Criterios para el diseño de pruebas objetivas de respuesta breve o de elección de alternativas

Carmelo Basoredo Ledo

Resumen

La utilización generalizada de pruebas de elección múltiple suscita el interés por el análisis de las pruebas objetivas. En el artículo se comparan éstas con las de respuesta breve y se propone la combinación de ambos tipos para centrar la atención tanto en la selección de los contenidos como en las condiciones del formato del instrumento. Se analizan los principales defectos de construcción y se aportan directrices para su mejora. En el caso de las pruebas de elección de alternativa se prioriza un modelo de 3 opciones por pregunta, sin necesidad de controlar la influencia del azar para aprovechar el conocimiento parcial e inseguro como estrategia de respuesta.

Palabras clave: Memoria de reconocimiento, evocación de recuerdos, conocimiento parcial, respuesta patrón, tabla de especificaciones, punto de corte, validez de contenido.

1. Introducción, objeto y finalidad

Es difícil determinar con exactitud las razones por las cuales en la actualidad se está incrementando el interés por la evaluación, prácticamente en todos los ámbitos de la actividad humana. Puede que sea un efecto de la progresiva reducción de los recursos naturales, de la adquisición de un mayor grado de conciencia ecológica, o cualquier otro tipo de justificación; pero lo cierto es que, al menos en las sociedades occidentales, los esfuerzos que se realizan en pro de la calidad y la eficacia de los procesos y productos humanos, y que inciden sobre cualquier tarea ligada con la evaluación, son cada vez más notables.

En los ámbitos educativos, del trabajo o del desarrollo de las organizaciones esta actitud incide, a su vez, en la generalización progresiva de las teorías y las aplicaciones prácticas de conceptos tales como el desempeño de tareas y la competencia, definida como el comportamiento eficaz de las personas conforme a aquellos criterios que se utilicen para describir lo que ha de entenderse por eficacia. De ese modo, es preciso replantearse una y otra vez los objetivos y contenidos de la evaluación, y también revisar la validez de los instrumentos que se utilizan con una finalidad evaluadora.

Este artículo versa sobre dos tipos de pruebas objetivas de evaluación de determinados aprendizajes, las pruebas escritas de respuesta breve y las pruebas escritas de elección de una respuesta entre varias alternativas.

La utilidad o posibilidad de generalización de estas dos clases de exámenes escritos alcanza a todos los ámbitos de la Enseñanza, la Psicología, los Recursos Humanos, y cualquier otra actividad donde sea preciso recoger información para diagnosticar o evaluar rasgos, procesos cognitivos y resultados del aprendizaje o del desempeño de tareas.

Sobre ambas clases de pruebas existe ya una ingente cantidad de investigaciones realizadas a lo largo de más de 80 años, en mayor medida sobre las pruebas de elección múltiple. Sin embargo son escasos los estudios en los que las consideran como dos elementos de un mismo continuo, planteamiento metodológico que servirá de base en esta ocasión, en la creencia de que esta estrategia refuerza la validez de ambos tipos de pruebas.

El objetivo principal del trabajo es realizar una síntesis de criterios para la utilización de estas dos modalidades de cuestionario, que incluye medidas concretas, ejemplos y pautas para el diseño y el análisis de los instrumentos. Esta tarea tiene dos finalidades claras, una primera, insistir en la necesidad de incrementar el empleo de las pruebas escritas de respuesta breve, a la vista de su mayor facilidad de diseño por comparación con las de elección de alternativa y, en segundo lugar, proponer algunos cambios substanciales en el uso habitual de las pruebas de elección múltiple, lo que también las facilita, sin merma alguna para su validez.

2. Conceptos, tipos y argumentos específicos

Una prueba objetiva, en este contexto, es un examen escrito o un "test de lápiz y papel" constituido por preguntas precisas, que tienen una respuesta específica y delimitada de antemano, dentro de un reducido margen de posibles variaciones (Parsons y Fenwick, 1999).

El carácter objetivo de estas pruebas queda restringido al criterio unívoco de su evaluación (Roback, 1921), dado que sólo se admite aquella respuesta a la que *a*

priori se le ha atribuido un único valor de certeza o a cualquier otra respuesta discrecional cuyo contenido de respuesta es reconocible dentro de un margen muy reducido y bien delimitado. Se trata de pruebas de preguntas con respuesta cerrada. Por contra, la selección del dominio de conocimiento, del formato o incluso de los criterios de evaluación tiene un carácter subjetivo, que es el propio de los responsables de su diseño y utilización (Parsons y Fenwick, 1999), obviamente.

El modelo de las pruebas objetivas está tomado de las Ciencias Exactas, donde es muy común encontrar algoritmos que dejan al margen el componente subjetivo. No obstante, aunque la objetividad sea un principio ideal de cualquier instrumento de evaluación (Green, 1978), muchas variables del comportamiento humano son difícilmente evaluables mediante pruebas objetivas, por el riesgo que supone caer en un reduccionismo conceptual, metodológico e instrumental (Hernández, 2007). En definitiva, las pruebas objetivas son instrumentos válidos para evaluar determinadas variables, pero no para otras.

2.1 Definiciones y modalidades de formato

Hay formatos variados y criterios de clasificación de las pruebas objetivas bastante distintos, si bien, para este trabajo, en relación con el formato, resulta de interés distinguir dos categorías generales, las pruebas de respuesta sugerida o de reconocimiento de una determinada información proporcionada dentro de un conjunto, de las pruebas que exigen construir la respuesta, bien sea como un producto creativo inmediato, bien como recuerdo de una información previamente almacenada en la memoria (Jordan y Mitchell, 2009). Las pruebas de elección de alternativa son una modalidad de pruebas de reconocimiento y las de respuesta breve pertenecen al tipo de respuesta elaborada.

Por tanto, una respuesta breve es aquella que se prepara para cada ocasión, sin ningún tipo de ayuda que no sean los recursos cognitivos de la persona y siempre que su extensión no exceda de 150 palabras, como criterio de diferenciación de las pruebas de ensayo escrito (Basoredo, 2008). Ellintong (1987) distingue tres subtipos de pruebas de respuesta breve, las de completar textos mutilados o cualquier información gráfica fragmentaria, las que tienen una respuesta única y las de respuesta abierta.

Por su parte, las pruebas de elección entre varias alternativas se caracterizan por sugerir la respuesta correspondiente a las preguntas. Este instrumento puede adoptar distintas formas, de elección única entre un número reducido de respuestas, de elección de más de una alternativa, de atribución de valor de verdad o falsedad a todas las alternativas de respuesta, de emparejamiento entre varias alternativas, etc. (Haladyna, Downing y Rodríguez, 2002). En este artículo solamente se trata la variedad de elección de una única respuesta entre un número inferior a cinco propuestas, por ser, probablemente, la más utilizada.

2.2 Aspectos generales sobre criterios de diseño y análisis

El mayor atractivo que suscita el empleo de pruebas objetivas es su facilidad de corrección en comparación con otras pruebas escritas de ensayo, de solución de problemas, etc., pero cuando se realizan diversos análisis sobre su validez se observan, al menos, dos tipos de problemas, unos que tienen que ver con el contenido de las pruebas y otros con el diseño y manejo de los instrumentos. O sea, el uso de las pruebas objetivas —como de cualquier otra— debe cuidar dos elementos fundamentales, ¿qué se evalúa?, esto es, el objeto y contenido, por una parte, y por otra, ¿cómo se evalúa? (Ebel, 1982), tanto en lo referido al propio formato del instrumento como a las operaciones de tratamiento e interpretación de los datos.

Los cuatro criterios fundamentales a los que debe responder el contenido de cualquier examen son, su importancia y representatividad, así como su grado de dificultad y su poder de discriminación entre las personas que obtengan un alto y un bajo rendimiento en la prueba. Los dos primeros son las claves para realizar las inferencias relativas a la validez del contenido (Binning y Barret, 1989). Por su parte, los dos últimos, o sea, el grado de dificultad y el poder de discriminación guardan mayor relación con el modo en cómo se evalúa, esto es con la fiabilidad de la prueba y su determinación exacta únicamente es posible mediante el análisis de items posterior a la aplicación de la prueba a una muestra de personas de la población a evaluar (Muñiz, Fidalgo, García-Cueto, Martínez y Moreno, 2005).

El problema reside en que muchas de las investigaciones realizadas al efecto suelen centrar más la atención en uno de estos dos conjuntos en detrimento del otro, cuando es absolutamente necesario considerar rigurosamente ambos.

Por ello se han elegido en este trabajo estas dos modalidades de pruebas, la de respuesta breve para subrayar el valor de la selección de los objetivos y contenidos de evaluación y la de elección múltiple para orientar el análisis hacia la problemática sobre la forma de evaluar.

En cierto modo, con independencia del tipo y subtipo de prueba objetiva que se trate, cabe la posibilidad de establecer una correspondencia entre una prueba de respuesta breve y cualquier otra prueba objetiva sobre la base de dos principios metodológicos, el de simetría cognitiva y el de contigüidad procesal:

a) El principio de simetría cognitiva se basa en la constatación de que todas las pruebas objetivas, con algunos matices diferenciales, prácticamente son adecuadas para evaluar unas determinadas variables y no otras, pero, mientras que las pruebas de respuesta breve exigen un esfuerzo cognitivo de evocación de un recuerdo y, tal vez, de cierta habilidad para el pensamiento creativo, en el resto de los tipos el proceso cognitivo es el propio de la memoria de reconocimiento.

b) La contigüidad procesal se explica por el hecho de que del resultado de las primeras operaciones de selección de objetivos y contenidos siempre se deriva un cuestionario de respuesta breve, con sus interrogantes, sus respuestas apropiadas y sus criterios. Sólo cuando esta tarea está terminada, a un nivel más o menos explícito, el siguiente paso será convertir este cuestionario en cualquier otro de una variedad distinta.

Por consiguiente, ésta es la estrategia más recomendable para el diseño y el análisis de cualquiera de los tipos y subtipos de pruebas objetivas, con independencia de las características diferenciales de los mismos.

