2

Construcción de tests y análisis de ítems

Un test está formado por una serie de ítems (o elementos, reactivos, preguntas, cuestiones,…) a los que cada individuo debe responder. Después de cuantificar las respuestas de una persona a los ítems del test, se pretende asignar una puntuación (a veces, varias) a esa persona respecto al constructo o atributo que se pretende medir. La puntuación asignada debería indicar su grado o nivel en el atributo, constructo o rasgo a evaluar. Vimos en el capítulo precedente que en las Ciencias Sociales y de la Salud es frecuente recurrir a indicadores para obtener la posición de la persona en un constructo. Se puede, por ejemplo, inferir su *posición social* tras preguntarle donde vive, cómo es su casa, cuánto gana… Para la medida de muchos constructos utilizamos también la medición mediante indicadores (Navas, 2001). Los ítems de un test de Responsabilidad, por ejemplo, serán los indicadores que nos permitan llegar al nivel de la persona en la variable latente Responsabilidad, a partir de un conjunto de respuestas.

En este tema vamos a estudiar cómo se construyen los tests y los indicadores de calidad psicométrica de los ítems. En capítulos sucesivos estudiaremos los indicadores de calidad del test como un todo.

El primer paso para la construcción del test es detallar minuciosamente los objetivos que se pretenden conseguir con su aplicación y las características fundamentales que debe tener. Cuando hayamos especificado ambas cosas, tendremos que decidir el tipo de ítem que resulta más apropiado. Estudiaremos los formatos más comunes y las normas de redacción que se recomienda seguir.

En el proceso de construcción de un test hay que elaborar más ítems de los que pensamos aplicar, con la idea de someterlos a un proceso de control de calidad que nos permita quedarnos con los más apropiados y conseguir así el mejor test posible. El proceso de control de calidad tiene dos partes: una cualitativa y otra cuantitativa. La cualitativa pre-

cede a la administración de los ítems y requiere que varios expertos comprueben que se han seguido correctamente todos los pasos en la construcción del test.

Estudiaremos también cómo se convierten en valores numéricos las respuestas dadas a los ítems. El siguiente paso del proceso es la aplicación piloto del test provisional (es decir, del test formado a partir de los ítems que se consideran adecuados tras el análisis cualitativo). Tras la administración piloto, se procede al estudio cuantitativo de las características de los ítems mediante un proceso denominado como *análisis de ítems*. A continuación, hay que decidir qué ítems concretos formarán el test definitivo. Se llama *ensamblaje* del test al proceso por el que se forma el test a partir de los ítems disponibles que han superado los controles de calidad.

Una ventaja de seguir un proceso sistemático es que se va a conseguir un test similar al que hubiesen obtenido otros expertos, e incluso a un segundo test que hiciéramos nosotros, si éste fuera el interés. El proceso de construcción requiere además que se haya pensado en todos los aspectos que afectan a la calidad del test resultante, lo que previsiblemente ayudará a conseguir un mejor resultado final.

Diseño del test

Lo primero a especificar es el constructo, atributo o característica psicológica a evaluar y el propósito del test. Hay que dar respuestas a tres preguntas (Navas, 2001): ¿Qué se va a medir con el test?, ¿a quién se va a medir? y ¿qué uso se piensa hacer de las puntuaciones? Podemos querer evaluar el nivel de Autoestima, Motivación, Inteligencia, el nivel de conocimientos en cierta materia, la calidad autopercibida del sueño, etc. Las teorías sobre los constructos suponen siempre un punto de referencia importante. Ciertamente son muchos los constructos que podemos querer evaluar, pero no son menos los propósitos de la evaluación. Por ejemplo, en un test educativo, Bloom, Hastings and Madaus (1971) han propuesto tres posibles propósitos: a) la evaluación inicial para diagnosticar puntos fuertes y débiles y ubicar a los estudiantes en el curso adecuado, b) la evaluación formativa para valorar el progreso en la instrucción y decidir qué y cómo enseñarles, y c) la evaluación sumativa para determinar el nivel de conocimientos adquirido en un curso por cada estudiante.

El propósito del test condiciona qué tipos de ítems pueden ser más apropiados. Por ejemplo, un test para la evaluación del dominio del inglés, a aplicar en las pruebas de acceso a la universidad, puede requerir ciertos tipos de ítems (por ejemplo, que evalúen la comprensión de textos científicos), distintos de los que pudiera utilizar un test a aplicar en procesos selectivos y cuyo propósito sea excluir del proceso a los candidatos que no alcancen un nivel mínimo prefijado de comprensión oral del idioma.

