Análisis de los ítems

Albert Bonillo

PID_00198631

Universitat Oberta de Catalunya

Índice

Introducción

El objetivo de este módulo es introducir al estudiante en el tema del análisis de ítems. Consideramos un error que este tema no esté presente en todos los planes docentes de la asignatura de *Psicometría*, ya que es suficientemente importante para que valga la pena tratarlo. Es cierto que, tradicionalmente, el estudio de la psicometría se ha focalizado más hacia las propiedades de los instrumentos de medida, que preguntan sobre aspectos opinativos y constructos psicológicos, que hacia los instrumentos que miden conocimiento o habilidad. Sin embargo, el psicólogo de a pie trabaja tanto o más con los segundos que con los primeros.

Queremos que el estudiante sepa, desde el principio, que este módulo no lo hará experto en ninguno de los aspectos que en él se tratan. Le proporcionará, deseamos y esperamos, una buena introducción a cada uno de los temas, pero es fácil que por placer –o necesidad profesional, o ambas– necesite profundizar en algunos de los aspectos tratados. Recomendaremos textos que sí los traten en profundidad.

El módulo se inicia precisamente distinguiendo entre instrumentos en función de su objetivo. En segundo lugar, y ya centrados en pruebas de ejecución máxima, mostraremos cuáles son los aspectos que hay que tener en cuenta en la construcción de sus ítems. En el tercer apartado veremos cómo analizar las propiedades psicométricas de la prueba y de los ítems a partir de la teoría clásica de test (TCT). Veremos los conceptos de dificultad y discriminación, y aprenderemos a valorar si un ítem es correcto o quizá necesita una revisión. En el cuarto apartado veremos una introducción a la teoría de respuesta al ítem (TRI), que es una alternativa de análisis a la TCT. Veremos la TRI de manera más sucinta que la TCT. El modelo de TRI resuelve problemas teóricos de la TCT, pero los cálculos de esta son más sencillos y fácilmente aplicables que los de aquella.

Dejaremos para el final las conclusiones que resuman todo lo presentado.

1. Tipos de pruebas

Es tradicional que, cuando desde el ámbito de la psicología hablamos de una prueba –o de un instrumento de medida– pensemos de inmediato en una encuesta de opinión, un test de personalidad o similar. Desde el punto de vista del tipo de prueba, estas que hemos mencionado no son distintas del cuestionario de satisfacción sobre el servicio que encontramos a la salida de muchos hoteles. Pretenden medir, en una persona, el valor determinado de un constructo cuya existencia se presupone.

Caso distinto es una prueba que pretenda ordenar a los mejores candidatos a un puesto de trabajo. En este contexto, donde es de suponer que existe un criterio –ser un buen trabajador para el puesto ofertado–, la medida del constructo puede pasar a un segundo plano. El objetivo del instrumento es que cada uno de los ítems optimice la correcta clasificación de las personas. Veamos, pues, qué características tienen las pruebas en función de lo que pretenden.

1.1. Pruebas de ejecución típica frente a pruebas de ejecución máxima

Si clasificamos las pruebas por su objetivo, distinguiremos entre dos tipos básicos. Denominamos pruebas de ejecución típica –o de ejecución de rasgos– a aquellas que miden aspectos no escalables, o dicho de otra manera, a aquellas cuyas preguntas no tienen respuestas correctas ni erróneas, sino que se trata de aspectos de opinión, de preferencia o similar. Por el contrario, llamamos pruebas de ejecución máxima a aquellas que evalúan constructos que sí son escalables, y que son aquellos en los que tiene sentido hablar de respuestas correctas y erróneas. Un examen, un test de inteligencia o cualquier instrumento que mida aptitud sería clasificado dentro de este epígrafe.

Aunque todos los conceptos que hemos visto hasta ahora en módulos anteriores –fiabilidad, validez y transformación de las puntuaciones obtenidas– son aplicables a ambos tipos de instrumentos, las estrategias para su estudio suelen variar ligeramente y se suelen estudiar aplicándolos a las pruebas de ejecución típica. Es cierto que, por ejemplo, un test de inteligencia debe ser fiable, pero puede no tener demasiado sentido administrarlo dos veces en unas pocas semanas, ya que los participantes podrían haber obtenido la respuesta correcta en el tiempo transcurrido y contaminar así los resultados. Sin embargo, sí tiene sentido repetir un test de personalidad con pocos días de diferencia y comprobar de ese modo si la medida del instrumento es tan estable como se supone que es el constructo medido. En definitiva, las características que se deben estudiar dependen, cómo no, del objetivo del instrumento.

En muchas ocasiones, el psicólogo profesional no utiliza instrumentos estandarizados, sino que debe crear él mismo el instrumento. Si el estudiante trabajara en el departamento de recursos humanos de una multinacional y esta le pidiera una prueba para ocupar un puesto muy específico, ¿qué haría? Tras comprobar que esta prueba no existe en el mercado debería crearla. Y debería hacerlo teniendo en cuenta qué se pretende hacer con esa prueba: seleccionar al mejor trabajador para ese puesto. ¿Y a partir de ahí? Supongamos que ese puesto requiere ciertos conocimientos. El psicólogo debería construir una prueba que, a partir del número mínimo de ítems, pueda seleccionar al mejor de los candidatos.

Aprendamos, pues, qué debe tenerse en cuenta cuando (no) hay (más remedio) que crear una prueba.

2. Directivas en la construcción de ítems

El estudiante ya sabe que el objetivo principal de este módulo es mostrar cómo medir la calidad de un test de rendimiento. Ahora bien, no debemos eludir explicar qué hay que hacer para construir correctamente una prueba. Creemos que el trabajo de un psicólogo no debe ser únicamente valorar si la prueba está mejor o peor hecha, sino que también tiene mucho que aportar en su construcción. Cambiando totalmente de ámbito: ¿no sería extraño que un arquitecto valorara edificios si no aprendiera primero a construirlos?

Existen varios trabajos que exponen de manera muy exhaustiva cuáles son las directivas que hay que seguir para construir correctamente una prueba de ejecución máxima. Uno de los primeros y más conocidos es el de Haladyna, Downing y Rodríguez (2002). Se trata, ni más ni menos, que de 31 criterios que seguir, clasificados por apartados. Estos criterios se refieren al contenido de la pregunta (por ejemplo, cada ítem debe medir un único conocimiento), al formato, al estilo (recomienda ítems cortos), al enunciado (tienen que evitar las negaciones) y a las opciones de respuesta (recomienda evitar la opción "Todas las anteriores son correctas/incorrectas").

Personalmente, preferimos los criterios de Moreno, Martínez y Muñiz (2004). Son menos (doce), son mucho más claros y más fáciles de aplicar. Como podéis ver en la tabla 1, ahora los aspectos que hay que valorar son tres: elección del contenido, su expresión y opciones de respuesta.

