¿MIDEN LO MISMO DISTINTOS FORMATOS DE PREGUNTA? UNA REVISIÓN SOBRE EL TEMA.

Elsy J. Urdaneta Durán
Elsy.urdaneta.d@gmail.com
Universidad de Los Andes, Núcleo Rafael Rangel. Trujillo,Venezuela

Recibido: 5/12/2012 **Aceptado**: 24/4/2013

Resumen

Los test o pruebas son los instrumentos empleados para la medición de variables del ámbito psicológico y educativo y los resultados de estos son utilizados para la toma de importantes decisiones que afectarán muy directamente la vida de los examinados. Esta es la razón por la que su estudio tiene una gran relevancia en la investigación científica. En tal sentido, este artículo recoge los hallazgos de los estudios más importantes acerca del formato de la pregunta en pruebas de desempeño y la influencia que tienen sobre la ejecución de la tarea, al investigar si los distintos formatos de ítems logran medir lo mismo o, si por el contrario, tienen un efecto diferencial sobre la conducta generadora de competencias del examinado al momento de afrontar una prueba.
Palabras clave: formato del ítem, elección múltiple, respuesta construida.

DO DIFFERENT ITEM FORMATS MEASURE THE SAME CONSTRUCT? A REVIEW OF THE SUBJETC

Abstract

The tests are the instruments used for the measurement of variables of psychological and educational field and the results of these are used for making very important decisions that will affect directly the lives of the examined. This is the reason that their study is of great significance in scientific research. As such, this article presents the findings of the most important studies about the item format in performance tests and their influence on the execution of the task, to investigate whether the different item formats measure achieve the same or, on the contrary, have a differential effect on the behavior of the considered generating skills when facing a test.
Keywords: item format, multiple choice, constructed response.

Los test como instrumentos de medición de constructos psicológicos y dominios académicos

Considerando que los métodos de la ciencia deben involucrar el uso de equipos de alta precisión, escrupulosos controles y procedimientos estandarizados para garantizar la objetividad y comparabilidad de las medidas obtenidas, se admite la utilización de los tests como instrumentos de medición. Si bien son diferentes de los instrumentos de medición utilizados en las ciencias naturales, se han desarrollado importantes teorías que justifican y avalan su aplicación, sobre la base de la interpretación que puede hacerse de las puntuaciones obtenidas con ellos.

La diversidad de características psicológicas y educativas medidas mediante tests es impresionante. Las exigencias de la sociedad, que requiere clasificar, diagnosticar, describir y evaluar, propician la elaboración y uso creciente de tales instrumentos y demandan investigación y

estudio cada vez más detallado. La toma de decisiones basadas en las puntuaciones obtenidas con un test o prueba, al afectar de modo directo la vida de las personas a las que se les aplican, representa un asunto muy delicado, importante y de gran sensibilidad social, razón suficiente para que la construcción de los mismos sea considerada como un área de estudio con una alta exigencia de rigor metodológico a fin de cumplir con los requerimientos, tanto científicos como sociales, que le demanda su papel en el mundo.

La satisfacción de estas exigencias quedará plasmada en la elaboración de cada uno de los elementos que componen una prueba y en todos los factores relacionados con su aplicación. Pero particularmente el ítem, su contenido y su forma, es el que debe generar la conducta que representa el constructo que busca ser medido; es decir, el que debe poner en marcha todos los mecanismos cognitivos y no cognitivos que posee el examinado y que le permitirán responder a la pregunta.

Tipos de ítem o pregunta según su formato

El ítem, tal y como lo define Osterlind (1998), es una unidad de medida con un estímulo y una forma preceptiva de responder para producir una respuesta del examinado, a partir de la cual puede ser estimado su nivel en el constructo psicológico medido, sea este conocimientos, habilidades, destrezas o rasgos de personalidad. El formato del ítem refiere a su diseño y disposición, lo cual hace que su categorización pueda resultar muy diversa.

En términos generales, se suelen considerar dos categorías amplias de formatos de ítems: de respuesta seleccionada y de respuesta construida (Bennett y Ward, 1993; Downing, 2006; Osterlind, 1998). Los ítems de respuesta seleccionada generalmente están compuestos por un enunciado que representa un problema, plantea una pregunta o recoge una afirmación, seguida por un conjunto de alternativas de respuesta, en donde una o varias representan la respuesta correcta y el examinado debe seleccionar la respuesta que considere acertada. En contraste, los ítems de respuesta construida no suministran ninguna respuesta, sino que el examinado la debe elaborar a través de la interpretación, análisis y utilización de la información suministrada por el ítem y de su conocimiento y experiencia previa.

