argumentos en deficiencias tales como que la escala para calificar las pruebas o productos no tiene una objetividad garantizada; la mayor demanda de tiempo no permite cubrir adecuadamente el dominio a evaluar, repercutiendo en la validez y limitando la generalizabilidad de las inferencias; además, su aplicación resulta muy costosa y tiende a centrarse más en el producto o proceso que en el dominio de interés (Martínez, 1999; Messick, 1998; Mumford et al, 1998) .

Son numerosos los estudios que han examinado el funcionamiento de estos tipos de formato y su contribución a la medición en función de sus propiedades psicométricas (Rodríguez, 2003;Traub, 1993). En estos estudios usualmente se presenta una comparación de los parámetros estimados para cada pregunta o de las puntuaciones obtenidas por cada sujeto en cada formato cuando las pruebas miden el mismo dominio cognitivo.

Martínez (1991) comparó las estimaciones de los parámetros de los ítems en dos formas equivalentes de un mismo test con formato de respuesta abierta y de elección múltiple, concluyendo que para sus datos los ítems de respuesta abierta tenían mayor dificultad y discriminación que sus contrapartes de elección múltiple. Sin embargo los resultados encontrados por Bridgeman (1992) no son tan decisivos. Por ejemplo, aunque a nivel de los ítems se encontraba mucha diferencia en cuanto a dificultad, las estimaciones de los parámetros de los sujetos con los distintos formatos eran muy similares; por otro lado, las curvas características de los ítems en algunos casos se superponían y en otros

eran muy distintas. Mas adelante Bridgeman y Rock (1993) comparan estos formatos a través de un análisis factorial confirmatorio (AFC) y encuentran que los ítems de respuesta abierta no parecen medir ninguna nueva dimensión, pero advierten que este resultado no debe ser generalizado.

El estudio de la precisión de las preguntas en función de su formato es usualmente realizado comparando el coeficiente alfa (e.g. Ackerman& Smith, 1988; Ayala, Yin, Shultz, &Shavelson, 2002; Bridgeman& Rock, 1993; Manhart, 1996; Martinez, 1991) o a través de las funciones de información (e.g., Jodoin, 2003; Lukhele, Thissen, &Wainer, 1994). Aunque los resultados obtenidos sugieren que la mayor precisión se obtiene con las preguntas o ítems planteados en formato de elección, los valores obtenidos para el coeficiente alfa en las preguntas de respuesta construida es variable, llegando en algunos casos a presentar mayor confiabilidad las pruebas de respuesta abierta que las de elección múltiple.

La equivalencia del constructo medido con diferentes formatos ha sido el foco de los esfuerzos de la investigación en este campo. La mayoría de los estudios han empleado modelos factoriales confirmatorios, usando diferentes enfoques. Algunos estudios han examinado la estructura factorial teórica del constructo evaluado por medio de análisis factoriales confirmatorios, el cual revela si el constructo es o no función del formato de prueba usado para medirlo (e.g., Ackerman& Smith, 1988; Ayala, Yin, Shultz, &Shavelson., 2002; Bridgeman& Rock, 1993; Hancock, 1994). En otros estudios se ejecuta una comparación explícita entre modelos: un modelo de un factor es comparado con un modelo de dos o más factores correlacionados, dependiendo del número de formatos de ítem incluidos en la investigación (e.g., Bennet, Rock, & Wang, 1991; Manhart, 1996).

Los resultados obtenidos en estas investigaciones no proveen evidencia concluyente acerca de la equivalencia psicométrica de las puntuaciones obtenidas con los distintos formatos. Mientras que en algunas investigaciones los hallazgos indican que independientemente del formato las pruebas pueden medir el mismo dominio cognitivo o constructo (e.g., Bennet, Rock, & Wang, 1991; Hancock, 1994), en

otros estudios los resultados sugieren que no es así, que cada formato puede medir aspectos distintos de la variable o característica evaluada (e.g., Ackerman y Smith, 1988; Manhart, 1996). Además, la mayoría de los estudios realizados comparan respuesta abierta y elección múltiple y son muy pocos los estudios donde se contrastan con otras formas de evaluación.