La elección del formato de respuestas de elección entre varias alternativas tiene, igualmente, su propia justificación. Una respuesta breve es siempre discrecional, ya que puede que quepan distintas formas para responder, mientras que la elección de alternativa se asimila a cualquier algoritmo, donde una excluye a las demás. Esto es, las respuestas breves de carácter abierto permiten cualquier formulación, dentro de unos límites de valor de certeza, o dicho de otro modo, el criterio general es aceptar cualquier enunciado que, de conformidad con los criterios de contenido, no sea falso, lo que, además, también permite una cierta valoración gradual. El paso siguiente consiste en "cerrar" las posibilidades de respuesta ofreciendo varias alternativas, plausibles todas, entre las cuales, al menos una de ellas sea verdadera o unívocamente considerada como la mejor respuesta.

3. Características y utilidades de estos dos tipos de pruebas objetivas

Ordinariamente, se considera que las pruebas objetivas son instrumentos de evaluación muy eficientes, fáciles de utilizar y con alto grado de fiabilidad estadística y validez de contenido. No obstante, su diseño es una tarea compleja, tanto por lo que se refiere a la selección de los objetivos y contenidos de los cuestionarios, como respecto al formato y los criterios de evaluación, siendo las de respuesta breve las más simples de todas ellas.

Por contra, entre las desventajas del uso de pruebas objetivas, amén de la ya referida complejidad del diseño, se alude a la superficialidad, trivialidad o irrelevancia del contenido, por el hecho de que son las preguntas que responden a estas características las más sencillas de redactar (Green, 1978), aunque tales deficiencias se podrían corregir, si los cuestionarios estuvieran bien construidos (Fuhrman, 1996).

Morales (2006) centra sus críticas al uso de las pruebas objetivas en a) la excesiva cantidad de preguntas memorísticas y descontextualizadas que contienen, b) la relativa falta de coherencia entre sus contenidos, objetivos y criterios de evaluación, y c) el perjuicio que pueden suponer para aquellas personas cuyo estilo cognitivo sea *dependiente de campo.* En la misma línea Hernández (2007) alude al hecho de que se trata de instrumentos de evaluación más dirigidos a comprobar la acumulación de conocimientos que a la solución de problemas concretos y, en consecuencia, ajenos a las tareas reales. Esto quiere decir que el abuso de pruebas objetivas deviene en aprendizajes poco significativos, carentes de sentido y aplicabilidad, que dificultan el conocimiento situado, como parte y producto de la actividad y el contexto en el que se desarrolla y emplea (Díaz-Barriga, 2003).

A la vista de las anteriores apreciaciones, dentro de cualquier proceso de evaluación de aprendizajes, evaluación de competencia profesional o del desempeño de tareas, conviene restringir el uso de las pruebas objetivas a aquellos contenidos y objetivos para los que se ha comprobado que son buenos instrumentos, combinándolas con otros instrumentos más adecuados, aún siendo de mayor complejidad, y siempre y cuando su diseño y utilización responda a unas normas rigurosas, producto de la investigación.

En definitiva, y ya centrándose en los objetivos preferentes para el uso de cuestionarios de respuesta breve o de elección de alternativa, cabe destacar aquellos que tienen mayor relación con la evaluación de conocimientos declarativos y contenidos factuales como datos, términos, definiciones o conceptos, y también con los procesos cognitivos de recuerdo, comprensión y análisis, en mayor medida (Basoredo, 2009). A éstos podrían añadirse la evaluación de conocimientos sobre las formas de proceder y algunos elementos de los procesos cognitivos de aplicación del conocimiento a la realización de tareas, pero no otros. Para la evaluación de objetivos referidos a la mayoría de los contenidos procedimentales y de los procesos cognitivos de aplicación, creatividad o síntesis, así como de las habilidades y destrezas, es preferible emplear otro tipo de pruebas escritas o de ejercicios de simulación y solución de problemas (Basoredo, 2010).

4. Fases del proceso de diseño y uso de las pruebas objetivas

Las pruebas objetivas tienen múltiples aplicaciones y la finalidad específica de las mismas, realmente, influye bastante en las condiciones y las circunstancias de su diseño y utilización. Jornet y Suárez (1996) distinguen varios tipos de uso, entre ellos, cuando se emplean como indicadores de resultados dentro del Sistema Educativo, por ejemplo, o cuando se trata de pruebas de diagnóstico o de nivel de

dominio, como podrían ser las de selección de personal y evaluación de la competencia profesional.

Tratándose de indicadores de resultados del aprendizaje, es probable que sean los propios docentes quienes las diseñen y utilicen dentro de su aula, con lo cual las tareas de análisis del contexto y de acotamiento de los dominios de conocimiento se reducen mucho, porque para cuando se van a emplear estos instrumentos ya consta una información muy precisa sobre tales aspectos. Además, en este caso, la trascendencia de la interpretación de los resultados es menor que en circunstancias de diagnóstico, máxime si se utilizan instrumentos complementarios de evaluación más variados, dentro de un modelo de evaluación continua.

Sin embargo, cuando la finalidad de las pruebas implica la toma de decisiones con consecuencias de naturaleza selectiva y efectos de discriminación entre las personas evaluadas, cual es el caso de pruebas estandarizadas de certificación y de admisión (Jornet y Suárez, 1996), hay que poner un mayor énfasis en la elección de instrumentos de evaluación de probada validez.

Esta circunstancia de mayor exigencia se traduce en dos medidas necesarias, que son, la intervención de agentes expertos, tanto en materia de los dominios de conocimiento como sobre metodología de evaluación, y también la descripción precisa y detallada de las fases y las condiciones de diseño y uso de estas pruebas.

El diseño y utilización de cualquier prueba objetiva incluye las siguientes fases, tres de la cuales son coincidentes con las propuestas por Soubirón y Camarano (2006):

Fase I: Análisis del contexto.
Fase II: Selección de los contenidos.
Fase III: Redacción del cuestionario.
Fase IV: Evaluación de las respuestas e interpretación de los resultados.

La primera fase requiere agrupar las tareas en dos direcciones, una se refiere al análisis de las condiciones y circunstancias del propio contexto en si, ámbito de estudio, finalidad y objetivos, análisis de las tareas, etc., y la segunda dirección trata del análisis de las características de la población a quien va dirigida la prueba. El argumento justificativo de estas tareas previas tiene que ver con la determinación de los grados de dificultad y de discriminación de cada uno los elementos del cuestionario (Basoredo, 2008; Jornet y Suárez, 1996), especialmente si no es factible realizar el análisis de items de una prueba piloto.

La selección de los objetivos y contenidos específicos propios de la segunda fase gira en torno a la determinación de la amplitud del dominio de conocimiento, sus límites y sus dimensiones (Jornet y Suárez, 1996), así como la identificación precisa del nivel de exigencia previsto y otros criterios de evaluación. La concreción y síntesis de todos estos elementos se realiza con ayuda de una tabla de especificaciones (Ministerio de Educación de Guatemala, 1971; Soubirón y Camarano, 2006; Basoredo, 2008).

El paso siguiente consiste en la redacción del cuestionario, a partir de un borrador inicial de preguntas de respuesta breve, con su correspondiente respuesta y justificación de la validez del contenido. El cuestionario definitivo, como sabemos, puede adoptar el mismo formato del borrador o transformarse en cualquier otro, estrategia de doble operación que se propone en este trabajo al considerar que es una medida de refuerzo de la validez del contenido del instrumento de evaluación.

La evaluación de las respuestas incluye, en primer lugar, la corrección de las mismas, luego la puntuación con la ayuda de los criterios especificados y, por último, la interpretación definitiva de los resultados.

5. El diseño de pruebas de respuesta breve

De los dos tipos de pruebas objetivas, objeto de análisis en este trabajo, ya es sabido que las pruebas de respuesta breve se caracterizan por la discrecionalidad y apertura de la misma, sin indicio alguno ni propuesta definida que la condicione más allá del planteamiento realizado en la propia pregunta.

Entre las tres variedades de respuestas breves señaladas por Ellintong (1987), la respuesta única, la respuesta mediante enunciados plurales y la respuesta de completar textos, aunque todas ellas puedan dar origen a operaciones cognitivas de carácter similar, las dos primeras tienen un formato idéntico de pregunta y de respuesta, que sólo exige comprender o interpretar los estímulos incluidos en el planteamiento inicial. Estas dos variedades implican dar respuestas en lenguaje natural, derivadas de diferentes formas de procesos cognitivos de memoria y evocación de recuerdos (Jordan y Mitchell, 2009). Por su parte, encontrar las palabras para cumplimentar un texto incompleto exige mayores esfuerzos de comprensión e interpretación de enunciados distintos, característica tal que, de este modo, aumenta la complejidad del diseño del formato de esta tercera variedad. Por tanto, en esta ocasión se centrará la atención en las dos primeras, en las que las dos fases más importantes del diseño y uso son la selección de los contenidos y la concreción de los criterios de corrección y calificación de las respuestas.

5.1 Ventajas e inconvenientes de las pruebas de respuesta breve

Las pruebas de respuesta breve son buenos instrumentos para evaluar dominios de conocimientos y procesos cognitivos de comprensión, análisis y aplicación de un nivel medio-bajo de dificultad (Ellintong, 1987). Suelen tener, no obstante, un grado de dificultad un poco mayor que las de elección de alternativa, al carecer del recurso que supone el reconocimiento de un contenido explícito entre varias opciones de respuesta.

Desde la perspectiva de los resultados del aprendizaje (Jonassen y Tessmer, 1996), este tipo de pruebas es útil para la evaluación de información relevante o clave, el conocimiento de textos, marcos de referencia teóricos o conceptos, el uso de procedimientos, reglas o principios, la definición y solución de problemas sencillos, o la construcción de analogías y argumentos simples.