Formato de respuesta seleccionada

Los ítems de elección múltiple y los de verdadero o falso son ejemplos típicos del formato de respuesta seleccionada; el primero es ampliamente utilizado en las pruebas de evaluación educativa a gran escala por las ventajas que proporciona, entre las cuales se destacan: su bajo costo

operativo, su facilidad de calificarlas objetivamente, su mayor precisión y la posibilidad de contar con muestras representativas del dominio evaluado (Mumford, Baughman, Supinski y Anderson, 1998; Osterlind, 1998; Ryan y Greguras, 1998). No obstante, durante las últimas décadas este formato ha perdido exclusividad en las pruebas utilizadas en macro encuestas educativas, debido a algunas debilidades que se han detectado en el mismo. Por un lado, se afirma que es limitado para evaluar las habilidades de nivel cognitivo superior y las destrezas para el aprendizaje a lo largo de la vida; así como también para poder llevar a cabo la más que conveniente integración de la evaluación con la instrucción, tal como las sugieren las teorías del aprendizaje y la psicología cognitiva que se manejan en la actualidad. Por otro lado se señala que para medir destrezas psicomotoras puede resultar inútil y que para evaluar tareas complejas tales como la ejecución de instrumentos o destrezas en disciplinas artísticas, solamente el formato de respuesta construida es el adecuado (Downing, 2006; Martínez Arias, 2010). Además, las respuestas en este formato son muy fáciles de ser copiadas fraudulentamente o de ser respondidas al azar.

Formato de respuesta construida

El formato de respuesta construida adopta una gran variedad de formas, tales como completar oraciones, reordenar frases o secuencias, responder con una palabra o un número, emitir una respuesta argumentada, resolver problemas, hacer experimentos o hacer muestras de escritura. Los ítems con este formato pueden, a su vez, ser clasificados en dos grandes categorías representadas por los ítems de respuesta corta y por los ítems de respuesta extendida. Los ítems de respuesta corta sólo exigen como solución una palabra, una frase o un número, mientras que los ítems de respuesta extendida requieren ejecuciones más complejas tales como escritura de ensayos, solución de problemas, realización de experimentos científicos u otras modalidades de lo que se conoce con el término anglosajón de *performance assessment* o evaluación de la actuación. Al igual que los ítems de respuesta seleccionada los ítems de respuesta construida tienen virtudes y debilidades.

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, los ítems de respuesta corta adolecen de muchas de las críticas realizadas a los ítems de elección múltiple, por cuanto también se les considera diferentes de lo que podría ser un problema de la "vida real".

Por su parte, los ítems de respuesta extendida resultan bastante atractivos, no sólo porque parecen capturar habilidades complejas, sino también porque permiten la observación de los

procedimientos llevados a cabo por un sujeto para emitir una respuesta, lo cual posibilita establecer diferencias cualitativas entre los evaluados y ayuda a obtener información para fines de diagnóstico. Particularmente los ítems tipo evaluación de la actuación permiten la alineación de destrezas y habilidades con las competencias importantes para la vida en contextos realistas (Martínez Arias, 2010). No obstante, este formato también tiene sus detractores, que basan sus argumentos en deficiencias tales como que la escala para calificar las pruebas o productos plantea problemas de objetividad y que la mayor demanda de tiempo no permite cubrir adecuadamente el dominio a evaluar, repercutiendo en la validez y limitando la generalizabilidad de las inferencias (Bennett, 1993; Martínez, 1999; Messick, 1998; Mumford et al, 1998). Además, su aplicación y calificación resulta muy costosa y tiende a centrarse más en el producto o proceso que en el constructo de interés

Los estándares para la evaluación psicológica y educativa (AERA, APA y NCME, 1999) indican que el formato del ítem debe especificarse en función del propósito de la prueba. Por tanto, es necesario conocer a fondo cómo funciona cada uno de los formatos que puedan ser empleados en los procesos de medida de los constructos psicológicos y educativos.