Tabla 1. Nuevas directrices para la construcción de ítems de elección múltiple

A. Elección del contenido que se desea evaluar
1. Debe ser una muestra representativa del contenido recogido en una tabla de especificación, evitando ítems triviales. 2. La representatividad deberá marcar lo sencillo o complejo, concreto o abstracto, memorístico o de razonamiento que deba ser el ítem, así como el modo de expresarlo.
B. Expresión del contenido en el ítem
3. Lo central debe expresarse en el enunciado. Cada opción es un complemento que debe concordar gramaticalmente con el enunciado. 4. La sintaxis o estructura gramatical debe ser correcta. Evitar ítems demasiado escuetos o profusos, ambiguos o confusos, cuidando además las expresiones negativas. 5. La semántica debe ajustarse al contenido y a las personas evaluadas.
C. Construcción de las opciones

Fuente: Tomado de Moreno, Martínez y Muñiz (2004)

6. La opción correcta debe ser solo una, acompañada por distractoras plausibles.
7. La opción correcta debe estar repartida entre las distintas ubicaciones.
8. Las opciones deben ser preferiblemente tres.
9. Las opciones deben presentarse usualmente en vertical.
10. El conjunto de opciones de cada ítem debe aparecer estructurado.
11. Las opciones deben ser autónomas entre sí, sin solaparse ni referirse unas a otras. Por ello, deben evitarse las opciones "Todas las anteriores" y "Ninguna de las anteriores".
12. Ninguna opción debe destacar del resto ni en contenido ni en apariencia.

Fuente: Tomado de Moreno, Martínez y Muñiz (2004)

En el contenido, deben preguntarse cosas fundamentales. Parece obvio, pero ¿cuántos exámenes recordamos en los que se nos preguntaban algunas cuestiones que aparecieron poco (o nada) en clase? Una prueba debería contener solo (pero todos) los conceptos fundamentales de la asignatura que valora. La creencia de que al preguntar cuestiones menores, en el fondo, estamos obligando al alumno a estudiar toda la materia es absurda y favorece el azar. Respecto al azar, recordad a aquellos alumnos que solo estudiaban medio programa –o menos– y confiaban en tener suerte el día del examen.

Sobre la expresión, las tres cuestiones apuntadas son obvias, pero de nuevo no siempre se cumplen.

Un ejemplo paradigmático de Moreno, Martínez y Muñiz (2004) muestra que es mejor redactar este ítem:

En física se denomina sublimación a un cambio de materia:

1. Sólida a gaseosa.
2. Líquida a sólida.
3. Gaseosa a líquida.

que este:

En física, sublimación:

1. Supone un cambio de materia sólida a materia gaseosa.
2. Se refiere a un cambio de materia líquida a materia sólida.
3. Consiste en un cambio de materia gaseosa a materia líquida.

Sobre las opciones de respuesta, destacaremos la recomendación de que las opciones sean independientes entre sí, lo que automáticamente conlleva no usar los célebres "Todas/Ninguna de las anteriores". Es obvio que para rechazar una opción como "Todas las anteriores son correctas" solo necesitamos saber que una de las otras opciones no lo es. Así, de un plumazo, podemos eliminar dos opciones de las posibilidades y la elección se facilita mucho. Si el test tiene tres opciones, ya conocemos la respuesta, y si tiene cuatro, incluso podemos arriesgarnos a contestar al azar entre las dos restantes.

Para saber más

Si deseáis profundizar en este tema, os recomendamos acudir al texto original de Moreno, Martínez y Muñiz (2004), en el que, en un tono muy didáctico y con ejemplos muy accesibles, encontraréis una explicación muy exhaustiva de cada uno de los criterios.

Ejemplo de prueba de ejecución máxima

A partir de las directrices mostradas, y para ilustrar con un ejemplo concreto y cercano los conceptos que se presentarán en este módulo, hemos construido el siguiente examen. Contiene diez preguntas sobre este mismo módulo y la opción correcta está resaltada en negrita.

1. La dificultad (ID) es un índice que indica la probabilidad de…

A. acertarlo.
B. fallarlo.
C. contestarlo.

2. El valor de discriminación de un ítem (ID) debe ser…

A. negativo.
B. distinto de 0.
C. positivo.

3. Un distractor debería tener discriminación…

A. positiva.
B. negativa.
C. cercana a 0.

4. Un test de personalidad es una prueba de...

A. ejecución máxima.
B. ejecución típica.
C. rendimiento.

5. La fórmula para calcular *IDc* es...

A. $\frac{A-\frac{E}{1-K}}{N}$.
B. $\frac{A-\frac{K}{E-1}}{N}$.
C. $\frac{\mathbf{A-\frac{E}{K-1}}}{\mathbf{N}}$.

6. El modelo de TRI calcula, a partir del conocimiento,...

A. la puntuación total esperada.
B. la probabilidad de acertar un ítem.
C. la discriminación del test.

7. Los parámetros *a*, *b* y *c* de la TRI indican, respectivamente,…

A. discriminación, dificultad, pseudoadivinación.
B. dificultad, discriminación, pseudoadivinación.
C. pseudoadivinación, discriminación, dificultad.

8. Si, por nivel de dificultad, solo pudiéramos tener ítems de un tipo, estos deberían ser, generalmente,…

A. fáciles.
B. difíciles.
C. medios.

9. Un ítem que pregunte sobre un aspecto del temario difícil debería ser...

A. fácil.
B. medio.
C. difícil.

10. La evaluación del sesgo pretende…

A. hacer más justas las pruebas.
B. evaluar la dificultad de los ítems.

Ved también

En el último apartado de este módulo se detalla un ejemplo de respuestas (ficticias) de un grupo de veinte alumnos a esta prueba, junto a los cálculos de la mayoría de los índices a los que haremos referencia en este texto.

C. aumentar la fiabilidad de la prueba.

3. Teoría clásica

Existen dos grandes modos de acercarse al análisis de ítems. Distinguiremos, pues, entre la teoría clásica de test (TCT) y la teoría de respuesta al ítem (TRI). La primera la estudiaremos en este apartado y la segunda, en el siguiente. ¿Qué supuestos tiene la TCT? Aunque se estudia en profundidad en el apartado de fiabilidad, se resumen en la ecuación

$$X = V + E$$

Esta implica que la puntuación que una persona obtiene al contestar un instrumento de medida (X) contiene el denominado "nivel verdadero" de esa persona (V) y una parte de error. El objetivo de la TCT es, pues, medir y minimizar ese error, lo que implica analizar la fiabilidad de la medida. Como no podía ser de otra manera, y siempre bajo estos supuestos, todos los indicadores de calidad de los ítems dependen de la muestra de personas que los han contestado.