Una de las ventajas de este tipo de pruebas, sobre el resto de variedades de pruebas objetivas, es que carecen de los defectos incluidos entre los contenidos de las alternativas de respuesta, que pueden aumentar de manera espuria la dificultad de la prueba o aportar indicios para la selección de la respuesta idónea (Case y Swanson, 2006). La característica principal de este tipo de pruebas es que permite a la persona examinada expresarse con sus propias palabras para dar respuesta al interrogante que se le plantea (Davis, 1967).

Sin embargo, por este mismo motivo, su fiabilidad suele ser algo menor que la de elección de alternativa (Bleske-Rechek, Zeug & Webb, 2007), dado el margen de discrecionalidad que cabe admitir en las respuestas cuando no sean unívocas. A esto hay que añadir, como desventaja, la incomodidad que supone para las personas encargadas de la corrección la lectura de textos cuya caligrafía dificulta su comprensión. Si, además, la selección de los contenidos, por descripción y delimitación de los dominios de conocimiento y categorización de los criterios de evaluación, se realizara a falta del debido rigor, como, por ejemplo, cuando se prescinde de los argumentos de justificación para aceptar o rechazar una

determinada respuesta, estas pruebas son difíciles de evaluar (Thyne,1978; Tenbrinck, 1984; Lafourcade, 1985).

5.2 Características de los agentes y del análisis del contexto

Una idea que ya se apuntaba anteriormente era la necesidad de elegir agentes expertos tanto para el diseño como para las tareas propias de la evaluación, particularmente cuando de la utilización de este instrumento se deriven consecuencias importantes para las personas.

La determinación de quienes hayan de responder del diseño y uso de pruebas objetivas va ligada a la primera fase de análisis del contexto, y su idoneidad y nivel de experticia guarda relación con tres ámbitos concretos, ya que deben ser expertos en la materia de la prueba objeto de examen y, además, mostrar habilidades desarrolladas para el análisis de las características de las personas a evaluar y para el uso de técnicas de evaluación cuantitativa y cualitativa de algunas de estas características. Un equipo eficiente es el formado, al menos, por tres especialistas, dos expertos en los dominios de conocimiento, por ejemplo, uno de ellos de orientación más teórica y el otro de carácter más bien aplicado, y un tercer miembro especializado en materia de metodología de análisis y evaluación (Basoredo, 2008).

La primera fase de análisis del contexto forma parte de lo que Muñiz y colaboradores (2005) consideran como directrices previas a la construcción de los enunciados y es absolutamente necesaria en la mayoría de las ocasiones, máxime cuando no se disponga de todos los elementos del dominio del contenido a evaluar, carencia ésta que obliga a realizar un muestreo de las materias por áreas representativas (Pelechano, 1988).

El análisis contextual centra su atención en dos direcciones, una es la concreción de la intención general del examen en términos de objetivos específicos detallados y la otra es la descripción de las características de la población a que va dirigida la prueba. A partir de ambos puntos de enfoque se deben obtener las características básicas del objeto de medida y los principales criterios de interpretación de las puntuaciones (Jornet y Suarez, 1996).

Si se trata de un examen de rendimientos del aprendizaje, los elementos de partida del análisis contextual son el plan de estudios y el programa de enseñanza y, en el caso de su uso en selección de personal o evaluación del desempeño, el estudio del contexto se sustancia con ayuda de la descripción del puesto de trabajo y el análisis de tareas.

Por su parte, dos son, básicamente, las características de las personas a evaluar que han de considerarse en la fase inicial de análisis del contexto, su amplitud y su grado de diversidad (Jornet y Suárez, 1996). La información precisa sobre estas variables permitirá estimar unos primeros niveles teóricos de dificultad y discriminación de la prueba, pues el empleo de una misma prueba en un colectivo reducido y homogéneo da origen a resultados muy distintos a los obtenidos por un colectivo numeroso y heterogéneo.

Los análisis contextuales, por tanto, aportan la configuración inicial de la tabla de especificaciones —que se describirá en una sección posterior— y su concreción definitiva es consecuencia de la selección de los contenidos y de los criterios de calificación, puntuación e interpretación de las puntuaciones.

5.3 Selección de los contenidos

Pelechano (1988) distingue dos situaciones para proceder a la selección de los contenidos de una prueba que respete los dos criterios de validez de contenido, a) cuando constan en detalle todos los elementos potencialmente evaluables del dominio y b) cuando el dominio es de carácter amplio y está escasamente delimitado. Un ejemplo del primer caso puede ser el contenido de una norma ISO de calidad, mientras que una muestra del segundo caso es, por ejemplo, las nociones básicas sobre Dietética y Alimentación.

El procedimiento general para la selección de contenidos de un dominio absolutamente delimitado consiste en atribuir un grado teórico de dificultad global a la prueba y específico a todos los elementos del dominio (Jornet y Suárez, 1996) y posteriormente realizar una selección aleatoria de los temas, por grupos de dificultad.

Sin embargo, en la segunda de las situaciones, que es la más compleja y la más común, es preciso, primero, realizar una delimitación del dominio, después, una concreción en detalle de los temas, epígrafes, etc., con la estimación de la importancia y del grado de dificultad correspondiente y, por último, un muestreo por áreas representativas (Pelechano, 1988; Case y Swanson, 2006).

Una estrategia muy recomendable consiste en formular inicialmente todos los interrogantes posibles sobre los temas, sin ningún tipo de limitaciones de formato, ni extensión.

A continuación se estima el grado teórico de dificultad global de la prueba, a la vista de la exigencia de sus objetivos y las características de la población de personas a examinar. Por ejemplo, un examen de dificultad de tipo medio puede constituirse a razón de 1/3 de preguntas fáciles, difíciles y de mediana dificultad y uno de dificultad alta o baja, respectivamente, a razón de ½ de preguntas difíciles, 4/10 de dificultad media y 1/10 de preguntas fáciles, o viceversa.

Por último, se da forma al primer borrador de preguntas directas, con sus soluciones correspondientes y las referencias exactas o argumentos que explican la validez de las mismas. Al conjunto de soluciones debidamente justificadas se le denomina la respuesta patrón (Basoredo, 2008).

5.4 Redacción del cuestionario de preguntas

La redacción de preguntas de respuesta breve ha de seguir las reglas básicas que se emplean para la elaboración de cuestionarios de recogida de cualquier tipo de información por medio de encuesta.

Algunas de las medidas más importantes para la redacción definitiva de las preguntas tienen que ver con la validez del contenido, con la claridad y el grado de comprensión que se necesita para evaluar los objetivos que se pretenden y con las limitaciones espacio-temporales que son necesarias para calibrar el esfuerzo requerido a las personas que han de realizar el examen.

Son muchos los autores que insisten en la necesidad de redactar las preguntas de forma muy clara, sin ambigüedades y expresadas en un lenguaje sencillo, fácilmente comprensible y gramaticalmente correcto (Pelechano, 1988; Furhman, 1996; Parsons y Fenwick, 1999; Muñiz y García-Mendoza, 2002; Moreno, Martínez y Muñiz, 2004 y 2006; Morales, 2006; Soubiron y Camarano, 2006).

Las preguntas es preferible redactarlas en forma interrogativa directa, evitando verborrea innecesaria con objeto de que las personas capten desde un principio el interrogante esencial y contribuyendo, así, a minimizar el tiempo de lectura (Muñiz

y García-Mendoza, 2002). También es posible redactar la pregunta en forma de un enunciado de tarea, iniciada siempre por un verbo que demanda una acción concreta, como, por ejemplo: *"Indicar los puntos cardinales que sigue la trayectoria del sol en el hemisferio Norte".*

Por otra parte, es necesario evitar material discutible y preguntas basadas en meras opiniones (Pelechano, 1988; Muñiz y García-Mendoza, 2002). Lo adecuado es limitarse a redactar preguntas sobre hechos, conceptos y modelos, leyes y principios, teorías, o procedimientos de aplicación y generalizaciones (Rebustillo y Moltó, citados por Hernández, 2007).

El número total de preguntas del cuestionario debe cubrir cada uno de los grandes apartados de la tabla de especificaciones, distribuidas mediante porcentajes por cada uno de los subapartados (Pelechano, 1988).

A continuación se indican algunas recomendaciones más para la redacción de las preguntas de respuesta breve, que la mayoría son igualmente válidas para la redacción del enunciado principal en la pruebas de elección de alternativa:

a) Redactar un número suficiente de preguntas para representar los aspectos principales de los contenidos que son objeto de examen.
b) Proceder pregunta por pregunta, cada una sobre un único motivo.
c) Redactar cada pregunta con tal claridad, concisión y ausencia de ambigüedad, que corresponda a una respuesta unívoca y específica para el interrogante planteado.
d) Emplear, preferentemente, preguntas directas o enunciados que demanden una tarea concreta.
e) Evitar formulaciones idénticas a las utilizadas en los manuales o en cualquier otra publicación de referencia.
f) Evitar expresiones que puedan ofender o dañar la dignidad de las personas.
g) Ordenar las preguntas con un determinado criterio, por dificultad, dominio de conocimiento, sección temática, etc.
h) Limitar el espacio disponible para contestar, como máximo al doble del que se estime necesario.
i) Limitar el número total de preguntas al periodo de tiempo destinado para la respuesta, a razón de 2 preguntas por minuto, aproximadamente.
j) Comprobar la relevancia de todas las preguntas *("¿son todas las que están?")*
k) Comprobar la representatividad de todos los contenidos importantes del dominio *("¿están todas las que son?").*
l) Comprobar la ausencia total de ambigüedades, tanto en las preguntas como en la delimitación de la respuesta patrón.
m) Realizar los cambios derivados de los resultados de todas las comprobaciones.

6. La evaluación de las preguntas de respuesta breve

La evaluación de las preguntas breves sigue la metodología general de evaluación de todos los exámenes de desarrollo escrito. O sea, cada pregunta debe tener su respuesta patrón específica unívoca o lo más precisa y delimitada posible, aunque pueda admitir algunas variaciones.

A su vez, se necesita una escala de evaluación acerca del grado de certeza y validez de las respuestas. Por tratarse de exámenes muy concretos, puede resultar idónea una escala de sólo 3 tipos, bien (1), regular (0,5) y muy deficiente (0).