En esta investigación se comparan tres formatos de pregunta: elección múltiple (EM), respuesta abierta (RA) y evaluación de la actuación (EA). El propósito de la investigación es examinar la equivalencia psicométrica de las puntuaciones obtenidas con pruebas compuestas por ítems de cada uno de los formatos estudiados. Para ello, se investiga si las puntuaciones obtenidas con las diferentes pruebas son igualmente validas y confiables utilizando una variedad de estrategias que nos permitan reunir evidencia acerca de su validez y confiabilidad.

MÉTODO

Sujetos

Para llevar a cabo esta investigación se utilizaron los datos provenientes de un estudio internacional conocido como TIMSS (las siglas vienen de Third International Mathematics and Science Study) en su edición del año 1994, en el cual se compara el rendimiento de los estudiantes de más de 40 países en las áreas de matemáticas y de ciencias. Para este fin, los estudiantes de estos países (de los cuales lamentablemente Venezuela no forma parte, pues hace bastantes años que el país no se mide en estudios internacionales) se sometían a pruebas diseñadas con formatos de preguntas de selección y de respuesta abierta corta. Adicionalmente, un pequeño grupo de países aplicó otra prueba en la cual los estudiantes realizaban experimentos y que se identificó como *performance assessment.*

En este estudio se seleccionaron 3116 estudiantes pertenecientes a cinco países (Canadá, Nueva Zelanda, Escocia, España y Suiza) que aplicaron la prueba de *performance assessment* y de este modo poder comparar tres formatos de pregunta. Es necesario señalar que la selección de estos países obedeció al hecho de que trabajaron con

idénticas especificaciones para la selección muestral y de que poseen un promedio en el área de ciencias (área seleccionada para esta investigación) bastante similar, todo esto de manera de garantizar la consistencia de la muestra obtenida.

Instrumentos

Los instrumentos utilizados para recoger la información son el cuestionario de rendimiento y la prueba de *performance assessment* administrados en el TIMSS.

El cuestionario de rendimiento, en su sección de ciencias consta de 102 preguntas de selección y 33 de respuesta abierta, distribuidas en ocho formas alternativas equivalentes del cuestionario de rendimiento (con una cobertura equivalente de los distintos contenidos a evaluar y dificultad aproximadamente igual). En este estudio se consideraron 92 preguntas de EM y 20 de RA dado que sólo esas fueron respondidas por los grupos a quienes se les aplicó la prueba de EA.

En la prueba de evaluación de la actuación los estudiantes diseñaron experimentos, manipularon materiales, probaron hipótesis, registraron resultados y elaboraron conclusiones cuando ejecutaban un conjunto de tareas de matemáticas y ciencias. La selección de las tareas que conformarían la prueba de evaluación de la actuación estuvo basada, entre otras, en cuestiones como la adecuada cobertura de contenidos, de acuerdo a la estructura curricular de los países participantes, la mayor cobertura posible de aspectos que posibilitaran la evaluación de destrezas, la viabilidad en la obtención de los materiales y administración de las tareas, el tiempo límite para completar la tarea y el nivel de dificultad. Cada tarea se evaluaba mediante un conjunto de ítems que los estudiantes respondían en forma escrita y que eran calificados mediante un sistema de códigos similar al utilizado para las respuestas abiertas. Cada grupo de tareas requería alrededor de 30 minutos para su ejecución y a cada estudiante se le asignaba una secuencia y una rotación particular de grupos cuya combinación determinaba qué tareas debería realizar.