Hay que atender a varias características de la población a evaluar, como la edad, el nivel educativo, la familiaridad con el medio de administración (por ejemplo, con el ordenador cuando se vaya a construir un test informatizado), la eventual presencia de discapacidades y de diferencias en el dominio del lenguaje. En estos dos últimos casos, habría que considerar la política de *acomodaciones* a aplicar y su equivalencia (comprobar que la puntuación en el test corresponde al nivel en el constructo, sin ventaja o desventaja atribuible a las acomodaciones). La acomodación más frecuente consiste en modificar el tiempo de administración, de manera que un evaluado con discapacidad motora, por

ejemplo, alcance la misma puntuación que otro sin discapacidad que tenga su mismo nivel en el constructo. A veces se preparan cuadernillos e instrucciones en distintos idiomas, o en tamaños de letra y formato diferentes. Otras veces se cambia el modo de administración, que puede pasar de colectivo a individual, o se leen o interpretan los ítems al evaluado para facilitarle su comprensión.

Schmeiser y Welch (2006) recomiendan prestar atención a lo que llaman *restricciones administrativas*. Las hay de distinto tipo: de tiempo, de coste, de medios (como aulas, ordenadores o vigilantes disponibles), etc. Los medios disponibles pueden condicionar el número de tests a construir si se quiere preservar la seguridad de la prueba. Por ejemplo, en contextos de evaluación educativa, la escasez de aulas o de vigilantes podría aconsejar la construcción de varios exámenes equivalentes, en vez de un único examen. De todas las restricciones, las más importantes son las relacionadas con el tiempo de administración. Dependiendo de la trascendencia del test (no tiene la misma un examen de una parte de la materia que uno con el que se consigue la acreditación para trabajar como médico, por ejemplo), de la edad de los evaluados o del tipo de ítems, será necesario un test con unas u otras características psicométricas, más largo o más corto y que requiera más o menos tiempo de aplicación. La longitud óptima del test es aquella que, siendo aceptable para los evaluados, proporciona puntuaciones con la calidad necesaria para justificar el uso previsto o las decisiones que se toman en el proceso de evaluación.

Vimos en el capítulo precedente que hay tests *referidos a normas* y *referidos al criterio*. El tipo de test también condiciona la prueba concreta que finalmente construyamos. También habrá que decidir si finalmente se va a dar a cada evaluado una o más puntuaciones, y, en el caso de sólo una, cómo contribuyen a ella las diferentes partes del test.

Además del propósito y de todo lo expuesto anteriormente, conviene construir *la tabla de especificaciones*, donde se detallan los contenidos del test, las destrezas cognitivas a evaluar y el porcentaje de ítems que debe corresponder a cada celdilla que resulta de cruzar los contenidos y las destrezas. En evaluación educativa, hay que analizar el programa del curso, preguntar a los profesores qué explican y qué tiempo dedican a cada parte, hacer una revisión bibliográfica, etc. En selección de personal, se han propuesto técnicas propias, como el análisis de puestos o la técnica de los incidentes críticos, que facilitan la especificación de los contenidos y destrezas.

La Tabla 2.1 (adaptada de la Tabla 9.2 de Schmeiser y Welch, 2006, p. 317) muestra la estructura de una tabla de especificaciones de un examen de Asesoramiento Psicológico. Incluye dos destrezas: Comprensión (de hechos, razones, relaciones, principios, fórmulas, gráficos y detección de errores en los procedimientos y en la práctica) y Aplicación (habilidad para seleccionar y aplicar principios y reglas, analizar e interpretar situaciones, extraer conclusiones y diagnosticar necesidades y problemas).

La tabla muestra que el 42% de las preguntas del test evaluarán la destreza Compresión frente al resto (58%) que evaluará Aplicación. En cuanto a contenidos, hay partes menos importantes, como Fundamentos del asesoramiento (un 6% de los contenidos), y otras más importantes, como Asesoramiento individual (30%). La tabla debe detallar los contenidos de cada parte. Por ejemplo, dentro de la parte de Fundamentos del asesoramiento, los modelos de asesoramiento representan el 1% del total del test y se va a evaluar en ellos exclusivamente la destreza Comprensión.

Tabla 2.1. Ejemplo de tabla de especificaciones

	Peso del contenido	Peso de las destrezas cognitivas	
		Compresión	Aplicación
Fundamentos del asesoramiento	**6%**	**4%**	**2%**
Modelos de asesoramiento	1%	1%	0%
Propósitos y objetivos del asesoramiento	2%	1%	1%
Consideraciones éticas y legales	2%	1%	1%
El equipo de asesoramiento	1%	1%	0%
Asesoramiento individual	**30%**	**10%**	**20%**
…			
Asesoramiento grupal	**10%**	**5%**	**5%**
…			
Desarrollo de carreras	**20%**	**13%**	**7%**
…			
Total del test	**100%**	**42%**	**58%**

Hay que detallar también las partes del constructo a evaluar cuando elaboremos un test en contextos no educativos. Por ejemplo, si queremos medir Dogmatismo, debemos establecer los diversos componentes o manifestaciones del mismo: dogmatismo ante la política, ante la educación de los hijos, ante la religión, en las relaciones familiares, etc. En un test de calidad de vida en diabéticos, el test deberá evaluar los determinantes de la calidad de vida en la población general (la situación familiar, social, laboral…) y, además, los específicos de la población de diabéticos a la que el test va dirigido (tolerancia al tratamiento, temor ante la evolución de la enfermedad…).