La síntesis de los contenidos de la respuesta patrón y los criterios de evaluación constituye la tabla de especificaciones de la prueba.

6.1 La tabla de especificaciones

Una tabla de especificaciones o matriz de valoración es un cuadro de doble entrada para representar los principales elementos del contenido, la respuesta patrón y los criterios específicos de evaluación de cualquier prueba o examen. Toda tabla de especificaciones se elabora a partir de una clasificación de dimensiones, subdimensiones, temas o epígrafes que se sitúan en el eje de ordenadas y otra clasificación con los criterios de importancia, dificultad y puntuación en el eje de abscisas.

Para elaborar una tabla de especificaciones se concreta el nivel de dificultad en que serán evaluados los contenidos, así como el peso o importancia que tendrán en relación con el conjunto de la prueba (Soubirón y Camarano, 2006). La tabla de especificaciones, por consiguiente, tiene la finalidad de expresar la importancia relativa de los distintos objetivos de la prueba entre si y con respecto al proceso de evaluación en su conjunto (Ministerio de Educación de Guatemala, 1971), que pudiera incluir una variedad de pruebas distintas.

La principal utilidad de una tabla de especificaciones es automatizar el procedimiento de evaluación de una prueba. Dado que cualquier tabla de especificaciones incluye las categorías a evaluar y la proporción de puntuación máxima que les corresponde a cada una sobre la puntuación total (Basoredo, 2008), la puntuación directa de un examen es la resultante de la suma de los productos de cada una de las respuestas corregidas por la proporción atribuida en cada una de la filas de la tabla.

*Ejemplo: Si la corrección de un examen de 5 preguntas con una puntuación máxima de 10 puntos, a razón de 2 puntos cada respuesta correcta, da como resultado 2 preguntas bien contestadas, 2 regular y 1 deficiente, de acuerdo a la escala de tres valores anteriormente indicada, la puntuación directa obtenida son 5 puntos [PD= (2*1*2)+ (2*0,5*2)+ (1*0)].*

La construcción de la tabla de especificaciones se inicia en la primera fase de análisis del contexto, se desarrolla en mayor media en la segunda fase de selección del contenido y se completa tras la finalización de la redacción definitiva del cuestionario y de la respuesta patrón. Su elaboración admite multiplicidad de variantes, incluidas distintas fórmulas de ponderación de las respuestas, por su importancia o dificultad.

El aspecto de mayor interés de cualquier tabla de especificaciones o matriz de valoración es que hace posible la comparación de los resultados de la corrección de la prueba con los criterios específicos de valoración de la misma y, de ese modo, obtener un resultado que permita la adopción de una decisión objetiva, justa y generalizable.

6.2 Criterios y procedimientos de evaluación

La síntesis de los criterios de evaluación se recoge, tal y como se dicho, mediante una tabla de especificaciones.

Un criterio es una norma o regla para comparar cualquier objeto o fenómeno acerca del cual es preciso realizar una evaluación o elaborar un juicio.

Toda evaluación ha de fundamentarse en datos, hechos y argumentos objetivos, dejando al margen conjeturas carentes de justificación. Por tanto, con carácter previo a la ejecución de cualquier procedimiento de evaluación, se necesita disponer de criterios y procedimientos para realizar las comparaciones entre la información a evaluar y la información del criterio o patrón.

En los ámbitos referidos en este artículo y desde la perspectiva del criterio, dos son los modelos posibles para la evaluación, uno es comparar las respuestas de las pruebas realizadas únicamente con la respuesta patrón y una segunda modalidad consiste en comparar las respuestas de cada individuo con las respuestas dadas por todas la personas evaluadas. Al primer modelo se le denomina *evaluación por criterio* y al segundo *evaluación por norma de grupo.*

La comparación de una respuesta dada con la respuesta patrón se realiza siempre, en todo caso. El elemento diferenciador entre ambos modelos es tomar una decisión basándose únicamente en esta primera comparación o bien realizar una segunda comparación con las respuestas dadas por todos, cuando el número de personas sea suficiente, por ejemplo, 30 o más individuos.

Generalmente, la razón para adoptar uno u otro modelo reside en la principal finalidad de la evaluación, de modo que tratándose de un proceso de evaluación de conocimientos, de carácter selectivo, es común emplear sólo la evaluación por criterio, determinando *a priori* un punto de corte para cuantificar el nivel de exigencia mínimo de superación de la prueba.

Toda evaluación se deriva de tres operaciones, la primera, es la corrección de la prueba, comparando las respuestas de las personas evaluadas con la respuesta patrón, la segunda es obtener la puntuación, aplicando las reglas y criterios de la tabla de especificaciones y la tercera consiste en hallar el resultado definitivo, que incluye la decisión correspondiente.

En el ejemplo explicado en la sección anterior, de las dos primeras operaciones había resultado una puntuación de 5 puntos sobre 10, pero no se había decidido nivel de exigencia alguno, con lo cual, dependiendo de cuál fuera el punto de corte de la prueba este resultado podría ser suficiente o no. Para un nivel de exigencia medio o medio bajo 5 puntos es un resultado suficiente, pero si el punto de corte de la prueba, que responde a un nivel de exigencia alto, fuera 7, este resultado es claramente deficitario.

La única condición exigida para legitimar cualquier proceso de evaluación como éste es informar previamente a las personas que van a ser evaluadas de los principales criterios y normas que afectan al mismo, como la escala de puntuación, el punto de corte de la prueba o el procedimiento preciso para su obtención, de modo que una vez conocido un resultado concreto éste sea objetivamente explicado a partir de la aplicación de los referidos criterios.

7. Las pruebas de elección de respuestas entre varias alternativas

Las pruebas de elección de una respuesta entre varias alternativas, tal y como se han planteado en este estudio, son una variante de las pruebas de respuesta breve que se caracterizan por activar los procesos de la memoria de reconocimiento. En realidad, son instrumentos adecuados para evaluar objetivos y contenidos análogos a los de las pruebas de evocación o recuerdo. Dolinsky y Reid (1984) resumen los objetivos a evaluar mediante este tipo de instrumento en los de reconocimiento de una información, la comprensión de un planteamiento, la aplicación de un procedimiento y el análisis de una situación. A estos Case y Swanson (2006)

añaden la interpretación de textos, con la posibilidad de proponer un juicio sobre los mismos.

La particularidad de este tipo de pruebas consiste en proponer un conjunto de varias alternativas para cada pregunta (Dolinsky y Reid, 1984; Moreno, Martínez y Muñiz, 2004), centrando el esfuerzo de la persona en encontrar entre ellas la respuesta que sea cierta o bien aquella a la que se le ha atribuido *a priori*, con carácter unívoco, el valor de ser la mejor entre todas. El resto de respuestas alternativas son los distractores. Esta singularidad hace que se trate de un instrumento utilizado con gran frecuencia por la ventaja que reporta la corrección automatizada de las mismas (Bush, 2006) y la aparente justificación de objetividad de las respuestas "correctas". A pesar de ello, en el mismo elemento en el que radica esta ventaja se halla también su mayor inconveniente, ya que la redacción de las alternativas de respuesta aumenta enormemente la dificultad de su diseño (Bush, 2006), por comparación con las pruebas de evocación de respuesta breve, a la vez de incrementar así la probabilidad de sesgos.

El modelo que se ha propuesto ya desde un principio aconseja que el proceso de diseño de este tipo de pruebas sea idéntico al de las pruebas de evocación de respuesta breve, salvo en la fase del diseño del cuestionario en la que, a su vez, hay que disponer de un procedimiento específico para la redacción de las alternativas de respuesta. Por su parte, para la evaluación de las mismas únicamente se requiere de una escala de 2 valores, acierto (1) y error (0), pudiendo ser el resto de criterios y especificaciones similar al de las anteriores.

Sobre esta clase de pruebas hay multitud de investigaciones, muchas de las cuales han analizado y propuesto medidas sobre los tres temas que son más recurrentes al respecto, como los defectos y errores del diseño, el número de alternativas de respuesta que sería más apropiado o la importancia de las respuestas dadas al azar y su mayor o menor necesidad de control.

7.1 Defectos y errores de las pruebas de elección de respuesta

En el diseño de las pruebas de elección de respuesta cabe distinguir las cuestiones referidas al contenido de aquellas otras que se refieren a la redacción de los enunciados de las preguntas y, sobre todo, de las opciones de respuesta (Haladyna, Downing y Rodríguez, 2002; Muñiz y García-Mendoza, 2002; Moreno, Martínez y Muñiz, 2004). A estas dos categorías habría que añadir también los aspectos relacionados con las instrucciones y la gestión de las sesiones en las que se realiza el examen.

Los posibles defectos y errores relativos al contenido son comunes para las dos modalidades de pruebas y los más importantes se han comentado en un apartado anterior; de modo que en esta sección se dedicará una atención especial a las deficiencias en la redacción de las preguntas y las opciones de respuesta, siendo los distractores en donde se da una mayor frecuencia de error de redacción (Frary, 1995).

Las deficiencias en la redacción de los enunciados de las preguntas y las alternativas de respuestas se agrupan en dos categorías generales, a) indicios de la respuesta válida y b) incremento espurio de la dificultad (Moreno, Martínez y Muñiz, 2006; Case y Swanson, 2006).

Entre los errores de diseño que dan muestra o indicios sobre cuál de las opciones es la respuesta válida cabe citar, la falta de concordancia gramatical entre el enunciado de pregunta y alguna de las opciones alternativas, la diferencia notable en la extensión o número de palabras de alguna de las opciones, que suele ser la mejor respuesta, la repetición desproporcionada de la mejor respuesta en la misma

posición entre los distractores, la convergencia de los elementos comunes de las opciones en la mejor respuesta, el uso indebido o desproporcionado de términos absolutos, los solapes entre las opciones, la repetición de términos relevantes en el enunciado y en las opciones, o las pistas de tipo lógico, cuando, por ejemplo, varias de las alternativas agotan todas las posibilidades de respuesta y se queda una al margen o cuando algunas pertenecen a una misma categoría y la respuesta válida pertenece a otra distinta.