El diseño de recogida de datos utilizado fue el de muestras matriciales múltiples, utilizando para ellos las distintas formas o cuadernillos en que se distribuyen las preguntas. Por consiguiente, la matriz final de datos se caracteriza por la presencia de numerosos datos faltantes, que se hace aún mayor al unir los archivos de datos correspondientes al cuestionario de rendimiento y al de la prueba de evaluación de la actuación. Este hecho va a determinar de forma muy directa la manera de proceder al realizar los subsiguientes análisis en el presente trabajo.

Análisis

Para el estudio de la equivalencia psicométrica de las puntuaciones obtenidas con los ítems de distinto formato (EM, RA y EA) se utilizaron variadas estrategias tanto para reunir evidencias acerca de su validez como para estimar la confiabilidad según cada tipo de formato. Siempre que fue posible los análisis se realizaron trabajando desde la óptica de la Teoría Clásica de los Tests (TCT) así como desde la Teoría de Respuesta al Ítem (TRI).

Estimación de los parámetros desde la TRI

Para las preguntas de EM y RA se utilizaron los parámetros estimados en el TIMSS publicados en el reporte técnico. Para la estimación de los parámetros de las preguntas de EA, cuyos valores no fueron reportados, se utilizó el mismo procedimiento que utilizaran en el TIMSS para obtener los parámetros de las preguntas en formato EM y RA, es decir, se empleó el modelo logia multinomial de coeficientes aleatorios de Adams, Wilson y Wang (1997) con el software ConQuest. Para obtener las estimaciones de los parámetros de estos ítems en la misma escala que las de los ítems de EM y RA –la escala base del TIMSS- se introdujo en el proceso de calibración las respuestas a los ítems del grupo A (los ítems con mejores propiedades del estudio), fijando el valor de los parámetros de esos ítems al de las estimaciones ya disponibles en la escala base. Por último, antes de obtener la estimación del rendimiento de los sujetos en la prueba de evaluación de la actuación, se evaluó el ajuste de los datos al modelo, examinando el índice proporcionado por el programa y revisando la invarianza de las estimaciones de los parámetros de los ítems.

Validez

Las líneas de obtención de evidencia acerca de la validez de las puntuaciones obtenidas con los distintos formatos empleados se orientan a examinar la equivalencia del rasgo medido, la dimensionalidad de las pruebas formadas por preguntas de distinto formato y la evidencia convergente y discriminante.

La *equivalencia del constructo medido* se estudió a través del examen de los coeficientes de correlación desatenuada entre las estimaciones del rendimiento de los estudiantes obtenidos con los distintos formatos de pregunta.

La evaluación de la *dimensionalidad* se llevó a cabo utilizando una estrategia de comparación de modelos desde la óptica de los modelos de estructuras de covarianza. El primero es un modelo de un factor donde se considera que todos los ítems, cualquiera sea su formato (EM, RA o EA) miden un único constructo, que en este caso sería el rendimiento en ciencias. En el segundo modelo se establecen tres factores asociados a cada tipo de formato y se considera que cada uno de ellos mide una dimensión diferente del constructo.

La estimación de los parámetros y de los índices de ajuste de ambos modelos se realizó con el software LISREL, liberando las correlaciones entre los errores de medida de las variables con ítems comunes.

La *evidencia convergente y discriminante* se examinó utilizando una matriz multirasgo-multimétodo (MMRMM) dentro del marco de los modelos de estructura de covarianza. Los métodos están representados en el presente estudio por los distintos formatos de pregunta considerados (EM, RA y EA); el rasgo de interés central es el rendimiento en ciencias y el rasgo adicional el rendimiento en matemáticas, que constituyó la otra área evaluada en el TIMSS. Para obtener las estimaciones del rendimiento en matemáticas con cada uno de los formatos estudiados se procedió de la misma forma que para obtener el rendimiento en ciencias. El software utilizado para la estimación fue el LISREL y se trabajó con el modelo de unicidades correlacionadas (Kenny y Kashy, 1992;

Marsh, 1989), dado que el modelo más clásico –el análisis factorial confirmatorio con factores de rasgo y de método- no está identificado al trabajar en este caso con sólo dos rasgos.