Construcción provisional del test

El siguiente paso para la construcción del test es la elaboración de los ítems. Se suele recomendar que al menos se elaboren el doble de ítems de los que se piensa que debiera tener el test. En los apartados sucesivos veremos que algunos ítems serán descartados por no satisfacer los controles de calidad. Al haber elaborado más ítems de los necesarios podremos quedarnos con los mejores. Debemos conseguir un conjunto de ítems ante los que quepa esperar respuestas distintas de los que tengan alto y bajo nivel en el constructo que queremos medir. Si, por ejemplo, pretendemos evaluar la actitud ante la inmigración, un ítem podría requerir al evaluado informar de si está o no de acuerdo con la frase: *La inmigración trae más ventajas que inconvenientes.* Cabe esperar que las personas con actitud positiva estén de acuerdo y las personas con actitud negativa, en desacuerdo. Si queremos medir el dominio de las ecuaciones de primer grado, podríamos proponer como ítem el siguiente: *Obtenga el valor de x en la ecuación* $2x - 4 = 2$. Los que dominen dichas ecuaciones encontrarán la solución correcta y los que no, no. El rendimiento en ambos ítems depende del nivel de dominio del constructo que se quiere evaluar.

Tipos y formatos de ítems

Como hemos visto, podemos hablar de tests de *rendimiento óptimo* y de *rendimiento típico*. En uno de rendimiento óptimo quien responde pretende obtener la máxima puntuación posible. Así ocurre ante un examen, una prueba de aptitudes, un test de inteligencia, etc. En un test de rendimiento típico quien responde debe hacerlo de forma que su puntuación informe de cómo es o de su posición personal sobre lo que se pregunta. En este caso, no tiene sentido hablar de rendimiento máximo. Los tests de rendimiento óptimo y típico difieren en las siguientes 3 características:

1. *El tiempo de administración del test.* En los tests de rendimiento óptimo, hay que determinarlo con mucho cuidado. Vimos en el capítulo precedente que se distingue entre tests de *velocidad* y de *potencia*. En los primeros se fija el tiempo de administración de forma sea muy difícil resolver todos los ítems. Con frecuencia la tarea es muy sencilla (por ejemplo, sumas de un solo dígito) y se trata de ver cuántas sumas es capaz de hacer cada persona en el tiempo disponible. En los tests de potencia, por el contrario, se suele dar tiempo suficiente para que sea posible dar una respuesta meditada a cada ítem. En los de velocidad, lo que importa es saber cuántos ítems es capaz de hacer el evaluado; mientras que en los de potencia se presta especial atención a las características de los ítems que hace. En la práctica, la mayoría de los tests de rendimiento óptimo son de potencia, si bien se fija el tiempo de administración de modo que no sobre demasiado. Eso hace que el test pueda resultar parcialmente de velocidad para los evaluados más lentos. Para considerar a una prueba como un test de potencia suele establecerse que una clara mayoría haya podido dar una respuesta meditada a un 90% o más de los ítems (Schmeiser y Welch, 2006).
   En los tests de rendimiento típico el tiempo de aplicación tiene escasa o nula relevancia. Muchos tests de personalidad, por ejemplo, no tienen un tiempo límite de aplicación y se permite que cada evaluado responda a su ritmo. Se dan a veces instrucciones del tipo "responda sin pensar demasiado" con la idea de obtener la primera respuesta a los ítems, no para indicar que el tiempo de administración es limitado.
2. *Tratamiento de las preguntas dejadas sin contestar.* En un test de rendimiento óptimo, cuando no se responde a un ítem, caben dos posibilidades. Puede ser un ítem que el evaluado no ha tenido tiempo de leer (en lo sucesivo, ítem *no alcanzado*) o puede que no haya querido dar la respuesta tras haberlo leído (*omisión*). Si el test es (puro) de velocidad, cabe suponer que los ítems sin respuesta son ítems no alcanzados. Si fuese (puro) de potencia, cabe suponer que ha tenido tiempo suficiente para estudiar todos los ítems y que ha omitido deliberadamente la respuesta, y son, por tanto, omisiones. Por lo general, como no suele haber tests puros, lo que se suele hacer es considerar como ítems no alcanzados por un evaluado los que siguen a su última respuesta, y como omisiones los no respondidos que preceden a su última respuesta. Tal proceder no está exento de cierta lógica, pero choca con las estrategias de respuestas de algunos evaluados. Por ejemplo, ciertas personas hacen una lectura rápida del test y responden a los ítems que les resultan fáciles. Después, pasan a responder, uno a uno, despacio, hasta donde lleguen. Con esta estrategia, los ítems dejados sin responder, previos al último ítem respondido en la primera pasada rápida, van a ser considerados como omisiones cuando han podido no ser vistos detenidamente (Schmeiser y Welch, 2006).