Por su parte, algunas formas de redactar aumentan indebidamente el grado de dificultad de la prueba. Entre todas, cabe señalar las siguientes: la complejidad o extensión innecesaria del enunciado de la pregunta o de las opciones, la ambigüedad o imprecisión de varios términos, como, por ejemplo, los que expresan frecuencia, el uso intrascendente de las partículas negativas, la mezcla de distintos formatos para expresar datos, la interdependencia de respuestas a preguntas distintas, la heterogeneidad de los criterios lógicos para redactar las opciones, la falta de discriminación entre unas opciones y otras o la carencia de lógica en el orden de presentación de las distintas opciones.

Es más, aunque entre las referencias consultadas no se ha encontrado información sobre este aspecto, también hay que señalar algunas deficiencias que, por experiencia, tienen que ver con la redacción de las instrucciones y con la gestión de la sesión en que se realiza la prueba. Entre ellas merece la pena citar la carencia o falta de comprensión sobre cuál es la tarea que deben de realizar las personas que van a ser examinadas o del procedimiento concreto para realizar el examen, una explicación insuficiente o excesiva de los criterios y el mecanismo de evaluación de la prueba, o la gestión deficiente del tiempo de respuesta por exceso o por defecto.

Unos y otros errores suelen combinarse en una misma pregunta, aumentando la probabilidad de los distintos sesgos hasta poder invalidar de forma absoluta la prueba, haciendo que ésta pierda el carácter objetivo que de ella se predica.

7.2 Número ideal de preguntas y opciones de respuesta

El número más apropiado de preguntas de una prueba objetiva es una variable escasamente estudiada y difícil de estimar con carácter general (Burton, 2006). La razón que este autor trata de explicar para ello es que la longitud de una prueba objetiva viene determinada por factores tan variados como el propio formato de la misma, el modelo de fiabilidad utilizado, el grado de dificultad de las preguntas y, sobre todo, la amplitud y heterogeneidad del dominio de conocimiento.

El principio psicométrico básico afirma que la longitud de cualquier test guarda una proporcionalidad directa con su índice de fiabilidad, que en sentido práctico se traduce en aumentar el número de preguntas cuando se quiere mejorar su consistencia. Sin embargo, la cuestión está en valorar el tamaño del efecto derivado del aumento de la longitud de la prueba, porque a partir de ciertos límites el incremento en los índices de fiabilidad es intranscendente.

Tratándose de un dominio de conocimiento no muy reducido, según Burton (2006) los límites para disponer de una prueba objetiva fiable oscilan entre 30 preguntas de respuesta breve y 300 items de verdadero-falso aproximadamente, si bien este autor no se atreve a establecer una longitud mínima adecuada con carácter general para un tipo de formato determinado.

No ocurre lo mismo en relación con el número óptimo de alternativas por pregunta en el caso de las pruebas de elección múltiple, ya que, tras 80 años de investigación y haber realizado multitud de estudios, se ha llegado a la conclusión de que 3 alternativas por pregunta es el número ideal tanto por razones de

dificultad, como de discriminación, fiabilidad o validez de la prueba (Rodríguez, 2005).

Afirma Rodríguez (2005) en su estudio de meta-análisis que incluye un total de 48 investigaciones anteriores realizadas entre 1944 y 1995, que la reducción de 5 alternativas por pregunta a 3, produce una disminución insignificante del 7% en su grado de dificultad, no afecta en modo alguno al poder de discriminación de la misma y disminuyen un 6%, por término medio, los índices de fiabilidad. Cuando se reduce de 4 alternativas a 3, el grado de dificultad disminuye un 4%, el índice de discriminación aumenta un 3% y la fiabilidad un 2%. Sin embargo una reducción mayor, de 5 ó 4 alternativas a 2, da como resultados alteraciones significativas del grado de dificultad, que disminuye entre un 19% y un 23%; el poder de discriminación también se reduce entre un 9% y un 11%, lo mismo que los índices de fiabilidad.

Lord (1977), desde la perspectiva del análisis de las curvas características de los items, concluye que una reducción del número de alternativas de respuesta por pregunta unida al aumento proporcional de la longitud de la prueba favorece el incremento en la eficiencia de la misma para las personas de nivel alto y perjudica a las personas de menor nivel. Este autor concluye, sin embargo, que se puede reducir el número de alternativas sin afectar el poder de discriminación y el modelo de 3 opciones le parece el mejor, en consonancia con las aproximaciones teóricas de otros autores como Tversky o Grier, pero no dedicó su estudio a confirmar la referida hipótesis.

Bruno y Dirkzwanger (1995), tras contrastar el modelo de la teoría de la información con distintas perspectivas de análisis para esta cuestión, como la que intenta maximizar el poder de discriminación, la de elección en tiempo restringido, la teoría de la respuesta al ítem y la maximización de la fiabilidad por el método de repetición, llegan a la misma conclusión de que 3 opciones por pregunta son mejores que 4 en términos de discriminación y de consistencia interna de la prueba.

Rogers y Harley (1999) insisten en el hecho del aumento de dificultad que supone la redacción del tercer distractor y los siguientes, estimando que las pruebas de 3 opciones son, al menos tan fiables como las de 4 o 5 alternativas. Además, afirman que la reducción de opciones también es buena porque disminuye la probabilidad de acierto por azar, cuestión que será objeto de análisis en la siguiente sección.

Por tanto, parece bastante claro que el modelo de 3 alternativas por pregunta, 1 verdadera y 2 distractores ambos plausibles, es el mejor, por su mayor facilidad de diseño y su mayor posibilidad de control de las respuestas dadas al azar, al mantener el resto de las características psicométricas similares a las de los modelos de 4 o más alternativas.

7.3 Las instrucciones y las estrategias de respuesta

En la sección anterior ya se había señalado la importancia que tiene disponer de unas buenas instrucciones para asegurar que las personas que van a ser examinadas conocen perfectamente qué deben de hacer y qué se espera de ellas.

Pues bien, algunos autores (Prieto y Delgado, 1999; Morales, 1995 y 2006) consideran que el efecto de las instrucciones es determinante en la elección de la estrategia que las personas van a seguir para realizar una prueba, dado que el procedimiento de corrección de la misma es uno de los mayores condicionantes del mecanismo de la decisión que se adoptará para seleccionar la que se considera mejor respuesta entre varias alternativas.

El resultado final de una respuesta es uno de estos tres, un acierto, un error o una omisión, pero a la hora de corregir la prueba el tratamiento a dar a los errores y a las omisiones puede ser muy variado. Budescu y Bar-Hillel (1993) recogen tres de los procedimientos más comunes para la corrección de las pruebas de elección de mejor respuesta, a saber: S1, que consiste en contabilizar los aciertos, valorándolos con 1 punto y atribuir 0 puntos a los errores y omisiones por igual; S2, que intenta controlar las respuestas dadas al azar, penalizando los errores, mediante el descuento correspondiente de la aplicación de la fórmula *A-[1/(k-1)],* siendo *A* el número total de aciertos y *k* el número de opciones de respuesta; S3, que intenta evitar las respuestas dadas al azar, para lo cual, bonifica las omisiones, añadiendo la parte proporcional correspondiente a la puntuación total de los aciertos de conformidad con la aplicación de la fórmula *A+(O/k),* siendo *O* el número total de omisiones.

La problemática ligada a la elección de las distintas estrategias de respuesta, en función del criterio que se haya adoptado *a priori* para la corrección de la prueba e informado a las personas que la van a realizar, ha sido objeto de múltiples estudios y perspectivas, la de las respuestas dadas al azar, la de la consideración o no del valor del aprendizaje parcial, la de la tendencia a adivinar cuando se carece de seguridad sobre la respuesta correcta, o la tendencia a adivinar la respuesta a partir de indicios espurios derivados de los defectos del instrumento.

Prieto y Delgado (1999) adoptan cuatro dimensiones para analizar esta variada problemática, la dimensión estratégica, relacionada con el método más racional que debieran seguir las personas para maximizar sus resultados, la dimensión psicológica, que tiene que ver con la dificultad de comprender todas las implicaciones derivadas de unas determinadas instrucciones a la hora de realizar la prueba, la dimensión psicométrica, la más interesante para los analistas, que está centrada en la características de fiabilidad y validez de la evaluación, y la dimensión educativa, ligada a los procesos cognitivos de respuesta automática o controlada a los diversos estímulos. Combinando los criterios de las cuatro dimensiones en contextos de evaluación de aprendizajes, estos autores optan por el procedimiento S3, de desaconsejar la respuesta al azar y bonificar las omisiones, porque, además de satisfacer mejor las exigencias de la tercera dimensión, la psicométrica, es el más coherente con el control consciente y razonado del aprendizaje. No obstante, admiten, desde la perspectiva psicológica, el valor del procedimiento S1, que considera sólo los aciertos, por ser el más sencillo de explicar y comprender.

La tercera dimensión plantea unas exigencias de mínimos, relativamente independientes del método de corrección, y las medidas más adecuadas en relación con esta dimensión son las propuestas por Cizek, Robinson y O'Day (1998), relativas a la eliminación de las preguntas disfuncionales a partir de un análisis *a priori* del borrador de la prueba con ayuda de un grupo de expertos, y *a posteriori*, pero antes de obtener las calificaciones definitivas, mediante el análisis de items. Estas medidas, con todo, son complejas y escasamente aplicadas en la práctica, con el inconveniente añadido de que la eliminación de preguntas disfuncionales puede alterar la validez de contenido por la escasez de muestra relativa a algún aspecto importante del dominio. La solución que da Bush (2006), quien propone una estrategia de mejora similar, es la creación y alimentación progresiva de un banco con preguntas que satisfagan los distintos criterios del análisis de items.