Veamos, ahora, las principales propiedades que se han de medir de un ítem.

3.1. Dificultad

El **índice de dificultad** de un ítem (ID) es la proporción de personas que lo contestan correctamente. Es decir,

$$ID = \frac{A}{N}$$

Lectura de la fórmula

A: Número de personas que aciertan el ítem.
N: Número total de personas que lo contestan.

Al tratarse de una proporción, ya que los acertantes son un subconjunto de los que contestan, es obvio que sus valores fluctúan entre 0 y 1, y frecuentemente se expresan como un porcentaje. Paradójicamente, los valores cercanos a 1 indican una baja dificultad –debería llamarse pues índice de facilidad y no de dificultad– y valores cercanos a 0 indican dificultad máxima. Podemos ver en la fila titulada como ID de la hoja de cálculo anexa las dificultades de los ítems del ejemplo de prueba de ejecución máxima.

La fórmula anterior presenta un problema: no tiene en cuenta que una parte de los aciertos se dan por puro azar. Al tratarse de preguntas con alternativas cerradas, es lógico pensar que una parte de los acertantes no conocían la respuesta y que si han acertado es "solo" porque han elegido una de las alternativas, no porque "sepan" la respuesta. El problema que se nos plantea es desconocer cuántos lo han hecho. La solución es muy intuitiva. Si cien personas que no saben japonés contestaran un examen redactado en esa lengua con

preguntas de cuatro alternativas, ¿cuántas preguntas esperaríamos que acertaran? La esperanza matemática –y el sentido común– nos dice que veinticinco. ¿Y si hubiera cinco alternativas de respuesta? Obviamente, veinte. Por lo tanto, el número de aciertos total –y por ende, la probabilidad de acertar una pregunta– depende en cierta medida del número de alternativas de respuesta.

Por tanto, es recomendable utilizar la siguiente fórmula.

$$IDc = \frac{A - \frac{E}{K-1}}{N}$$

Lectura de la fórmula

A: Número de personas que aciertan el ítem

E: Número de personas que fallan el ítem

K: Número de alternativas (u opciones) de respuesta

N: Número total de personas que lo contestan

Inmediatamente, observamos que la diferencia entre ambas fórmulas es que en la segunda se resta de *A* un número que se obtiene de dividir los errores (*E*) entre el número de alternativas erróneas (*K*-1).

Para aprehender este concepto, observemos los resultados del ítem 4. Lo han acertado 8 de los 20 participantes. Su dificultad sin corregir es por tanto del 40%. Ahora bien, al corregir el índice por el acierto al azar [(12/[(3-1)):20]=3], obtenemos un 30%, esto es, podemos suponer que el 30% de las personas lo han acertado de casualidad. Vale la pena observar que con el primer cálculo obtendríamos una dificultad del 40% (podríamos etiquetarlo como de dificultad media), mientras que con el segundo obtendríamos una *IDc* = 10% (alta dificultad).

Aunque conceptualmente no tiene sentido, puesto que sigue siendo una proporción, el *IDc* puede ser inferior a 0: en ese caso, se asigna *IDc* = 0. Esto es lo que ocurre con el ítem 10, que tiene una dificultad corregida de –5%, lo que no tiene sentido. Observemos también cómo los ítems 1 y 3 son perfectamente inútiles, el primero por fácil y el segundo por difícil.

Un ítem que todos aciertan o que todos fallan no sirve para nada más que para perder el tiempo contestándolo. Si todo aquel que responde acierta, es como si regaláramos a todos los alumnos una parte de la puntuación. Y si todos lo fallan, es como si los penalizáramos. Supongamos una prueba que tiene 10 ítems y una puntuación teórica entre 0 y 10. En el primer caso que hemos expuesto, la puntuación real podría fluctuar entre 1 y 10, y en segundo entre 0 y 9. Está claro que esto no habla bien de las propiedades de la prueba.

Una vez que sabemos la dificultad de un ítem, planteémonos, ¿cómo deberían ser las dificultades de todos los ítems de una prueba? Como dice la directriz dos de Moreno, Martínez y Muñiz (2004), la dificultad de un ítem debe relacionarse con la del concepto que recoge. Esto es, si un contenido es fácil, el ítem debe ser fácil. Por tanto, una prueba que mide contenidos diversos debería tener ítems de todas las dificultades, y éstas deberían corresponderse a la dificultad de los conceptos medidos.

Una propuesta, realizada por nosotros (Bonillo, 2012) es mostrar en un gráfico la dificultad (eje Y) de los ítems de una prueba (ordenados de menor a mayor dificultad en el eje X) y observar si la pendiente es cercana a los 45º. Esto es, la línea que une los puntos de estas dificultades debería cruzar en diagonal el gráfico.

La siguiente figura muestra esta idea aplicada a ítems de los exámenes de acceso a la formación sanitaria especializada (las célebres pruebas de médico, farmacéutico y enfermera interno residente, MIR, FIR y EIR, respectivamente, para dos años, el 2005 y el 2006). Si sumamos ambas convocatorias, se presentaron a ellas más de 23.000 aspirantes y sus resultados son muy relevantes, ya que los aspirantes que las superan serán los futuros médicos, farmacéuticos y enfermeras especialistas.