Confiabilidad

Para la obtención de medidas de confiabilidad se utilizó el coeficiente alfa para el análisis desde la TCT y la función de información para el estudio desde la TRI.

La utilización del diseño de muestras matriciales múltiples en el estudio hacía imposible estimar un único valor del coeficiente alfa por formato de pregunta, ya que eran muchos los ítems que por diseño no habían sido administrados a los sujetos. Por ello, se calculó el coeficiente alfa separadamente para los ítems de elección múltiple y respuesta abierta incluidos en cada una de las siete formas equivalentes del cuestionario de rendimiento y para los ítems de evaluación de la actuación se procedió de modo semejante, considerando en este caso el número de secuencia y rotación de las tareas que le correspondió al estudiante examinado.

Seguidamente, dado que la confiabilidad clásica depende de la longitud del test y dado que ésta variaba notablemente de un formato a otro de ítems, se procedió como sigue para poder comparar la confiabilidad de los ítems de distinto formato. Se optó por trabajar con pruebas cuya resolución llevara alrededor de media hora de trabajo. Conociendo a través del reporte técnico que el tiempo estimado para responder las preguntas del TIMSS es aproximadamente de 1 minuto para EM, 2 minutos para RA corta, 5 minutos para RA larga, 15 minutos para tareas cortas y 30 minutos para tareas largas, se redefinió la prueba de EM con 30 preguntas, la de RA con 10 preguntas y la de EA con 8 preguntas y se obtuvieron los correspondientes coeficientes de confiabilidad corregidos, aplicando la ecuación de Spearman-Brown con los coeficientes anteriormente calculados. A continuación se promediaron según formato los coeficientes corregidos y se logró tener una estimación de la confiabilidad para cada formato estudiado.

Al abordar la confiabilidad desde la TRI, hubo que recurrir al programa MULTILOG ya que el ConQuest no estima las funciones de información. Para comparar la cantidad de información proporcionada por los ítems de cada formato se realizó un planteamiento totalmente paralelo al seguido al estimar la fiabilidad desde la TCT: se trabajó con subtests constituidos por ítems de idéntico formato cuyo tiempo de realización estuviera en torno a los 30 minutos.

RESULTADOS

Evidencias en relación a la validez

Equivalencia del constructo medido

Los resultados obtenidos al analizar la correlación entre las estimaciones del rendimiento obtenidas con los distintos formatos de pregunta muestran la falta de equivalencia del constructo medido. Los coeficientes de correlación obtenidos luego de eliminar el error, es decir, corregidos por desatenuación, tienen valores que se alejan de la unidad ($\hat{\rho}_{EMRA} = 0{,}696$; $\hat{\rho}_{EMEA} = 0{,}533$; $\hat{\rho}_{RAEA} = 0{,}335$), sugiriendo que las puntuaciones obtenidas no representan el mismo constructo.

Dimensionalidad

Las cargas factoriales para el modelo de 1 factor y el modelo de tres factores (véase Tabla 1) muestran valores superiores en las saturaciones para el modelo de tres factores. Además, los índices de ajuste global apoyan el modelo de tres factores($\chi^2 = 63.269$, $p = 0.002$; RMSEA = 0.035; GFI = 0.979)mientras que los índices obtenidos para el modelo de un factorfit($\chi^2 = 227.545$, $p < 0.0005$; RMSEA = 0.224; GFI = 0.687) evidencian un ajuste deficiente. Todo esto indica que la matriz de varianza-covarianza está mejor reproducida por el modelo de tres factores y en consecuencia, el constructo no parece ser unidimensional.