El procedimiento de corrección para reducir o eliminar las respuestas dadas al azar (S2) es simplista porque atribuye una distribución de la tendencia a adivinar por azar homogénea entre todas las opciones, suponiendo que todas son igualmente atractivas (Budescu y Bar-Hillel, 1993). Estos autores opinan que es muy raro que alguien utilice esta estrategia de respuesta con carácter general ante todas las preguntas cuya respuesta desconoce, ya que el conocimiento sobre un tema es

gradual, desde un conocimiento absoluto hasta un total desconocimiento, pasando por un determinado nivel de conocimiento parcial o inseguro.

Hutchinson (1985), por su parte, ya había puesto en valor el conocimiento parcial o inseguro, dando también argumentos muy consistentes para tenerlo en cuenta. García-Pérez y Frary (1991), desde la perspectiva de la teoría de la respuesta al ítem, plantean la parametrización de las distintas probabilidades de cada respuesta, verdadera, errónea u omitida, en pruebas de elección de alternativas, que tiene en cuenta múltiples variables, como el número de opciones, el distinto grado de atracción de los distractores, etc., donde dejan clara la influencia del conocimiento parcial o inseguro en la estrategia de respuesta.

Rogers y Ping Yang (1996) exponen un modelo para maximizar la elección de las respuestas acertadas, arriesgando en la decisión a favor de alguna de las opciones, que contradice la pretendida homogeneidad del modelo de respuesta emitida por puro azar. Es más, Angoff (1989) ya había dado pruebas de que la tendencia a la adivinación, mediada por el nivel de conocimiento parcial o inseguro, favorece claramente a quienes obtienen los mejores resultados, porque quien más sabe más arriesga y logra una mayor cantidad de aciertos; así como que esta tendencia perjudica a quienes tienen un menor nivel de conocimiento, porque, probablemente, estas personas intenten contestar por azar o utilizando otras estrategias derivadas tanto de los defectos de diseño como de otros criterios independientes y distintos de los que es posible atribuir a los efectos de un conocimiento parcial o inseguro (Rogers y Ping Yang, 1996).

Morales (2006) realiza una síntesis bastante completa de los anteriores planteamientos, combinando los distintos criterios de procedimiento para dar la solución a los dos grandes problemas con las instrucciones que serían más adecuadas en cada caso.

Identifica este autor, por una parte, los problemas de eliminación o reducción de la influencia del azar en la respuesta y la necesidad de tener en cuenta el conocimiento parcial o inseguro, con independencia del contexto y los objetivos del examen. Por otra parte, asocia a los distintos procedimientos de control del azar, S2 y S3, las instrucciones que orientarán las distintas estrategias de respuesta en estas tres direcciones: a) Abstenerse y no responder en caso de duda para evitar la penalización correspondiente, (S2), que es el criterio más extendido; b) Contestar lo más probable si se puede eliminar con seguridad alguna de las opciones falsas y si no abstenerse (S3); c) Contestar lo más probable, aunque no se pueda eliminar con seguridad ninguna opción falsa (S1), porque cuando se cree que se tiene un buen conocimiento de la materia la probabilidad de acertar es bastante alta.

Entiende Morales (2006) que es imposible dar instrucciones que beneficien a todo el mundo, porque, por ejemplo, como ya sabemos, los criterios que tienen en cuenta el conocimiento parcial o inseguro (S1) favorecen a las personas más competentes. Es más, animar a responder cuando se carece de seguridad para eliminar las opciones falsas puede suponer un problema de tipo ético, a sabiendas de que esta recomendación es perjudicial para las personas menos dotadas.

Las discrepancias entre optar por evitar la influencia del azar o permitir los efectos del conocimiento parcial tienen que ver con el control de los distintos tipos de sesgos. Quienes defienden la oportunidad de contabilizar únicamente las respuestas acertadas centran su atención en la reducción de los errores sistemáticos, al estar comprobado el valor positivo del conocimiento parcial o inseguro, mientras a quienes se interesan por aminorar los efectos del azar les preocupan más los errores de tipo aleatorio (Morales, 2006).

Por último, a pesar de que Morales opina que todavía no hay conclusiones definitivas sobre estas cuestiones, se inclina por recomendar la elección de la respuesta más probable y la aplicación del procedimiento S1, si bien para un mejor tratamiento del conocimiento parcial o inseguro propone otros procedimientos diferentes a los tres analizados, como serían la eliminación de todas las respuestas probablemente falsas, sin necesidad de cualquier otra elección, o la elección de todas las respuestas probablemente verdaderas. Pero estos procedimientos obligan a adoptar otro formato de prueba distinto a la elección de una alternativa unívoca.

8. La calificación de las pruebas objetivas

La última operación de la cuarta fase del uso de pruebas objetivas, correspondiente a la evaluación y obtención de los resultados, es el establecimiento de los criterios y el procedimiento de calificación de la prueba, que se aplicarán después de haberla corregido y realizado las operaciones correspondientes, conforme a lo determinado en la tabla de especificaciones.

En general, suele ser necesario distinguir entre una puntuación directa y una calificación definitiva, por múltiples razones. Por ejemplo, es probable que el número total de preguntas sea variable, mientras que la escala de presentación de los resultados tenga un formato estandarizado para varias pruebas que no coincide con el total de preguntas de cada prueba. Otra razón puede ser que el nivel de exigencia mínimo sea variable, en función de la mayor o menor dificultad estimada de los contenidos de la prueba, cuando en la escala de presentación de los resultados el referido nivel de exigencia mínimo se sitúa normalmente en el punto medio. También pudiera ser que las distintas partes de la prueba pudieran tener una distinta ponderación, que habrá quedado representada en la tabla de especificaciones. Situaciones como éstas obligan a un emparejamiento con criterios de proporcionalidad entre la escala de puntuaciones directas de la prueba y la escala de presentación de las calificaciones definitivas.

Los dos principales problemas a resolver, antes de realizar la calificación definitiva de una prueba, son la determinación del nivel de exigencia mínimo, de donde se deriva la estimación de un punto de corte en la escala de puntuaciones directas y la elección del criterio de proporcionalidad directa con respecto a la escala de presentación de los resultados. La solución más sencilla para ambos problemas es determinar que el grado de exigencia mínimo corresponde con el punto medio de la escala de puntuaciones directas, y, por tanto, también con el punto medio en la escala de presentación de los resultados.

Ahora bien, cuando es necesario utilizar otro criterio o se emplea la norma de grupo para establecer los resultados definitivos, a estos dos problemas hay que darles solución por separado.

8.1 Determinación del nivel mínimo de exigencia

Si se utiliza la norma de grupo, este nivel mínimo estará relacionado con cualquiera de las medidas de tendencia central, la media aritmética, la mediana o la moda, o con una combinación de éstas y alguna otra de dispersión.

En las pruebas que siguen únicamente el modelo de evaluación por criterio, el nivel mínimo de exigencia habrá de ser coherente con la previsión provisional estimada al principio del diseño. En una prueba de dificultad media el nivel mínimo, lógicamente, se situará en el valor medio de la escala y en cualquier otra circunstancia oscilará ligeramente hacia una exigencia superior para aumentar la dificultad o inferior para disminuirla.

La determinación del nivel mínimo de exigencia de las pruebas de elección de alternativa, a su vez, es un momento apropiado para tratar de paliar la influencia de las respuestas dadas al azar, considerando también el valor del conocimiento parcial o inseguro. En este caso únicamente se tendrían en cuenta los aciertos (S1). De esta manera, una prueba de dificultad media, cuyo punto de corte inicialmente se situaría en el 50% del valor de la escala de puntuaciones directas, exigiría aumentar el punto de corte hasta el 62'5%, si son 4 las alternativas de respuesta por pregunta y el 66'6% en las de 3 alternativas. Este intento de neutralización se basa en que las probabilidades de acierto por puro azar son 0,25 y 0,33 para 4 y 3 alternativas de respuesta por pregunta, respectivamente, que se transforman en el incremento correspondiente sobre el 50%.

8.2 Correspondencia entre las escalas de puntuación

La proporcionalidad de dos escalas diferentes, en las que sus puntos medios señalen el nivel mínimo de exigencia, es directa entre cualquiera de sus puntos. Pero toda variación, por mínima que sea, en el sentido de aumentar o disminuir el nivel de exigencia de una prueba altera el carácter biunívoco de esta relación.

De este segundo supuesto se deducen dos escalas diferentes: la escala *"e"* de presentación de resultados cuyo mínimo de exigencia coincide con su valor medio (min_e) y la escala *"i"* de puntuaciones directas de la prueba, con un punto de corte (PC_i) situado en cualquier valor superior o inferior al 50% de la misma y una puntuación máxima concreta (MAX_i). Además, cada persona que ha superado la prueba obtiene una puntuación X_i en la misma. La operación de traslación de los puntos de una a otra escala exige también conocer la diferencia en términos absolutos entre el punto de corte y la máxima puntuación alcanzable ($df_{PCi\text{-}MAXi}$).

Así pues, la fórmula 1 para la traslación de las puntuaciones directas obtenidas a la escala de puntuación final (PF), aplicable en los casos en que se considere una prueba superada es la siguiente:

$$PF = min_e + ((Xi - PC_i) * min_e / df_{PCi\text{-}MAXi})$$

Consecuentemente, para el tramo de puntuaciones inferiores habrá que adaptar la fórmula anterior respecto a la diferencia entre la puntuación mínima posible y el punto de corte.

Ejemplo: Una prueba de 77 preguntas de elección entre 3 alternativas tiene su nivel mínimo de exigencia establecido en 51 aciertos. La escala de presentación de resultados es de 0 a 10 puntos con un mínimo de 5. La puntuación directa obtenida por una persona concreta son 59 puntos. ¿Cuál es su calificación definitiva?

PF=5 + (((59-51)(5/(77-51))) Resultado: 6,53 puntos sobre el total de 10.*

Este procedimiento es útil para cualquier tipo de prueba, con independencia de su formato, cuando el punto de corte de la misma no coincida exactamente con el 50% del valor máximo alcanzable en la misma y la escala de presentación de las calificaciones sea otra distinta. No obstante, su empleo es más propio de las pruebas de elección de alternativa que de las de respuesta breve. La razón para esta aplicación diferencial reside en que cualquier alteración sobrevenida, como puede ser la eliminación de una o varias preguntas disfuncionales de la propuesta, altera la proporcionalidad directa prevista entre las dos escalas y obliga a utilizar algún tipo de traslación como el anterior.