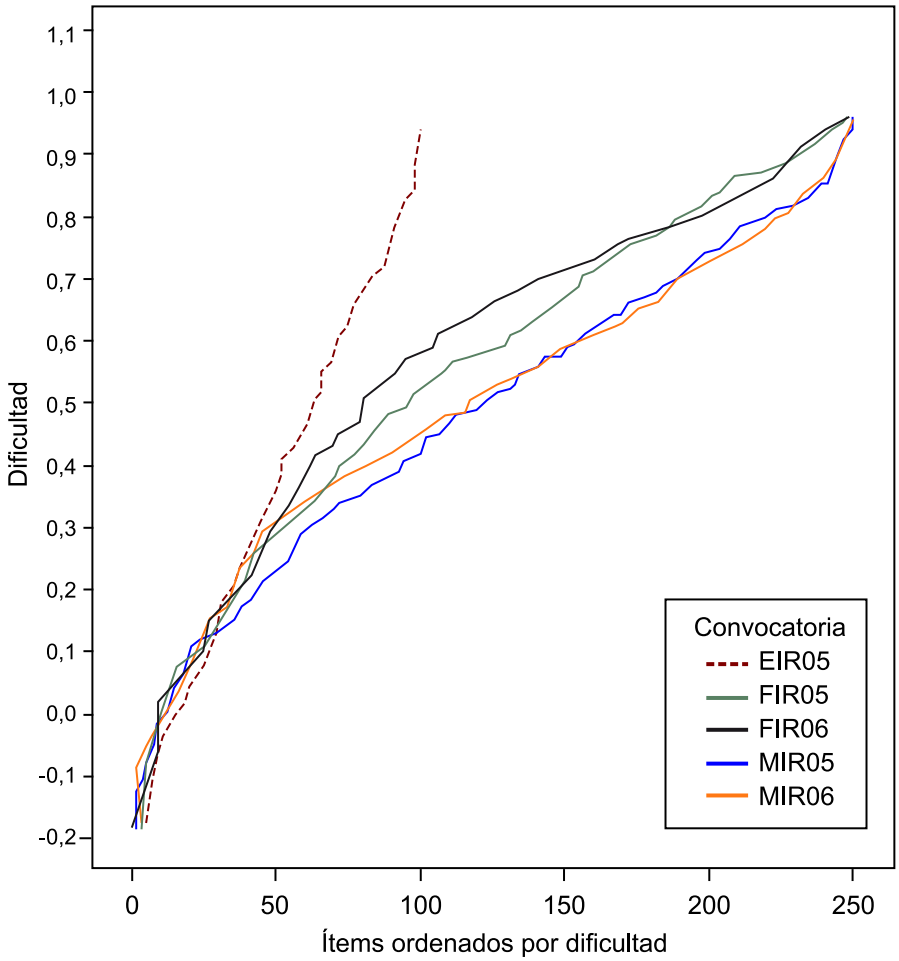

Figura 1. Ítems y dificultad de las pruebas FSE

Podemos ver en el gráfico varias cuestiones que merecen la pena destacarse. En primer lugar, observamos que hay dificultades negativas, y que estas se explican por aplicar la corrección del azar a ítems muy difíciles. En segundo lugar, cabe tener en cuenta que la prueba de enfermería cuenta con 100 ítems (frente a los 250 del resto de las convocatorias), lo que explica la evidente diferencia en las pendientes. En tercer lugar, las curvas de las pruebas de farmacia están,

Para saber más

Si deseáis conocer con más detalle la propuesta, que aquí solo apuntamos, podéis consultar el artículo original (Bonillo, 2012).

en el gráfico, "más altas" que las de medicina, lo que indica que son pruebas más fáciles. En cuarto lugar, las curvas son muy semejantes intraprogramas, es decir, la dificultad de las pruebas es muy semejante año tras año.

3.2. Discriminación

¿Es suficiente saber si un ítem es fácil o difícil para decidir si es adecuado o no? Intuitivamente, podríamos pensar que sí, pero estaríamos equivocados. De hecho, si tuviéramos que destacar una propiedad psicométrica de los ítems sobre el resto, esta sería la discriminación. Si un ítem no discrimina, no es útil para la medición, y ese es el objetivo para el que fue redactado.

Como su nombre indica, entendemos como discriminación la capacidad de un ítem de distinguir entre las personas que tienen un buen rendimiento en el test, respecto a las que lo tienen malo.

¿Quiénes deben contestar correctamente una pregunta de examen? No es tan importante si son muchos o pocos alumnos como que los acertantes sean, en general, "de los buenos alumnos". ¿A qué nos referimos cuando decimos "los buenos"? A aquellos que tienen una alta puntuación en la prueba. Es decir, un ítem debe ser más acertado entre aquellos que han obtenido una alta puntuación en la prueba que entre los que no la tienen. Obviamente, una pregunta no puede ser buena si solo la aciertan los peores alumnos: debe ocurrir lo contrario.

El índice de discriminación más popular es el índice D, conocido también como índice basado en las proporciones de aciertos.

$$D = Pa - Pb$$

Lectura de la fórmula

P_a: Proporción de personas del grupo de alto rendimiento que acierta el ítem

P_b: Proporción de personas del grupo de bajo rendimiento que acierta el ítem

Las proporciones se calcularían como hemos visto en la primera fórmula presentada y, de nuevo, puede expresarse como porcentajes. Pero ¿quiénes son los alumnos de alto y bajo rendimiento? Existen varias maneras de definir el punto de corte de la puntuación total en la prueba para hacer esta clasificación. Por un lado, es frecuente utilizar la media de la puntuación total en la prueba, lo que crea dos grupos de igual tamaño. Esta estrategia tiene como ventaja que todos los participantes participan en el cálculo, pero tiene como claro inconveniente que los grupos son poco extremos. Intuitivamente comprobamos que dos personas con rendimiento muy semejante pueden estar en grupos diferentes solo por una pequeña diferencia.

Es preferible utilizar grupos más extremos para poder estudiar correctamente este índice. Kelley (1939) recomienda utilizar los percentiles superior e inferior del 27%. ¿Por qué 27% y no 25%? Aunque el artículo original demuestra que el 27% es ligeramente mejor que el 25%, en el ejemplo con respuestas ficticias

que mostramos se utiliza el 25% como criterio para separar el grupo de rendimiento alto –se interpretaría como aquel que obtiene puntuaciones superiores al 75% de sus homólogos– del bajo –que reúne el 25% de las puntuaciones más bajas. Calcular el percentil 27 no siempre es sencillo, mientras que el 25 sí lo es, y las variaciones entre uno y otro son muy menores.

¿Cuáles son los límites de D? Es obvio que, teóricamente, puede fluctuar entre 1 y –1. El primer valor se daría solo cuando todas las personas del grupo superior acertaran y todas las del inferior fallaran. En valor –1 solo podría darse en el caso contrario, y entonces deberíamos sospechar si la respuesta considerada como correcta lo es. Ninguna de estas dos situaciones suele darse en la realidad.

¿Cómo debemos pues interpretar este índice? En primer lugar, solo valores positivos indican discriminación. Está claro que un ítem debe ser más acertado entre los mejores. Pero ¿qué valores indican una buena discriminación? Ebel (1965) propuso la siguiente clasificación, que debe ser tomada como orientación:

Tabla 2. Puntos de corte de los valores (en %) de discriminación (D) y su interpretación

D	Interpretación de la discriminación
> 40	Alta discriminación
30-40	Aceptable
20-30	Baja: se sugiere revisar el ítem
0-20	Mala: se elimina el ítem o se reforma profundamente
< 20	Inaceptable: eliminar el ítem

Un motivo para tomar la tabla anterior con precaución es que el índice *D* depende –y mucho– de la dificultad. Si un ítem es muy difícil, tendrá pocos acertantes (por definición), incluso en el grupo de alto rendimiento. Si P_a es baja, la *D* solo puede ser baja. No parece justo comparar *D* de ítems de dificultades muy diferentes. Una alternativa propuesta al índice *D* es calcular la diferencia de proporciones relativa, en lugar de la absoluta. Es decir,

$$Dr = (Pa - Pb) \big/ P$$

Calculemos la discriminación de los ítem 4 y 5, que tienen la misma dificultad.