Tabla 1
Cargas factoriales para los modelos de 1 y 3 factores

Variables	Modelo de 1 factor	Modelo de 3 factores		
EM1	0.650	0.822	-	-
EM2	0.650	0.830	-	-
EM3	0.538	0.735	-	-
EM4	0.665	0.847	-	-
EM5	0.531	0.727	-	-
EM6	0.537	0.727	-	-
RA1	0.269	-	0.776	-
RA2	0.367	-	0.967	-
RA3	0.341	-	0.877	-
EA1	0.885	-	-	0.956
EA2	0.846	-	-	0.938
EA3	0.891	-	-	0.989

Evidencia convergente y discriminante

La inspección de la MMRMM (véase Tabla 2), con los criterios expuestos originalmente por Campbell y Fiske (1959), no provee evidencia convergente de validez, pues los coeficientes monorasgo-heterométodo (en negrita en la tabla) no tienen valores muy altos (desde 0.119 a 0.529), indicio de que los distintos formatos no miden el mismo constructo. Por otra parte se observa una alta correlación entre los rasgos cuando son evaluados con el formato de elección múltiple, lo que en principio podría ser considerado como una evidencia discriminante poco satisfactoria.

Tabla 2
Matriz multirasgo-multimétodo

	EM_{CS}	RA_{CS}	EA_{CS}	EM_{MA}	RA_{MA}	EA_{MA}
EM_{CS}	-					
RA_{CS}	**0.383 (0.316)**[a]	-				
EA_{CS}	**0.305 (0.360)**	**0.142 (0.159)**	-			
EM_{MA}	0.563 (0.597)	0.284 (0.315)	0.276 (0.335)	-		
RA_{MA}	0.244 (0.392)	-0.055 (0.230)	0.176 (0.201)	**0.342 (0.529)**	-	
EA_{MA}	0.273 (0.277)	0.085 (0.073)	0.055 (0.156)	**0.386 (0.371)**	**0.119 (0.249)**	-

[a]Los valores entre paréntesis corresponden a los resultados con las estimaciones obtenidas desde la TRI.

Estos hallazgos son reforzados por los resultados del análisis realizado con el modelo de unicidades correlacionadas, que indican que el modelo presentó un buen ajuste a los datos, con unos índices de bondad del ajuste con valores adecuados, tanto para el análisis realizado con los datos provenientes de la TCT ($\chi^2 = 6.78$, p = 0.238; RMSEA = 0.0258; GFI = 0.998) como para los datos obtenidos desde la TRI ($\chi^2 = 17.57$, p = 0.004; RMSEA = 0.0312; GFI = 0.999). Además, al examinar la matriz de pesos factoriales se observa que, aunque todas las saturaciones son significativas, las cargas factoriales para las estimaciones del rendimiento tanto en ciencias como en matemáticas con los formatos RA y EA no son suficientemente altas, lo cual, como ya se adelantó al estudiar directamente la MMRMM, cuestiona la validez y puede implicar que el constructo no es unidimensional (véase Tabla 3).

Tabla3
Matriz de pesos factoriales

	TCT		TRI	
Variable	Ciencias	Matemáticas	Ciencias	Matemáticas
EM_{CS}	0.898	-	0.843	-
RA_{CS}	0.399	-	0.375	-
EA_{CS}	0.366	-	0.427	-
EM_{MA}	-	0.963	-	0.950
RA_{MA}	-	0.363	-	0.567
EA_{MA}	-	0.385	-	0.393

Por último, dado que las correlaciones entre las unicidades de las medidas de distinto rasgo tomadas con el mismo formato tienen valores muy bajos (de -0.167 a -0.054), no parece haber apenas varianza atribuible al formato y las diferencias observadas pueden provenir más bien de la no equivalencia del constructo medido por cada formato.