9. Verificación del grado de idoneidad de las pruebas objetivas

Una labor, relativamente compleja, pero ligada a la última fase del diseño de una prueba objetiva es argumentar y aportar la información necesaria que asegure la calidad de la misma (Bush, 2006).

Este quehacer tiene que ver con el concepto psicométrico de validez o demostración de que un instrumento de medida mide lo que afirma medir.

La validez científica se deriva de un modelo en el cual los resultados de una prueba están relacionados con el constructo que asegura probar, explicando tales relaciones mediante distintas formas de inferencias, como la capacidad de predicción de distintos eventos conforme a determinados criterios, la relevancia y representatividad del contenido, etc. (Binning y Barret, 1989).

La validez de contenido es la inferencia básica inicial en el caso del diseño de las pruebas objetivas que se realizan para comprobar el grado de conocimiento de alguien sobre un dominio o área concreta. Para asegurar la calidad de una prueba de elección de alternativas (Bush, 2006), la estrategia de dos pasos, como son el análisis previo a su utilización de los enunciados que configuran el borrador de la prueba y el análisis de items sobre los datos recogidos tras su aplicación posterior (Cizeck, Robinson y O'Day, 1998) es la más efectiva.

Esta estrategia de análisis previo y análisis posterior es de aplicación análoga a las pruebas de respuesta breve, con sus respuestas patrón y especificidades correspondientes y su rigor y grado de necesidad están determinados por la importancia de las consecuencias que se derivan de su uso. Cuando, por ejemplo, la prueba objetiva es un ejercicio más dentro de una multiplicidad de pruebas diferentes de evaluación continua del aprendizaje, habrá que preguntarse por la eficiencia de un proceso de validación como éste. No obstante, en procesos de naturaleza competitiva, como selección de personal o reconocimiento de la competencia, la necesidad de validación de los instrumentos tienen una mayor justificación.

9.1 Revisión previa al uso del borrador del cuestionario

Para proceder a una revisión del borrador de un cuestionario de preguntas de respuesta breve o de elección de alternativa la hipótesis de partida más razonable es considerar la posible presencia de defectos o errores, que son de corrección inmediata.

El procedimiento consiste en utilizar una lista de control similar a la del anexo 1, que se ha realizado con ayuda de diversas guías existentes (Haladyna, Downing y Rodríguez, 2002; Muñiz y García-Mendoza, 2002; Moreno, Martínez y Muñiz, 2006; Case y Swanson, 2006).

La misma persona responsable del diseño puede revisar su propuesta inicialmente, pero un procedimiento más riguroso es el que realice un equipo de 3 personas expertas independientes, primero de manera individual, sin necesidad de tomar ninguna decisión de cambios y luego de forma colectiva, considerando aquellas propuestas de modificación sobre las que se hayan puesto de acuerdo.

Para evaluar las posibles discrepancias y acuerdos sobre las propuestas de mejora conviene utilizar algún método de validez de contenido (Lawshe, 1975).

9.2 Análisis posterior de items

Un análisis de items es un estudio de las características del cuestionario, por ver si responde a los parámetros de dificultad, discriminación, estructura de los

distractores, y fiabilidad, validez y dimensionalidad del cuestionario en su conjunto (Muñiz, Fidalgo, García-Cueto, Martínez y Moreno, 2005).

Hay dos grandes modelos y diversas técnicas estadísticas para analizar las preguntas y las respuestas de un cuestionario como éste, uno es el enfoque clásico y otro es la teoría de respuesta al ítem, que profundiza mucho más en sus análisis, al considerar cada pregunta como la unidad básica de estudio. No obstante, cualquier técnica de éstas permite mantener aquellas preguntas y sus opciones que cumplen los criterios establecidos en cada parámetro y rechazar o reformular algunas otras, si es que son objeto de mejora.

Siguiendo el enfoque clásico (Muñiz, Fidalgo, García-Cueto, Martínez y Moreno, 2005) a continuación se resumen los principales elementos del análisis de items.

a) El índice de dificultad de una pregunta se obtiene dividiendo el número de personas que la han acertado por el número total de personas que la hayan respondido. En realidad, es más bien un índice de facilidad, ya que las preguntas más difíciles son las que únicamente las acierta menos de un 20%, o sea en una proporción entre 0,1 y 0,3, mientras que las menos difíciles son las que aciertan en un intervalo entre 0,8 y 1. Un nivel de dificultad apropiado estriba entre 0,4 y 0,6. Para reducir la influencia de la elección por azar, siempre es conveniente obtener los índices de dificultad corregidos. Esta operación se realiza restando de la proporción de aciertos (p) el cociente de dividir la proporción de fallos (q) entre la diferencia del número de alternativas (k) menos 1. La fórmula es: ID= p — (q**/**(k-1)).

b) Tan importante como la dificultad o más, desde la perspectiva del diseño de cuestionarios, es el poder de las preguntas para diferenciar entre las personas que obtienen unos rendimientos muy altos y muy bajos en la prueba, lo que se hace mediante el índice de discriminación. El índice de discriminación de una pregunta es la diferencia entre dos proporciones, la de acertantes en el 27% del total de mejores rendimientos y la del 27% de los de menor rendimiento. Un buen índice de discriminación tiene que ser igual o mayor a 0,3. Las preguntas con índices inferiores es preferible no utilizarlas o mejorarlas.

c) Otro de los indicadores de la validez de las preguntas es el grado de atracción de cada una de las opciones de respuestas, tanto la verdadera, como los distractores. Este indicador se averigua mediante una prueba de independencia (Chi^2), de resultado no significativo para todos los distractores, lo que quiere decir que la elección de cada uno de ellos se ha distribuido más o menos por igual. Cuando la prueba de independencia es significativa quiere decir que alguna de las opciones se ha elegido más de lo debido por alguna razón, con lo cual, debería rechazarse la pregunta o mejorar los distractores, salvo cuando se trate de la opción verdadera.

Las preguntas preferibles para ser eliminadas son las que presenten valores inferiores del índice de discriminación, en primer lugar, aquellas que comprometen la fiabilidad de la prueba en segundo lugar y, por último las que son extremadamente fáciles o difíciles.

Las operaciones de análisis de items pueden automatizarse, de modo que quepa la posibilidad de eliminar las preguntas disfuncionales, incluso, antes de obtener la relación de puntuaciones directas. El único cuidado que debe tener el equipo de personas expertas es que la reducción del cuestionario no afecte de manera significativa a la representatividad del dominio de conocimiento o a un área concreta del mismo.

10. Conclusiones

Por medio de este estudio se han recopilado y analizado los problemas, recursos metodológicos y criterios principales del diseño de las pruebas de respuesta breve y de elección de una alternativa entre varias opciones de respuesta. No obstante, muchas de las orientaciones y medidas consideradas son de aplicación también en el resto de los formatos de pruebas objetivas.

De todo este ensayo cabe destacar las siguientes conclusiones:

I. Se ha puesto de manifiesto el valor de las pruebas de respuesta breve por su virtualidad para activar los procesos cognitivos de evocación de recuerdos, juicio o elaboración particular de la respuesta y para centrar la atención en las dos principales condiciones de la validez de contenido, su relevancia y representatividad respecto a un dominio de conocimiento.

II. De manera análoga, el tratamiento dado a las pruebas de respuesta alternativa pone el acento en la importancia debida al formato del instrumento de evaluación por su capacidad de condicionar los resultados, con independencia de la selección del contenido.

III. Ambos tipos de pruebas son de distinta naturaleza, ya que la de respuesta alternativa prioriza los procesos cognitivos de atención y memoria de reconocimiento sobre los demás. Pero como los dos formatos son igualmente apropiados para la evaluación de objetivos similares puede decirse que son, por una parte, simétricos respecto de los procesos cognitivos de memoria que intervienen y, por otra, poseen una cierta continuidad procesal, porque el primer paso en el diseño de cualquier prueba objetiva debe de ser la elaboración de una prueba de respuesta breve.

IV. Las verdaderas particularidades de estos dos tipos de pruebas inciden sobre las diferentes fases del proceso de diseño, ejecución y evaluación de las mismas. Así, en las de elección de alternativa la complejidad reside en su diseño, y, concretamente, en la tercera fase que corresponde a la redacción de las opciones de respuesta, mientras que en las de respuesta breve el esfuerzo se traslada a las operaciones de corrección, dentro de la fase de evaluación, sobre todo si se han descuidado las dos primeras de análisis del contexto y selección de los contenidos. Únicamente la automatización de los procesos de ejecución, corrección y evaluación justifican el empleo masivo de las pruebas de elección de alternativa en detrimento de las de respuesta breve de carácter abierto, sin que se entre a ponderar las ventajas e inconvenientes de una y de otra.

V. En las pruebas de elección de alternativa el número ideal de opciones es 3. Esta es la conclusión a la que llegan investigaciones realizadas desde planteamientos muy diversos. Ahora bien, esta medida va a exigir un cierto esfuerzo de cambio cultural, ante la creencia extendida de que éste reducido número de opciones disminuye el grado de dificultad de la prueba. En este sentido, parece claro que para aumentar la dificultad del examen es preferible aumentar el número de preguntas en lugar del número de opciones por pregunta.

VI. Respecto a la disyuntiva entre reducir la influencia del azar en las respuestas o aprovechar el conocimiento parcial como mecanismo para la elección de las opciones, si bien no existe una posición tan definitiva

como la referida al número de alternativas, se sugiere utilizar este conocimiento parcial, sin necesidad alguna de penalizar los errores ni bonificar las respuestas omitidas, porque son las personas más competentes las que utilizan con mayor eficacia tal estrategia. Una forma de minimizar la influencia del azar podría ser aumentar el nivel mínimo de exigencia.

VII. Finalmente, se ha hecho hincapié en la necesidad de proceder a la revisión sistemática de los borradores de las pruebas y al análisis de items, como procedimientos para validar el contenido en los dos tipos exámenes.