Ítem 4: $D = 4/5 - 0/0 = 80\%$ y $Dr = \frac{0{,}8 - 0}{0{,}8} = 100\%$

Ítem 5: $D = 2/5 - 0/0 = 40\%$ y $Dr = \frac{0{,}4 - 0}{0{,}4} = 100\%$

Para el ítem 4: el primer valor (D = 80%) debería interpretarse de la siguiente manera: los mejores aciertan el ítem un 80% más que los peores. O interpretarlo como una diferencia de probabilidades: es un 80% más probable acertar el ítem cuando se tiene un alto rendimiento (que si se tiene bajo). El segundo valor (Dr = 100%) se interpretaría como que los buenos aciertan un 100% más (respecto a los malos). En este caso, ambos valores hablan

bien de la capacidad discriminativa del ítem. Para el ítem 5, los datos son parecidos (D = 40% y Dr = 100%). Así pues, ambos ítems serían más que aceptables.

Supongamos ahora un ítem muy difícil, que solo sea acertado por 1 de cada 20 participantes, y supongamos que este pertenece al grupo de alto rendimiento. Está claro que D = 1/6-0/6 = 0,17 = 17%. Según los criterios de Ebel tendríamos que eliminarlo, pero Dr = 100%, y por tanto tiene discriminación relativa máxima. ¿Qué deberíamos entonces hacer? No existe una respuesta absoluta a esta pregunta.

Si, por las características de la prueba, es aceptable tener un ítem tan difícil, este debería mantenerse ya que discrimina tanto como puede discriminar. Si, por el contrario, no es conveniente que tan pocas personas lo acierten, debería reformarse, pero la decisión –en este caso– depende por completo de la dificultad y no de la discriminación.

En resumen, hay que tener claro que la discriminación depende de la dificultad y no debe interpretarse *per se*. En términos estadísticos, la discriminación depende de la variancia de la dificultad.

Existen otros índices alternativos a los presentados para medir la discriminación. Uno de los más utilizados es la correlación ítem-test. Habitualmente, se utiliza el índice de correlación biserial-puntual, ya que permite cuantificar la relación entre una variable binaria –el acertar o no el ítem– y una variable de escala –la puntuación total de la persona en la prueba, idealmente sin tener en cuenta el ítem analizado. La fórmula y la lógica de esta prueba se encuentran ampliamente explicadas tanto en Muñiz (2003)[1] como en Martínez Arias (1992)[2], pero es muy sencillo ver que una alta correlación –cercana a 1– indica una gran discriminación del ítem, que valores cercanos a –1 indican lo contrario (donde los buenos fallan el ítem y los malos lo aciertan) y que valores cercanos a 0 indican que nada tiene que ver acertar este ítem con el conocimiento que mide el conjunto de la prueba.

(1)Muñiz (2003, p. 220).

(2)Martínez Arias (1992, p. 556).

3.3. Discriminación de los distractores

Un aspecto clave para el buen funcionamiento de un ítem es que sus distractores[3] realmente lo sean. Una alternativa que no es elegida por nadie –o casi– no está confundiendo a las personas que responden, y por tanto no es útil. Y a la inversa: una alternativa incorrecta que es muy elegida por los mejores quizá no es tan incorrecta como pensó el que la redactó.

(3)Distractores es el nombre que se le da a las alternativas de respuesta incorrectas.

¿Cómo se estudia el comportamiento de los distractores? El índice habitual vuelve a ser el índice D que ya hemos visto, pero en lugar de calcularlo a partir de quienes aciertan y fallan, se hace con quienes eligen cada una de las alternativas de respuestas.

Observad atentamente los resultados de administrar el ítem 5. La discriminación de la opción de respuesta B es más alta que la de la respuesta correcta, la C. Debemos interpretarlo como que los mejores eligen en mayor medida un distractor –como es la C– que la respuesta correcta –como es la B. Esto implica que debemos comprobar si la pauta pudiese contener un error. En nuestro caso no es así, y podemos atribuir al azar que la opción C haya sido tan ele-

gida. Ahora bien, quizá deberíamos recalcar a los alumnos qué significa cada elemento de la fórmula (puesto que es lo que se pregunta en el ítem 5) y fortalecer así el aprendizaje.

¿Qué propiedades matemáticas debe tener un distractor? Obviamente, tener discriminación negativa, es decir, ser más elegido entre los peores que entre los mejores. Además, sería óptimo que todos los distractores tuvieran una discriminación parecida, ya que indicaría que sus capacidades de atracción son semejantes. Conseguir esto es especialmente difícil, y esta dificultad crece exponencialmente con las alternativas de respuesta. En resumen: es mucho más difícil redactar tres distractores que sean efectivos que dos. Por ello, la mayoría de los estudios realizados recomiendan usar como mucho tres alternativas de respuesta.

No debemos creer que las propiedades que hasta ahora hemos presentado son independientes entre sí: nada más alejado de la realidad. Si un ítem tiene una opción de respuesta inverosímil (por ejemplo, Maradona como autor de *El Quijote*), el ítem será más fácil y necesariamente discriminará peor.

Como ya hemos hecho cuando hemos estudiado la dificultad, ahora nos planteamos si existe alguna manera de estudiar el conjunto de las discriminaciones de los ítems de una prueba.

De nuevo, la propuesta es nuestra (Bonillo, 2012) y consiste en mostrar en un gráfico, denominado diagrama de cajas, la discriminación de la opción correcta y de cada uno de los distractores, ordenados de mayor a menor.

La figura siguiente muestra esto aplicado a, de nuevo, los exámenes de acceso a la formación sanitaria especializada.