Relación entre formato de la pregunta y ejecución de la tarea

La evaluación de los resultados obtenidos por los estudiantes representa una fase de suma importancia en el proceso de enseñanza-aprendizaje, de manera que ésta debe hacerse con los instrumentos más confiables y apropiados. La meta a lograr es elaborar un buen test y, considerando que los ítems son la unidad básica del mismo, la construcción de un buen test pasa por escribir buenos ítems. Un buen ítem será aquel que nos permita obtener una puntuación que dé cuenta del constructo que estamos interesados en medir. Para que esto sea posible, el conjunto de ítems que conforman el test deben abarcar todos los aspectos relevantes a ese constructo. En este sentido, el formato de la pregunta que se utiliza en los exámenes es un elemento de interés relevante, pues no está claro que todos los formatos permitan evaluar cualquier constructo o, cuando menos, que lo hagan con la misma adecuación; además, pueden poner en juego procesos cognitivos diferentes al requerir del sujeto conductas distintas para responder e, igualmente, pueden no resultar igual de motivadores para todos los sujetos. Es decir, se debe tener muy en cuenta el formato de ítem a utilizar, pues como bien afirma Rodríguez (2002), esta elección lleva implícitos aspectos relacionados con el tema de la validez.

Snow (1993) certeramente afirma que la equivalencia del rasgo o variable a medir establecida a través de criterios psicométricos no tiene por qué ser la misma que la establecida a

través de criterios psicológicos. En tal sentido, es necesario tratar de determinar con claridad el potencial de cada formato para medir un determinado constructo y examinar no sólo la equivalencia psicométrica sino también la equivalencia psicológica de las medidas obtenidas al administrar ítems de distinto formato.

La ejecución de una tarea podría estar relacionada con el contexto y la forma en como se presenta la misma. Un conjunto de experimentos, realizados unos desde la perspectiva cognitiva y otros desde la perspectiva diferencial, sugieren que las instrucciones del test, las condiciones de aplicación, el orden de los ítems, el formato de respuesta o el tiempo de aplicación pueden tener efectos significativos en las puntuaciones obtenidas (Embretson, 1993; Embretson y Reise, 2000; Gorin, 2006; Haertel y Wiley, 1993; Snow y Peterson, 1985).

Snow (1989) ha señalado que el patrón de influencia de las variables afectivas y conativas en el rendimiento académico varía en función de características situacionales, tal como podría ser el formato de respuesta en un escenario de evaluación. En el mismo orden de ideas, Pintrich y Schunk (2002) afirman que el tipo de tareas tiene una importante influencia en la motivación y en la cognición, de tal manera que es lógico considerar que la tarea en sí misma representa un factor que puede influir en el momento de responder a una pregunta o ítem.

Siendo así, es posible afirmar que el formato de la pregunta puede representar un factor decisivo de la actuación del sujeto frente a una tarea particular, al constituirse en elemento que puede estar asociado a diferencias individuales en las variables tanto cognitivas como no cognitivas, que se activan durante la ejecución (Martínez, 1999; Snow, 1989). En tal sentido, el estudio del formato refiere al asunto de la validez por cuanto existe la preocupación de examinar si distintos formatos son capaces de medir un constructo al mismo nivel de amplitud y profundidad.

En atención a lo anterior, se puede señalar que la elaboración de las pruebas académicas debe estar soportada por la explicación de los procesos y estructuras del conocimiento que se desea evaluar, pero también debe considerar aspectos de tipo volitivo y motivacional e incluso aspectos contextuales relacionados con la forma de administración, formato y organización del test (Snow, 1993; Nickerson, 1994).

Estudios sobre el formato del ítem y equivalencia del constructo medido

Esta es la razón por la cual la selección del formato de la pregunta ha motivado diversas investigaciones que intentan averiguar si miden lo mismo aquellas preguntas en las que el

examinado debe seleccionar una respuesta entre varias o aquellas en las que el estudiante debe elaborar su respuesta, ya sea ésta una simple palabra o frase, o una respuesta larga.

La revisión de estos trabajos conduce a pensar que, en general, los ítems de elección múltiple no miden lo mismo que los ítems de respuesta construida. Se presume una posible influencia del formato del ítem no sólo en lo que respecta a los factores cognitivos, sino también en lo relacionado con los aspectos motivacionales y afectivos que podrían ser mediados por la forma de presentación de la tarea y las estrategias metacognitivas que pueden ser utilizados según el formato en que se presente la pregunta.