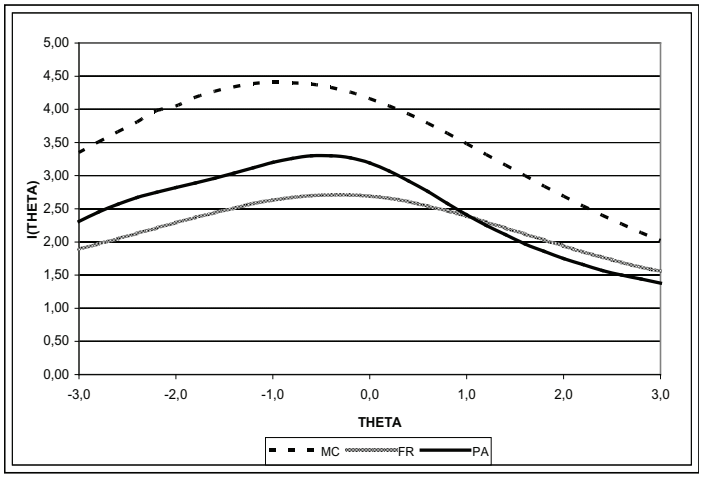
Evidencias en relación a la confiabilidad

Los valores promedio de los coeficientes alfa obtenidos para cada formato estudiado y corregidos con la ecuación de Spearman-Brown para pruebas de una duración de 30 minutos son de 0.709 para EM, 0.536 para RA y 0.591 para EA.

Estos valores permiten afirmar que para pruebas de una duración similar (30 minutos) se obtendrá una mayor fiabilidad con el formato de elección múltiple y que los ítems de respuesta abierta son los que reportan una menor fiabilidad, si bien el valor obtenido es bastante próximo al del formato de evaluación de la actuación.

Los resultados obtenidos al trabajar desde la óptica de la TRI son totalmente consistentes con la evidencia proporcionada por el coeficiente alfa, tal como puede verse en la Figura 1. La figura presenta la información promedio proporcionada a lo largo de la escala de rendimiento por cada tipo de formato de ítems, calculada a partir de subtests de aproximadamente media hora de duración. Tal como se observa en esta Figura 1, los subtests formados por ítems con formato de EM (MC en la leyenda de la figura) resultan más informativos que los de RA (FR en la leyenda de la figura) o EA (PA en la leyenda de la figura) para cualquier nivel de rendimiento. Se observa también que la información proporcionada por los subtests de EA está por encima de la de los subtests de RA para un amplio rango de niveles de rendimiento: el formato de RA es ligeramente más informativo que el de EA sólo para valores theta por encima de 1. El rango de theta para el cual los tests son más informativos está aproximadamente entre -1 y 0 para cualquier formato.

Figura 1. Curva con la función de información para cada formato



DISCUSIÓN

El conjunto de los análisis realizados apuntan hacia la conclusión de que los formatos de pregunta estudiados tienen características psicométricas distintas. Para aproximarse a esta conclusión se utilizaron variadas estrategias y enfoques a fin de tratar de superar debilidades de los procedimientos y observar posibles inconsistencias en los resultados según la perspectiva teórica empleada, lo cual permite una mayor confianza en los hallazgos de la investigación.

En el análisis de las correlaciones desatenuadas se pone de manifiesto que el constructo medido con los distintos formatos no es totalmente equivalente. Estos resultados son reforzados por el estudio de dimensionalidad de las pruebas a través del análisis factorial confirmatorio, donde se observa que el modelo de tres factores presenta un mejor ajuste que el de un factor, siendo mayores las cargas factoriales de las variables asociadas a un determinado formato en el modelo de tres factores; estos resultados coinciden con los hallazgos de Manhart (1996) y Ackerman y Smith (1988). Adicionalmente, el análisis de la MMRMM apunta en el mismo sentido, es decir, no apoya la existencia de una estructura unidimensional de los datos reflejando más bien que cada tipo de formato evalúa aspectos distintos del rendimiento en ciencias y revela, además, una fuerte relación entre el rendimiento en ciencias y en matemáticas. Estas evidencias en torno a la validez pueden ser interpretadas como que cada formato mide una dimensión diferente del rendimiento en ciencias, lo que podría ser explicado ya sea por la distribución de los dominios de contenido específicos entre los distintos formatos o porque la demanda de procesamiento cognitivo difiere según el formato del ítem.