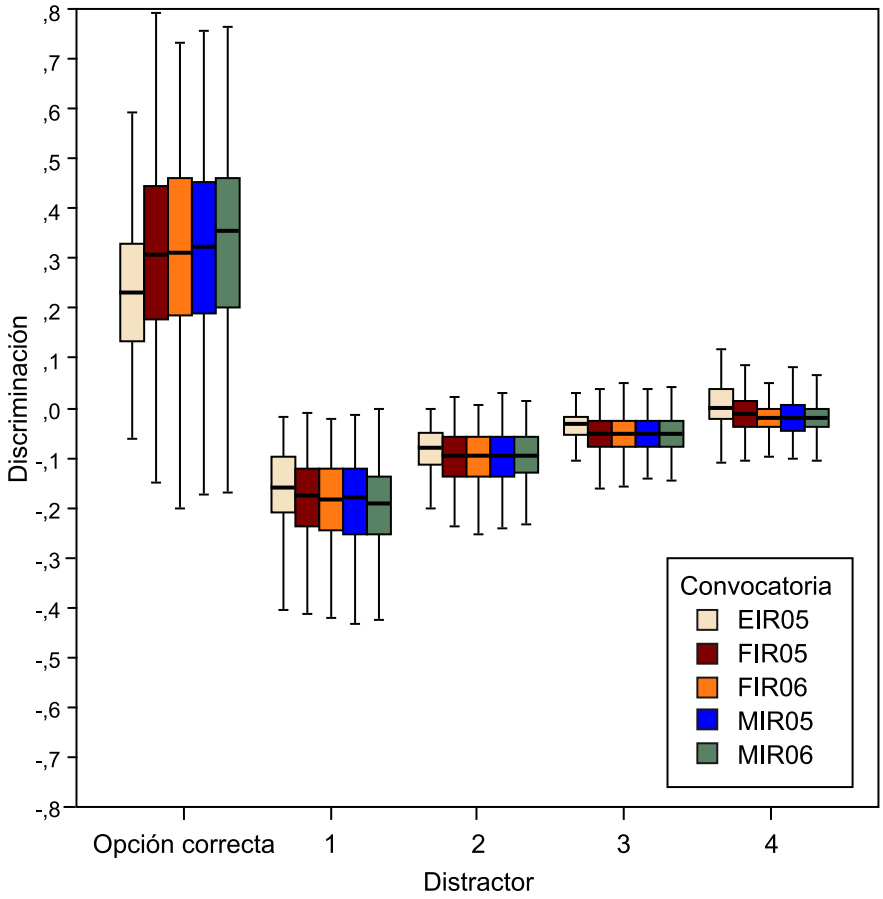

Figura 2. Discriminación de los distractores de las pruebas FSE

Debemos tener en cuenta que en esta figura aparecen las cinco convocatorias analizadas (esto es, [250 ítems × 2 programas × 2 años + 100 ítems de EIR] × 5 alternativas = 5.500 valores). Así, y para cada ítem, el distractor 1 es el más discriminativo, y el 4 el que menos. Como suele ocurrir en estos gráficos, las cajas muestran la media –en trazo grueso– y los cuartos –en los límites de las cajas. Las patillas (*whiskers*) muestran los valores mínimos y máximos no alejados ni extremos. Los alejados se muestran con puntos y los extremos, con asteriscos.

Se observa que las discriminaciones de las alternativas correctas son semejantes entre especialidades y convocatorias. Destacan de las demás las discriminaciones relativas a la prueba de EIR, que son más bajas y menos dispersas. En el análisis de los distractores se observa que existe un escalado entre estos, pero que se reduce cuantas más alternativas se contemplan; es decir, la diferencia entre la tercera y la cuarta alternativa es mucho menor que entre la primera y la segunda. También se observa que las alternativas tres y cuatro –recordemos que son ordenadas por su discriminación y que no deben identificarse con alternativas de respuesta D y E, por ejemplo– tienen discriminaciones muy bajas o casi nulas. Si consideramos que el límite superior de las cajas de la última alternativa es superior a 0, podemos decir que más del 25% de los ítems tienen una alternativa de respuesta con discriminación positiva –es decir, más elegida por el grupo con rendimiento alto. Además, la última alternativa presenta muchos valores extremos y alejados, esto es, ítems en los que, por su

Para saber más

De nuevo, si queréis conocer con detalle esta propuesta, podéis consultar el artículo original (Bonillo, 2012), ya que en este módulo no podemos extendernos mucho más de lo que ya hemos hecho.

alta discriminación positiva, sería discutible si la opción dada como correcta verdaderamente lo es, o es la única que lo es. Conclusión que hay que extraer: con tres opciones sería más que suficiente.

3.4. Valoración del sesgo

Un aspecto crucial cuando se crea –y cuando se valora– tanto un ítem como un instrumento de medida es que estos sean no sesgados. ¿De qué hablamos cuando nos referimos al sesgo en un instrumento de medida? Una báscula estará sesgada si siempre infravalora el peso de un objeto frente a otro cuando sabemos que ambos pesan exactamente lo mismo. En el contexto de las pruebas de ejecución máxima, entendemos que un ítem –o un test– está sesgado cuando grupos, por ejemplo hombres y mujeres o ricos y pobres, que tienen el mismo conocimiento sobre la materia medida, no obtienen valores iguales, sino que uno de los grupos es sistemáticamente "perjudicado".

Como podéis imaginar, los instrumentos sesgados pueden tener graves implicaciones sociales. Si un examen, como la selectividad, favoreciera sistemáticamente a un alumno con nivel socioeconómico alto frente a uno que no lo tiene, y siempre y cuando sepamos que ambos saben exactamente lo mismo sobre el tema, la selectividad no sería socialmente justa.

¿Cómo valorar el sesgo? Actualmente, se utiliza el concepto de **funcionamiento diferencial de los ítems** y el principal índice que se deriva es el DIF[4]. Decimos que un ítem "tiene o presenta" DIF[5] cuando se dan diferencias estadísticamente significativas en la puntuación de un ítem en dos grupos diferentes que deberían tener, de buena lógica, el mismo nivel. Para evaluar el DIF pueden utilizarse diferentes procedimientos matemáticos[6], pero en este texto solo hablaremos del método de Mantel-Haenszel (1959). Este es, en el marco de TCT, el más utilizado por su sencillez de cálculo y sus buenos resultados.

(4)La sigla DIF proviene de las iniciales del término en inglés.

(5)En terminología psicométrica coloquial, decimos que existe DIF.

(6)En Muñiz (2003, pp. 239-253) se puede ver una excelente revisión.

Lectura recomendada

Recomendamos encarecidamente la consulta del ejemplo que presenta Muñiz (pp. 247-249) y que aquí describiremos, pero no desarrollaremos matemáticamente.

El modo de cálculo y la idea que subyace al DIF son muy sencillos. Supongamos que deseamos saber si nuestra prueba no está afectada por el sexo de los que responden. Es obvio que no debería estarlo, y que si lo está deberíamos corregirlo, ya que podría calificarse de sexista.

Se trata entonces de dividir a los sujetos en grupos en función de sus puntuaciones totales (por ejemplo, en cinco grupos). Luego, habrá que elaborar una tabla por grupo en la que observaremos si la variable sexo se asocia a acertar más. Finalmente, este resultado se agregará en el estadístico de Mantel-Haenszel y se comparará con el χ^2 de referencia. Si el resultado es significativo, es que existen tramos de puntuación total en los que un sexo parece tener ventaja sobre otro, ya que acierta más. Entonces podemos decir que la prueba no

es justa. Cuando creemos una prueba deberíamos comprobar que es independiente de variables como el género, la raza y cualquier otra que pueda llevar, implícitamente, a la discriminación de las personas.