Entre los trabajos que tratan acerca de las características, naturaleza y potencialidades de los distintos formatos del ítem se puede mencionar las de Downing y Haladyna (2006), Haladyna,(1997, 1999), Kane y Mitchell (1996), Osterlind (1998), así como también estudios acerca de la equivalencia del constructo cuando es medido con diferentes formatos. Estos últimos trabajos intentan averiguar si distintos formatos de pregunta son capaces de medir el mismo constructo o si por el contrario miden aspectos diferentes de ese constructo o dominio de interés.

Ackerman y Smith (1988) establecieron la estructura factorial teórica del constructo a medir y luego compararon los resultados de los análisis factoriales confirmatorios, a fin de observar el grado de correlación entre cada factor en función del formato del ítem, encontrando que al menos en escritura el constructo medido es función del formato del ítem, puesto que los puntajes obtenidos con los distintos métodos probados en su investigación (ensayo, respuesta abierta y elección múltiple) proporcionan diferente información cognitiva y concluyen que una medida más confiable de este constructo puede requerir el uso de varios formatos.

Ayala, Shavelson, Yin y Shultz (2002) realizaron un estudio con el propósito de determinar si pruebas de evaluación de la actuación podían ser explícitamente diseñadas para medir tres dimensiones del razonamiento en la ciencia (conocimiento básico, razonamiento mecánico espacial y razonamiento cuantitativo) y, además, examinar la consistencia entre tres medidas del rendimiento, obtenidas mediante ítems de elección múltiple, respuesta abierta y evaluación de la actuación. Encontraron una confiabilidad muy baja en los ítems de respuesta abierta y por lo tanto no los consideraron para el resto de sus análisis. También hallaron que los puntajes obtenidos con los ítems de elección múltiple y de evaluación de la actuación no convergen en las tres dimensiones hipotetizadas del razonamiento, obteniendo además correlaciones muy moderadas que sugieren que podrían medir aspectos diferentes del rasgo considerado.

Bennett, Rock y Wang (1991) evaluaron la equivalencia del constructo medido con ítems de elección múltiple y de respuesta abierta en el área de computación, a través de un modelo de dos factores correlacionados (el primer factor para los ítems de elección múltiple y el segundo factor para los ítems de respuesta abierta), encontrando que un único factor provee el ajuste más parsimonioso. Sin embargo, los autores advierten que la evidencia encontrada es limitada y los resultados no pueden ser generalizados.

Bridgeman (1992) comparó distintos formatos de respuesta en lo concerniente a dificultad de la pregunta, discriminación del ítem y estructura correlacional utilizando la teoría de respuesta al ítem y encontró resultados contradictorios. 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.

Gadalla (1999) comparó la puntuación total obtenida en exámenes con formatos de respuesta de elección múltiple y respuesta abierta aplicados al área de matemáticas a estudiantes de cinco grados consecutivos encontrando efectos significativos del formato sólo para los dos primeros grados.

Hancock (1992) investigó los niveles de complejidad, dentro del dominio estrictamente cognitivo, evaluados con los formatos de elección múltiple y respuesta abierta usando como marco de trabajo la taxonomía de Bloom; examinó el grado en que ambos formatos miden la misma habilidad cognitiva y halló altas correlaciones en todos los niveles taxonómicos entre ambos formatos de respuesta. Sin embargo, advierte que su estudio podría ser válido sólo para algunos objetivos de enseñanza y áreas de conocimiento.

Jodoin (2003) realizó una investigación con el objetivo de proveer evidencia empírica de la confiabilidad de dos tipos de ítem, el clásico de elección múltiple y un tipo nuevo de respuesta construida que se deseaba probar en el examen computarizado del *Microsoft Certified Systems Engineer*. La confiabilidad fue evaluada usando las funciones de información de los ítems y se pudo concluir que los ítems de formato novedoso proveen considerablemente más información que los ítems de elección múltiple para todos los niveles de habilidad. No obstante, responder estos ítems lleva más tiempo y en consecuencia proveen menos información por minuto que los de elección múltiple, lo que puede tener mucha importancia al momento de seleccionar el tipo de ítem más adecuado para una determinada evaluación.