La inspección de la matriz de correlaciones entre unicidades, obtenida al analizar la MMRMM revela una diferencia importante en el error de medida asociado a cada formato: se observa que la precisión de las puntuaciones obtenidas con el formato EM es sensiblemente mayor que la de los formatos RA y EA. Se puede aducir que el número de ítems de EM era considerablemente mayor que el de RA y de EA y que eso podría explicar la mayor confiabilidad de las medidas obtenidas con el formato EM. Sin embargo, la estrategia utilizada en este estudio para

comparar la fiabilidad de las medidas proporcionadas por los distintos formatos fue trabajar con unidades similares en tiempo -con subtests de media hora de duración para cada formato- y comparar la confiabilidad de estos subtests examinando el coeficiente alfa y sus correspondientes funciones de información. Los resultados obtenidos concuerdan con lo anterior y permiten concluir que los distintos formatos de ítems dan medidas con distintos niveles de precisión, con valores notablemente superiores para los ítems de EM y con los valores más bajos para los de RA en pruebas de 30 minutos de duración. Estos resultados confirman lo obtenido en otras investigaciones (Jodoin, 2003; Luckele, Thissen y Wainer, 1994; Wainer y Thissen, 1993) en las que se concluye que el formato EM produce medidas de mayor precisión.

Los resultados obtenidos dejan poco margen para dudar de la falta de equivalencia del constructo cuando es medido con preguntas que tienen distinto tipo de formato. Por una parte, los datos provienen de un programa de alto perfil cuya rigurosidad está fuera de duda. Por otra parte, las evidencias fueron reunidas usando un variado tipo de perspectivas y técnicas, las cuales apuntan en su totalidad en la misma dirección, siempre dando lugar a resultados consistentes.

A la vista de estos resultados, cabría plantearse que, si los distintos formatos de las preguntas tienen propiedades métricas distintas, lo ideal sería utilizar el formato (o los formatos, como es el caso de TIMSS) que permita medir la variable de interés del mejor modo posible. Para producir inferencias válidas acerca de un examinado el examen aplicado debe reflejar adecuadamente el área de conocimiento a evaluar; por lo tanto, es importante que el conjunto de preguntas que conforman la prueba puedan capturar el espectro completo de contenidos y demandas cognitivas necesarias para medir ese conocimiento, habilidad o destreza, lo que no siempre parece ser logrado a través del uso de un solo tipo de formato. Ahora bien, la toma de decisiones en cuanto al formato muchas veces está más marcada por otras consideraciones tales como el costo, el tiempo y los recursos disponibles para realizar la aplicación de la prueba y su posterior corrección.

Las pruebas, como se mencionó ya, son una herramienta muy útil para tomar decisiones en cuanto a cupos en la universidad o en

relación a puestos de trabajo. Es obvio que si los puestos son limitados, es necesario una selección y una prueba de calidad es la mejor forma de obtener una evaluación objetiva de los conocimientos, destrezas o habilidades de una persona en determinado dominio académico o laboral. De la misma manera, los resultados de encuestas sobre preferencias electorales tienen una importancia suprema, pues es sabido que estos pueden afectar la decisión sobre el voto, fundamental en la vida democrática de las sociedades. Asimismo, es posible afirmar que los resultados de las macroencuestas socioeconómicas pueden orientar decisiones sobre políticas y programas sociales. En consecuencia, antes que prescindir de los cuestionarios, es preferible profundizar en la investigación que permita la optimización de los mismos.

Si se avanza en el estudio de la contribución a la medida que puede hacer cada uno de estos tipos de formatos de pregunta, será más sencillo determinar cuál sería el más adecuado para medir el dominio de interés con un determinado objetivo y para un grupo de sujetos y contexto de aplicación particulares. Esto debe contribuir a obtener evaluaciones que tengan una confiabilidad adecuada y que aporten el mayor número de evidencias que respalden su validez y de esta manera orientar la toma de decisiones basados en evidencia objetiva, replicable y con credibilidad tal que quede fuera de toda la duda que la selección ha sido justa e imparcial.