4. Teoría de respuesta al ítem

La **teoría de respuesta al ítem**[7] parte de una perspectiva totalmente distinta. Critica a la TCT al afirmar que estudia las propiedades de un test particular en una muestra –también particular– de personas. Desde la TCT, es cierto que dos tests distintos que miden el mismo constructo tendrán propiedades diferentes. Y también es cierto que esas propiedades dependerán de si las personas utilizadas para calibrar el test de rendimiento tienen o no un alto rendimiento. La TRI permite superar estos problemas, pero a costa de complicar enormemente los cálculos y que resulte más difícil de utilizar al ser más "cajanegrista".

(7)En inglés, *item response theory* (*IRT*).

La TRI se basa en el cálculo, para cada ítem, de una serie de parámetros, que asumen un modelo matemático muy concreto. El objetivo fundamental es la medición del rasgo latente[8], a partir de tres parámetros:

(8)El rasgo latente es el nombre que recibe el constructo que se va a medir, por ejemplo, el conocimiento de una asignatura.

- la discriminación del ítem,
- su dificultad y
- el acierto al azar.

Estos tres parámetros pueden observarse en la figura siguiente, que resulta clave para entender qué es la TRI[9]: se conoce como la curva característica del ítem (CCI).

(9)En inglés, *item response function* (*IRF*).

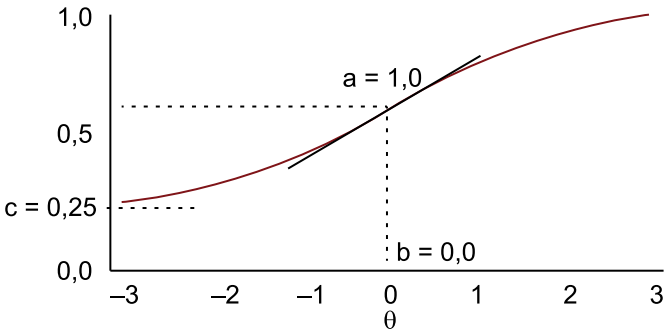

Figura 3. Ejemplo de curva característica de un ítem (CCI)

La CCI muestra, en el eje de las ordenadas (el eje *Y*), la probabilidad de acertar el ítem a partir de la magnitud del rasgo latente (eje abscisas o *X*). Esta probabilidad sigue una función sigmoidal (en forma de *S*, también llamada logística) y el rasgo latente se indica como θ (*theta*). La función logística tiene como propiedad que no puede ser menor de 0 ni mayor de 1, y θ suele estar comprendida entre –3 y +3.

A partir de la CCI podemos medir los tres parámetros clave de la TRI. El parámetro indicado como a) mide la discriminación del ítem. Una curva muy plana expresaría que no es importante tener un alto conocimiento del rasgo para

aumentar la probabilidad de acierto. Es decir, cuanto mayor sea la pendiente, mayor será la discriminación. El del gráfico mostrado es 1, valor positivo y aceptable.

El parámetro indicado como *b* mide la dificultad a partir del punto de corte del eje *X*, que corresponde a una probabilidad de acierto del 50%. Se interpretaría como el nivel de rasgo latente necesario para tener un 50% de probabilidades de acertar el ítem. Cuanto mayor sea la *b*, más difícil será el ítem, ya que hará falta más conocimiento para poder llegar a ese 50% de probabilidad deseado. El valor mostrado en el gráfico es 0, que se interpretaría como que es un ítem de dificultad (exactamente) media. El tercer parámetro, indicado como c), mide el nivel de azar y se conoce también como *índice de pseudoadivinación*. Gráficamente, corresponde al valor de *X* que corta el eje *Y*. Contempla, lógicamente, la probabilidad de acertar cuando el conocimiento del ítem es nulo. El valor del gráfico indica que este es alto: del 25%.

Estos cálculos se realizan con software muy específico. Mostrar con qué herramientas hacerlo y cómo excede los objetivos de este módulo.

Enlace recomendado

La siguiente página contiene un conjunto de programas gratuitos que permiten aplicar la TRI: http://www.psychology.gatech.edu/unfolding/FreeSoftware.html. Ahora bien, debemos recordar que aplicar la TRI no es sencillo y que requiere mayor formación que la que necesita la aplicación de la TCT.

Ejemplo de evaluación de las propiedades de los ítems a una muestra de alumnos ficticia

En la siguiente tabla se muestran las respuestas de los veinte sujetos (por filas) a las diez preguntas de la prueba (columnas). En negrita se marcan las respuestas correctas, cuya pauta aparece en la primera fila.

Pauta	**A**	**C**	**B**	**B**	**C**	**B**	**A**	**C**	**C**	**A**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**
S1	**A**	**C**	C	**B**	**C**	**B**	**A**	**C**	**C**	B
S2	**A**	**C**	C	**B**	B	**B**	**A**	**C**	**C**	A
S3	**A**	**C**	C	**B**	B	**B**	**A**	**C**	**C**	**A**
S4	**A**	**C**	C	A	**C**	**B**	**A**	**C**	**C**	**A**
S5	**A**	**C**	A	**B**	B	**B**	**A**	**C**	**C**	**A**
S6	**A**	A	A	A	A	A	B	A	B	B
S7	**A**	**C**	A	**B**	**C**	A	B	B	B	B
S8	**A**	**C**	A	A	V	A	B	B	B	B
S9	**A**	A	A	**B**	**C**	A	B	B	B	C
S10	**A**	**C**	A	**B**	A	A	B	B	B	B

Pauta	**A**	**C**	**B**	**B**	**C**	**B**	**A**	**C**	**C**	**A**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**
S11	**A**	A	A	**B**	**C**	**B**	B	**C**	**C**	**A**
S12	**A**	A	A	A	A	**B**	B	**C**	**C**	**A**
S13	**A**	A	A	A	**C**	**B**	B	B	B	C
S14	**A**	A	A	A	**C**	**B**	B	B	B	B
S15	**A**	A	A	A	**C**	**B**	B	B	B	C
S16	**A**	A	A	A	B	C	B	B	**C**	B
S17	**A**	A	A	A	B	A	B	B	B	B
S18	**A**	A	A	A	A	C	B	A	A	C
S19	**A**	A	A	A	A	A	B	A	A	C
S20	**A**	A	A	A	A	A	B	A	A	C

La tabla reproduce la anterior, pero codificando la respuesta como 1, si ha acertado, y 0, si no lo ha hecho. La columna titulada P. total contiene la puntuación total de cada persona, y corresponde a la suma de los unos de la tabla. En la parte inferior vemos dos filas, denominadas ID e IDc. Como sabemos, la primera corresponde al índice de dificultad y la segunda también a este pero en su versión corregida. Observad la gran diferencia entre ambos y recordad que el segundo corrige el acierto por azar que el primero obvia.