Mislevy (1993) discute un marco analítico para evaluar la contribución de los diferentes tipos de ítem al proceso de enseñanza. El estudio se centra en lo que puede conocerse de un individuo a partir de las observaciones que se le hacen e intenta proponer un marco de trabajo a través de una especie de diagrama de flujo para orientar a los educadores sobre qué formato de ítem es mejor aplicar en función del objetivo de enseñanza y de las características del sujeto, sin tratar directamente sobre el debate elección versus construcción, sino más bien de establecer un modelo que permita seleccionar a priori el tipo de instrucción y formato de evaluación más adecuado para cada estudiante, escenario y competencias requeridas.

En un estudio sobre pruebas de acceso a la Universidad, Sternberg y colaboradores (2004) encontraron que los tests en formato de elección múltiple conformaron un factor, independientemente del objetivo de los mismos, lo que los llevó a plantear la importancia del estudio del formato. Los autores consideran que si el constructo medido con diferentes tipos de formatos no es equivalente, se debe analizar si el formato representa una fuente de varianza irrelevante al constructo (varianza método) o si verdaderamente los ítems elaborados en formatos diferentes están midiendo cosas distintas representativas del mismo constructo.

Urdaneta (en prensa) comparó la validez y la confiabilidad de tres formatos distintos de pregunta: elección múltiple, respuesta corta y desarrollo. Para el estudio de la validez se reunió evidencia en relación con 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 cuanto a la validez denotan que cada formato mide una dimensión distinta del constructo y en relación con la confiabilidad se concluye que para pruebas de igual tiempo de duración se obtienen puntuaciones de mayor precisión con el formato de elección múltiple.

También pueden hallarse investigaciones como las de Hogan (1981), Traub (1993), Martínez (1999) o Rodríguez (2002, 2003), que recogen los resultados de variados estudios empíricos sobre formatos de ítem. En todos estos trabajos se señala la importancia del estudio del formato en la determinación de las diversas fuentes de variación, la existencia de diferencias y, muy importante, que la elección del formato del ítem está estrechamente vinculado con la representación del

constructo a medir, en consecuencia, es un asunto crucial para garantizar la validez de las inferencias.

Hogan (1981) realizó una amplísima revisión de investigaciones sobre estas cuestiones y concluye que los ítems planteados en formato de elección miden lo mismo que los formulados en formato de construcción. Sin embargo, acota que no se deben hacer generalizaciones debido a que la literatura seleccionada para la revisión está limitada al campo de la medición de lo cognitivo, fundamentalmente conocimientos en dominios académicos distintos a lectura y escritura. También matiza sus resultados advirtiendo acerca de la poca diversidad de la población estudiada y establece otras limitaciones de su estudio relacionadas con los métodos que fueron utilizados para evaluar la equivalencia.

Traub (1993) seleccionó nueve estudios sobre la equivalencia del rasgo medido, considerados por el autor los trabajos mejor diseñados y conducidos de las publicaciones más recientes para la fecha de su investigación. El propósito del estudio fue identificar los hallazgos consistentes y recomendar futuras líneas de investigación. La evidencia encontrada no es suficientemente sólida para afirmar que distintos formatos puedan medir lo mismo -en algunos dominios se revela equivalencia, mientras que en otros no- y tampoco hay hallazgos que indiquen dónde reside la diferencia. Recomienda para las próximas generaciones de estudios no considerar específicamente el asunto de la equivalencia sino más bien el del efecto del formato.

En su investigación, ampliamente citada en la literatura sobre el tema, Martínez (1999) trabaja sobre la hipótesis de que los ítems de selección y de respuesta construida difieren no sólo en la demanda cognitiva sino también en el rango de cognición a que cada uno de ellos puede dar lugar. Basándose en los resultados de trabajos empíricos previos señala la estrecha relación entre el formato del ítem y lo que se mide con éste, concluyendo que no existe un formato apropiado para todo propósito en toda ocasión. Afirma que usar una combinación de formatos puede minimizar la varianza método atribuible a un formato específico.