Referencias Bibliográficas

Ackerman, T. A. y Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice and free response. *AppliedPsychologicalMeassurement*, *12*, 117 - 128.

Adams, R. J., Wilson, M. y Wang, W. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Meassurement, 21*, 1 - 23.

Ayala, C. C., Yin, Y. y Shultz, S. (2002). *On science achievement from the perspective of different types of test: a multidimensional approach to achievement validation* (CSE Technical Report 572).California: CRESST/Stanford University.

Bennet, R. E., Rock, D. y Wang, M. (1991).Equivalence of free-response and multiple-choice items.*Journal of Educational Measurement, 28*, 77 – 92.

Bridgeman, B. (1992). A comparison of quantitative question in open-ended and multiple choice formats.*Journal of Educational Measurement*, *29*, 253 – 271.

Bridgeman, B. y Rock, D. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30*, 313 – 329.

Campbell, D. T. y Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix.*Psychological Bulletin, 56*, 81-105.

Hancock, G. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test format.*Journal of Experimental Education, 62*, 143- 157.

Harmon, M., Smith, T., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. et al. (1997).*Performance Assessment in IEA's Third International Mathematics and Science Study (TIMSS).*TIMSS International Study Center. Chestnut Hill, MA: Boston College.

Jodoin, M. (2003).Measurement efficiency of innovative item format in computer-based testing. *Journal of Educational Measurement, 40*, 1 - 15.

Kenny, D. y Kashy D. (1992).Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychologhical Bulletin, 112*, 165-172.

Lukhele, R., Thissen, D. y Wainer, H. (1994). On the relative value of multiple-choice, constructed response and examinee-selected items on two achievement test. *Journal of Educational Measurement, 31*, 234-250.

Manhart, J. J. (1996, Abril).*Factor analytic methods for determining whether multiple-choice and constructed response tests measure the same construct*. Documento presentado en la reunión anual del National Council on Measurement in Education, New York.

Marsh, H. (1989). Confirmatory factor analysis of multitrait-multimethod data: many problems and a few solutions. *Applied Psychological Measurement, 13*, 335-361.

Martinez, M. (1991).A comparison of multiple-choice and constructed figural response item. *Journal of Educational Measurement, 28*, 131-145.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychological, 34*, 204 – 218.

Messick, S. (1998).Alternative modes of assessment, uniform standards of validity.En M. Hakel (Ed.) *Beyond multiple choice: evaluating alternatives to traditional testing for selection*(pp. 59-74). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mumford, M. D., Baughman, W. A., Supinski, E. P. y Anderson, L. E. (1998). A construct approach to skill assessment: procedures for assessing complex cognitive skills. En M. Hakel (Ed.), *Beyond multiple choice: evaluating alternatives to traditional testing for selection* (pp. 75-112).Hillsdale, NJ: Lawrence ErlbaumAssociates, Inc.

Osterlind, S. J. (1998). *Constructing test items: multiple-choice, constructed-response, performance, and others formats* (2ª ed.). Boston: Kluwer Academic Publishers.

Rodriguez, M. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *Journal of Educational Measurement, 40*, 163-184.

Ryan, A. M. and Greguras G. (1998). Life is not multiple choice: reactions to the alternatives. En M. Hakel (Ed.), *Beyond multiple choice: evaluating alternatives to traditional testing for selection*(pp. 183- 202). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Traub, R. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. En: R. E. Bennett y W. Ward. (Eds.), *Construction versus Choice in Cognitive Measurement: issues in constructed response, performance testing and portfolio assessment* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H. y Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: toward a marxist theory of test construction. *Applied Measurement in Education, 6*, 103 – 118.