Referencias

Angoff, W.R.(1989). Does guessin really help? *Journal of Educational Measurement, 26,* 323-336.

Basoredo L., C. (2008). El examen de desarrollo escrito, sus tipos y sus procedimientos de diseño y evaluación. *Quaderns Digitals, 55.*

Basoredo L., C. (2009). ¿Cómo formular objetivos para el aprendizaje y el desarrollo de competencias? *Quaderns Digitals, 58.*

Basoredo L., C. (2010). Herramientas de análisis de contenido de utilidad en los ámbitos del aprendizaje y la educación. *Quaderns Digitals, 61.*

Binning, J.F. & Barret, G.V. (1989). Validity of Personnel Decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74(3),* 478-494.

Bleske-Rechek, A.; Zeug, N. & Webb, R.M.(2007). Discrepant performance on multplice-choice and short answer assessments and the relation of performance to General Scholastic Aptitude. *Assessment & Evaluation in Higher Education, 32(2),* 89-105.

Bruno, J.E. & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55(6),* 959-966.

Budescu, D. & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoric view of formula scoring. *Journal of Educational Measurement, 30(4),* 277-291.

Burton, R.F.(2006). Sampling knowledge and understanding: Hoe long should a test be? *Assessment & Evaluation in Higher Education, 31(5),* 569-582.

Bush, M.E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education, 14(4),* 398-404.

Case, S.M, Swanson, D.B.& National Board of Medical Examiners (2006). *¿Cómo elaborar preguntas para evaluaciones escritas en el área de ciencias básicas y clínicas?* Philadelphia: NBME.

Cizeck, G.; Robinson, K.L. & O'Day, D.M.(1998). Non functioning options: a closer look. *Educational and Psychological Measurement, 58(4),* 605-611.

Davis R.A.(1967). Short answer and essay examinations. *Peabody Journal of Educations, 44(5),* 269-271.

Díaz-Barriga, F. (2003). Cognición situada y estrategias para el aprendizaje significativo. *Revista Electrónica de Investigación Educativa- REDIE, 5(2).*

Dolinsky D. & Reid., V.E. (1984). Types of classroom tests: Objective cognitive measures. *American Journal of Pharamceutical Education, 48(3),* 285-290.

Ebel, R.L.(1982). Proposed solutions to two problem of test construction. *Journal of Educational Measurement, 19,* 267-278.

Ellington, H. (1987). *Short Answer Questions, teaching and learning in Higher Education.* Alberdeen (Scotalnd): RGIT-CICED.

Frary, R.B.(1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation, 4(11).*

Fuhrman, M. (1996). Developing good multiple-choice tests and test question. *Journal og Geoscience Education, 44,* 379-384.

García-Pérez, M.A. & Frary, R.B. (1991, abril). *Item Characteristic Curves: A new theoretical aproach.* Comunicación presentada en el Annual Educational Research Association, Chicago.

Green B. F.(1978). In defense of measurement. *American Psychologist, 33(7),* 664-670.

Haladyna, T.M., Downing, S.M. & Rodríguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15(3),* 309-334.

Hernández-Nodarse, M. (2007). Perfeccionando los exámenes escritos: reflexiones y sugerencias metodológicas. *Revista Iberoamerican de Educación, 41(4),* 1-25.

Hutchinson, T.P. (1985, julio). *Predictin performance in variants of the multiple-choice test.* Comunicación presentada en el Annual Meeting of the Psichometric Society and Classification Societies, Cambridge, UK.

Jonassen, D. & Tessmer, M. (1996). An outcomes-based taxonomy for instructional systems design, evaluation and research. *Training Research Journal, 2,* 11-46.

Jordan, S. & Mitchell, T. (2009). e-Assessment for learning? The potential of short answer free-text questions with tailored feedback. *British Journal of Educational Technology, 40(2),* 371-385.

Jornet M., J.M. & Suárez R., J.M.(1996). Pruebas estandarizadas y evaluación del rendimiento: usos y características. *Revista de Investigación Educativa, 14(2),* 141-163.

Lafourcade, P. (1985). *Evaluación de los aprendizajes.* Madrid: Cincel.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28,* 563-575.

Lord, F.M.(1977). Optimal number of choices per items: A comparison of four approaches. *Journal of Educational Measurement, 14,* 33-38.

Morales, P. (1995). *Las pruebas objetivas. Cuadernos monográficos del ICE, núm. 4.* Bilbao: Universidad de Deusto.

Morales, P. (2006). *Las pruebas objetivas: normas, modalidades y cuestiones discutidas.* Universidad Pontificia de Comillas: Facultad de Ciencias Humanas y Sociales. Madrid.

Ministerio de Educación (1971). *Tabla de especificaciones de pruebas objetivas en la Enseñanza Primaria, para las asignaturas de Estudios Sociales, Idioma Español, Matemáticas y Ciencias Naturales.* Guatemala: José Pineda Ibarra.

Moreno, R., Martínez, R. & Muñiz, J. (2004). Directrices para la construcción de items de elección múltiple. *Psicothema, 16(3),* 490-497.

Muñiz, J., Fidalgo, A.M., García-Cueto, E., Martínez, R. & Moreno, R. (2005). *Análisis de items.* Madrid: La Muralla.

Muñiz, J. & García-Mendoza, A. (2002). La construcción de items de elección múltiple. *Metodología de las Ciencias del Comportamiento (monográfico),* 416-422.

Parsons, J. & Fenwick, T. (1999). Using objective tests to evaluate. *Guides, classroom, teacher.* Department of Secondary Education. University of Alberta.

Pelechano B., V. (1988). *Del psicodiagnóstico clásico al análisis ecopsicológico. Vol. II.* Valencia: Alfaplús.

Prieto, G. & Delgado A.R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment, 15(2),* 143-150.

Roback, A.A.(1921). Subjective tests versus objective tests. *Journal of Educational Psychology, 12(8),* 439-444.

Rodríguez, M.(2005). Three options are optimal for multiple-choice items: A meta-anaysis of 80 years of research. *Educational Measurement: Issues and Practice, 24(2),* 3-13.

Rogers, W.T. & Harley, D.(1999). An empirical comparison of three and four-choice items and tests: susceptibility to test-wiseness and internal consistency reliability. *Educational and Psychological Measurement, 59(2),* 234-247.

Rogers, W.T. & Yang, P.(1996). Test-Wiseness: Its nature and application. *European Journal of Psychological Assessment, 12(3),* 247-259.

Soubirón, E. y Camarano, S. (2006). *Diseño de pruebas objetivas.* Facultad de Química de la Unversidad de la República de Uruguay.

Tenbrick, T.D. (1983). *Evaluación, guía práctica para profesores.* Madrid: Narcea.

Thyne, J.M. (1978). *Principios y técnicas de examen.* Madrid: Anaya.

Anexo 1: Lista de control para la revisión de pruebas

I. Análisis general del cuestionario			
Nº		SI	NO ¿?
1.-	¿Están suficientemente claros los objetivos a los que responde esta prueba?		
2.-	¿Se ha realizado una tabla de especificaciones para la selección de las preguntas y su distribución?		
3.-	¿Se ha redactado un cuestionario guía de preguntas directas con su respuesta correspondiente?		
4.-	¿Se han elaborado los argumentos o elegido las referencias que justifican la bondad o exactitud de las respuestas?		
5.-	¿Cada pregunta incluye un contenido específico, importante e independiente de las demás preguntas?		
6.-	¿Hay preguntas sobre todos los aspectos importantes, exceptuándose datos triviales o excesivamente específicos?		
7.-	¿Se han identificado suficientes preguntas fáciles, de dificultad media, y difíciles, conforme al colectivo a examinar?		
8.-	¿Se han ordenado las preguntas siguiendo algún criterio lógico?		
9.-	¿Están las respuestas correctas situadas en distinta posición, según una distribución homogénea de todas las opciones?		
10.-	¿Todos los enunciados de las preguntas y las opciones de respuesta se entienden fácilmente en la primera lectura?		

II. Análisis de las instrucciones			
11.-	¿Se han programado instrucciones para dar a conocer por escrito y explicar, a su vez, *de viva voz*?		
12.-	¿Las instrucciones incluyen alusiones oportunas sobre el contenido y el formato del examen?		
13.-	¿Se indica lo que ha de hacerse con el contenido de la prueba?		
14.-	¿Se describe el mecanismo de respuesta?		
15.-	¿Se describe el mecanismo de corrección y de calificación del cuestionario?		
16.-	¿Se ha previsto la duración de la sesión de examen a razón de 1 a 2 minutos por pregunta?		

III. Análisis de cada pregunta y sus opciones de respuesta

17.- ¿Se nota claramente qué se pretende evaluar mediante la pregunta?

18.- ¿Contiene el enunciado la información principal que es estrictamente necesaria?

19.- ¿El formato es una pregunta directa o un enunciado afirmativo?

20.- ¿La redacción incluye alguna negación que pueda ser evitada?

21.- ¿Se utilizan expresiones de carácter absoluto como "todos", "ninguno", "nunca", etc.?

22.- ¿Se utiliza la copia literal de términos o expresiones directamente extraídos de textos o libros?

23.- ¿Hay algún defecto de ortografía o concordancia gramatical entre el enunciado de pregunta y las opciones de respuesta?

24.- ¿Se ha utilizado alguna expresión que puedan resultar ofensiva para un determinado subgrupo de la población?

25.- ¿Se ha confirmado y argumentado que existe una única opción de respuesta válida o mejor que el resto?

26.- ¿Son las opciones de respuesta independientes entre si?

27.- ¿Todas las opciones tienen una estructura homogénea por contenido y formato, además de una longitud similar?

28.- ¿Se puede descartar por lógica alguna de las opciones de respuesta, porque es imposible?

29.- ¿Alguna de las opciones es imprecisa o induce a engaño?

30.- ¿La combinación de términos o repetición de éstos en el enunciado y en las opciones ofrece posibilidades de acierto?