Rodríguez (2002, 2003) realiza un interesante trabajo de revisión meta-analítica que recoge más de 60 investigaciones sobre el formato del ítem realizadas en alrededor de los últimos 80 años, basando la importancia de este tipo de investigaciones en tres razones: (1) las interpretaciones varían de acuerdo con el formato del ítem, (2) las diferencias en el costo son importantes y (3) las consecuencias de usar un formato dado pueden afectar la instrucción. En la revisión de los trabajos, que emplean la correlación u otras técnicas para el estudio de la equivalencia del formato, encuentra

valores altamente heterogéneos que no permiten ser concluyente. Partiendo de estos resultados hace algunas consideraciones en relación con la selección del formato, relacionadas con el costo y con aspectos políticos, para finalmente concluir que lo fundamental es diseñar los tests de forma tal que los ítems puedan dar lugar a la clase de conductas especificadas en la definición del constructo en una vía en la cual la validez quede garantizada, subrayando de esta forma la importancia de la investigación acerca del significado de las puntuaciones aportadas por un formato en particular de ítem.

Otro material fundamental en este tema es el libro editado por Bennett y Ward (1993), en el que se recoge un conjunto de 14 trabajos cardinales sobre el formato de los ítems. Messick (1993), por ejemplo, aconseja realizar estudios que posibiliten distinguir entre ítems que en función de su formato puedan presentar diferencias en relación con demandas de procesamiento de información. Snow (1993) propone un conjunto de 8 hipótesis rivales en las que plantea que las diferencias en los formatos pueden o no estar en la actitud hacia los mismos, en la ansiedad que pueden producir al examinado durante la ejecución, en la motivación que generan, en la forma como influencian la instrucción y el aprendizaje, en las habilidades y estructuras de conocimiento que son capaces de medir, en la profundidad del procesamiento cognitivo y en la adecuación psicométrica de cada uno a las condiciones de prueba. Bennett (1993) explora el concepto de respuesta construida atendiendo a los significados que pueda tener desde el punto de vista descriptivo, de su validez y su fiabilidad. En fin, este libro provee una útil compilación de referencias sobre el tema y ofrece un marco para el desarrollo de la investigación sobre el formato de los ítems como un elemento de gran importancia en la evaluación cognitiva.

Conclusiones

La revisión de estas publicaciones conduce a pensar que la investigación en este campo todavía tiene mucho que aportar. Pareciera que los ítems de selección no miden exactamente lo mismo, ni con la misma precisión que los ítems de respuesta construida; que la naturaleza de las diferencias entre unos y otros no ha sido todavía comprendida en toda su extensión; que la profundidad cognitiva que puede evocar cada tipo de formato difiere y que las variables afectivas y conativas funcionan diferencialmente en función del formato. Se percibe de lo anterior un llamado a continuar profundizando en la investigación acerca del formato del ítem, por el papel tan importante que juega en la investigación sobre validez.

En todos estos trabajos revisados, aun cuando se logra establecer la existencia de diferencias, no se profundiza en la razón de esta diferencia. Como señalan Pearson y Garavaglia (1997), es importante avanzar en el estudio de la contribución a la medida que puede hacer cada uno de estos tipos de formatos de evaluación, pues conociendo la naturaleza de lo que logra medir cada formato será más sencillo determinar cuál sería el más adecuado para medir el constructo de interés con un determinado objetivo y para un grupo de sujetos y contexto de aplicación particulares, para producir puntajes que tengan una confiabilidad adecuada y que aporten el mayor número de evidencias que respalden su validez y de esta manera contribuir a través del estudio del formato al perfeccionamiento de la medición de constructos psicológicos.

Así, si las teorías de los tests terminan por moverse definitivamente hacia el planteamiento donde la calidad global de la prueba (la validez y también la confiabilidad) se evalúe en función de la interpretación y del uso que se vaya a hacer de las puntuaciones obtenidas con esta, la tarea de construir los ítems -que en definitiva son las piezas clave del test- resulta ser de enorme importancia y, por tanto, la elección del formato del ítem debe fortalecer la conformación de un instrumento de medida sólido que aporte resultados que maximicen la calidad de la información proporcionada.

Referencias

Ackerman, T. y Smith, P. (1988). A comparison of the information provided by essay, multiple-choice and free response. *Applied Psychological Measurement, 12* (2), 117 -128.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ayala, C., Shavelson, R., Yin, Y. y Shultz, S. (2002). Reasoning dimensions underlying science achievement: the case of performance assessment. *Educational Assessment, 8*, 101-121.