	Acierto										**P. total**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**	
S1	1	1	0	1	1	1	1	1	1	0	8
S2	1	1	0	1	0	1	1	1	1	1	8
S3	1	1	0	1	0	1	1	1	1	1	8
S4	1	1	0	0	1	1	1	1	1	1	8
S5	1	1	0	1	0	1	1	1	1	1	8
S6	1	0	0	0	0	0	0	0	0	0	1
S7	1	1	0	1	1	0	0	0	0	0	4
S8	1	1	0	0	0	0	0	0	0	0	2
S9	1	0	0	1	1	0	0	0	0	0	3
S10	1	1	0	1	0	0	0	0	0	0	3
S11	1	0	0	1	1	1	0	1	1	1	7
S12	1	0	0	0	0	1	0	1	1	1	5
S13	1	0	0	0	1	1	0	0	0	0	3
S14	1	0	0	0	1	1	0	0	0	0	3

	Acierto										P. total
S15	1	0	0	0	1	1	0	0	0	0	3
S16	1	0	0	0	0	0	0	0	1	0	2
S17	1	0	0	0	0	0	0	0	0	0	1
S18	1	0	0	0	0	0	0	0	0	0	1
S19	1	0	0	0	0	0	0	0	0	0	1
S20	1	0	0	0	0	0	0	0	0	0	1
ID	100,0%	40,0%	0,0%	40,0%	40,0%	50,0%	25,0%	35,0%	40,0%	30,0%	
IDc	100,0%	10,0%	–50,0%	10,0%	10,0%	25,0%	–12,5%	2,5%	10,0%	–5,0%	

Los puntos de corte de los cuartiles de la puntación total son, respectivamente, 1,75 y 8. Esto es, qué puntuación mínima hay que tener para estar por encima del primer y tercer cuartiles. Los valores comprendidos entre ambos corresponden al grupo denominado intermedio, que contiene el 50% de las personas con valores alrededor de la media. A partir de estos valores se calcula la columna que podemos ver en la derecha de la tabla, y que clasifica a las personas en tres grupos.

	P. total	Grupo Rend.
S1	8	Superior
S2	8	Superior
S3	8	Superior
S4	8	Superior
S5	8	Superior
S6	1	Inferior
S7	4	Intermedio
S8	2	Intermedio
S9	3	Intermedio
S10	3	Intermedio
S11	7	Intermedio
S12	5	Intermedio
S13	3	Intermedio
S14	3	Intermedio
S15	3	Intermedio
S16	2	Intermedio
S17	1	Inferior
S18	1	Inferior

	P. total	Grupo Rend.
S19	1	Inferior
S20	1	Inferior

Las celdas de la tabla siguiente muestran los cálculos relativos a la discriminación de cada uno de los ítems. En las celdas tituladas % acierto grupo superior se muestra el porcentaje de personas del grupo de rendimiento superior que han elegido, respectivamente, las alternativas A, B y C. Idéntico cálculo se realiza en las celdas tituladas % acierto del grupo inferior. ¿Qué hacemos posteriormente con estos valores?

Gracias a este cálculo podemos obtener de manera sencilla la discriminación de cada una de las opciones de respuesta. En las celdas tituladas Índice D alternativas vemos la discriminación de la alternativa A, y lo mismo para las B y C con las filas inferiores. En negrita se muestra la discriminación de la alternativa correcta. Observad que la negrita coincide con la letra que, en la primera tabla de este apartado, aparecía en la pauta de respuestas correctas.

	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
% acierto grupo superior										
A	100,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%
% acierto grupo inferior										
A	100,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
B	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
C	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Índice D alternativas										
A	0,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%

Resumen

A lo largo de este texto hemos expuesto una multitud de aspectos sobre los instrumentos de medida. Sobre una prueba hemos visto qué estudiar, cómo y cuándo hacerlo, y hemos sido –especial y conscientemente– insistentes en el porqué hacerlo. La idea principal que nos gustaría haber podido transmitir es que lo que no podemos hacer es no hacer nada. Es decir, la peor de las situaciones posibles que podemos imaginar –en este contexto, claro– es aplicar una prueba sin estudiar ninguna de sus propiedades. No hacerlo convertirá el error en norma y no en excepción.

Esta situación descrita es, lamentablemente, la realidad. En el ámbito universitario, que es el que nos resulta más cercano, son contados los profesores que estudian sus exámenes tras administrarlos. Incluso los que lo hacen rara vez publican los resultados, lo que permitiría a los alumnos contrastar la justicia de la prueba con la que se les examinó. ¿Por qué ocurre esto? Creemos que es más achacable a la falta de formación que a la falta de transparencia. Al profesor no se le forma en cómo debe valorar sus pruebas, como tampoco se le forma en cómo debe realizar una clase. En nuestro sistema educativo, y estamos refiriéndonos a todas las etapas, la cultura de trabajar con evidencias no está implantada.

No podemos no centrarnos en el ámbito educativo, que es al que pertenecemos, pero ¿no creéis que deberían publicarse los datos que permitan valorar la justicia de unas oposiciones, como son las pruebas de acceso de la formación sanitaria especializada? De hecho, en el artículo ya citado (Bonillo, 2012) esta fue nuestra principal propuesta. Así, los opositores podrían impugnar preguntas, proponer otras correcciones y, en definitiva, todos podríamos estar más tranquilos al saber que las pruebas son justas y premian de verdad a los mejores.

En este módulo se han presentado muchos conceptos novedosos y se ha realizado muy brevemente. ¿Consideramos que el estudiante está preparado para aplicarlos mañana? En absoluto. Nos gustaría pensar que, para aquel que necesite un día crear y/o valorar una prueba, este texto es el primero de otros. Esos otros están citados o pueden buscarse alternativas.

Para aquel estudiante que no necesite lo que aquí se expone, que será la gran mayoría, confiamos en que los conceptos expuestos se entiendan. Cuando esto ocurre, y en el futuro se necesita aplicarlos, recuperarlos de la memoria y de la biblioteca es sencillo.

En cualquier caso, y para todos, esperamos haber insuflado espíritu crítico y deseo de trabajar con y por las evidencias. Al final, estos dos elementos son los que separan la psicología de otras cosas que nos deben ser ajenas.

Bibliografía

Baker, F. (2001). *The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation*. Maryland: University of Maryland.

Bonillo, A. (2012). Pruebas de acceso a la formación sanitaria especializada para médicos y otros profesionales sanitarios en España: examinando el examen y los examinados. *Gaceta Sanitaria*, 26 (3), pág. 231-235

Downing S. M. (2005). The effects of violating standard item writing principles on test and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10, 133-143

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood: Prentice-Hall.

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

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30 (1), 17-24. Disponible en: 10.1037/h0057123.

Mantel, N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Martínez Arias, R. (1996). *Psicometría: Teoría de los Tests Psicológicos y educativos*. Madrid: Síntesis.

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

Muñiz, J. (2003). *Teoría Clásica de los Tests*. Madrid: Pirámide.