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

Bennett, R. E. (1993). On the meaning of constructed response. En R. E. Bennett y W. Ward. (Eds.), *Construction versus Choice in Cognitive Measurement: issues in constructed response, performance testing and portfolio assessment* (pp. 1 - 27). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bennett, R. E. y Ward, W. (Eds.) (1993). *Construction versus Choice in Cognitive Measurement: issues in constructed response, performance testing and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Downing, S. M. (2006). Selected-response item formats in test development. En S. M. Downing y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287- 301).Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Downing, S. M.y Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Embretson, S. (1993). Psychometrics models for learning and cognitive processes. En N. Frederiksen, R. Mislevy e I. Bejar (Eds.), *Test theory for a new generation of tests.* (pp. 125–150).Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Embretson, S. E. y Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gadalla, T. (1999, Abril). *Multiple-choice versus constructed response tests in the assessment of mathematics computation skills*. Documento presentado en la reunión anual de la American Educational Research Association, Montreal, Québec.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21 – 35.

Haertel, E. H. y Wiley, D. E. (1993). Representations of ability structures.En N. Frederiksen, R. Mislevy y I. Bejar (Eds.), *Test theory for a new generation of tests.*(pp. 359 – 384).Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Needham Heights, MA: Allyn y Bacon.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2ª ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hancock, G. (1992). Cognitive complexity and the comparability of multiple-choice and constructed-response test format. *Journal of Experimental Educational, 62,* 143- 158.

Hogan, T. P. (1981). *Relationship between free-response and choice-type test of achievement: a review of the literature (*Reporte Nº 70). Washington, DC: National Institute of Education.

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

Kane, M.T. y Mitchell, R. (1996). *Implementing performance assessment. Promises, problems and challenges*. Mahwah, NJ: LEA.

Martínez Arias, R. (2010). La evaluación del desempeño. *Papeles del Psicólogo, 31,* 85– 96.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207 – 218.

Messick, S. (1993). Construct validity 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. 61-73). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Messick, S. (1998). Test validity. A matter of consequences. *Social Indicators Research, 45*, 35-44.

Mislevy, R. (1993). A framework for studying differences between multiple-choice and free response tests items. En: R. E. Bennett y W. Ward. (Eds.), *Construction versus Choice in Cognitive Measurement: issues in constructed response, performance testing and portfolio assessment* (pp. 75-106). 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 Erlbaum Associates, Inc.

Nickerson, R. S. (1994). The teaching of thinking and problem solving. En R. J. Sternberg (Ed.), *Thinking and problem solving* (pp. 409-449). San Diego, California: Academic Press, Inc.

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

Pearson, P. D. y Garavaglia, D. R. (1997). *NAEP Validity Studies: Improving the information value of performance items in large scale assessments. Working paper series* (NAEP Report Nº NCES-WP-2003-08). Washington, DC: National Center for Education Statistics.

Pintrich, P. R. y Schunk, D. H. (2002). *Motivation in Education. Theory, Research, and Applications.2nd Edition.*Upper Saddle River, N J: Merrill Prentice Hall.

Rodríguez, M.(2002). Choosing an item format. En G. Tindal y T. Haladyna (Eds.), *Large- scale assessment programs for all students: validity, technical adequacy, and implementation.* (pp. 213-231). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Rodríguez, 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. y 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.

Snow, R. (1993). Construct validity 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. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher, 18*(9), 8-14.

Snow, R. E. y Peterson, P. L. (1985). Cognitive analyses of tests: implications for redesign. En S. Embretson (Ed.), *Test Design. Developments in Psychology and Psychometrics* (pp. 149 - 166). San Diego, California: Academics Press, Inc.

Sternberg, R. J., the Rainbow Projects Collaborators y University of Michigan Business School Project Collaborators (2004). Theory-based university admissions testing for a new millennium. *Educational Psychologist, 39*, 185-198.

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.

Urdaneta, E. (En prensa). Equivalencia psicométrica de tres formatos de pregunta. *Ágora.*

LA AUTORA
Elsy J. Urdaneta Durán
Ingeniera egresada de la Universidad de los Andes.
Doctora en Metodología de las Ciencias del
Comportamiento por la Universidad Autónoma de Madrid.
Profesora en la Universidad de Los Andes, Núcleo Rafael Rangel.
cuya actividad de investigación se refiere
a formatos de ítem, confiabilidad y validez.
Trujillo, Venezuela.
E mail: *elsy.urdaneta.d@gmail.com*