En los tests de rendimiento típico las no respuestas tienen otro significado. Suelen indicar que el ítem no se entiende o hay desinterés y falta de motivación en el evaluado. Algunos tests dan pautas sobre qué hacer con las no respuestas. Por ejemplo, en el test de los cinco grandes factores de la personalidad NEO-FFI (Costa y McCrae, 1999) se recomienda la no calificación de quien tenga más de 10 ítems sin respuesta en los 60 ítems del test, y se dan pautas concretas de cómo puntuar los ítems dejados sin respuesta cuando son menos de 10.

3. *Respuestas al azar y sesgos de respuesta.* En los tests de rendimiento óptimo con ítems de opción múltiple (en los que se ha de elegir una respuesta entre varias) es posible obtener aciertos, respondiendo al azar, no sabiendo la respuesta correcta. Al final del capítulo estudiaremos distintas estrategias y soluciones a este problema. En los tests de rendimiento típico no caben las respuestas al azar si se está respondiendo al test con seriedad, pero sí podemos encontrar *sesgos* de *respuesta*, como la *tendencia a utilizar las categorías extremas*, la *aquiescencia* y la *deseabilidad social* (Guilford, 1954). En los ítems en los que el evaluado ha de emitir su respuesta eligiendo una categoría, nos podemos encontrar que dos evaluados, de nivel similar de rasgo, difieran en su tendencia al uso de las categorías extremas; uno podría utilizarlas en casi todas sus respuestas, mientras que otro podría no utilizarlas apenas. La aquiescencia es la tendencia a responder afirmativa o negativamente a un ítem independientemente de su contenido. Para evitar este sesgo de respuesta resulta eficaz la redacción de ítems *directos* e *inversos*; en los primeros, se espera una respuesta afirmativa de los que tengan alto nivel de rasgo, mientras que en los segundos, se espera negativa. El problema de la deseabilidad social y del falseamiento de las respuestas en los tests, dada su importancia en determinados contextos de evaluación psicológica, se considera con detalle en el capítulo 15 de este libro.

Formatos de ítems en tests de rendimiento óptimo[1]

En los tests de rendimiento óptimo pueden elaborarse preguntas abiertas (formato de respuesta construida) o preguntas con opciones preestablecidas (formato de respuesta seleccionada). Los dos formatos más comunes de los ítems con respuesta seleccionada son los ítems de verdadero-falso y los de opción múltiple.

a) *Verdadero-falso*: Se muestran dos alternativas y se ha de elegir la que se considera correcta. Por ejemplo, un ítem de un test de Historia Moderna podría ser:

Pi y Margall fue presidente de la 1ª República Española.
V( ) F( )

[1] En este capítulo expondremos los tipos de ítems de respuesta de uso más frecuente, tanto en tests de rendimiento óptimo como típico. Sin embargo, conviene advertir que en los últimos años han surgido formatos innovadores de ítems, por ejemplo los que utilizan las posibilidades del ordenador, y que permiten evaluar constructos que los ítems tradicionales de lápiz y papel no pueden evaluar o no lo hacen con la misma eficacia (Olea, Abad y Barrada, en prensa). En el capítulo 15 se mostrarán ejemplos de estos nuevos formatos.

b) *Opción múltiple*. Un ítem de opción múltiple consta de un enunciado y de tres o más opciones de respuesta, de las que sólo una es correcta. Por ejemplo, un ítem de un test de aptitud verbal puede ser:

Coche es a volante, como bicicleta es a...
a) Pedal
b) Sillín
c) Manillar
d) Cambio

Hay también varios tipos de ítems de respuesta construida (Navas, 2001). Los hay que requieren solo completar una frase (*Las provincias que integran la Comunidad Autónoma de Extremadura son…..*); otros requieren una respuesta más extensa, aunque breve, como responder en un párrafo de pocas líneas; o mucho más extensa, como hacer una redacción o construir una maqueta. Un ítem que requiere una respuesta corta sería *Exponga en no más de 10 líneas las dos principales características de la pintura de Goya*, y una extensa *Detalle la influencia de los escritores latinoamericanos en la novela española del siglo XX*. Otro tipo de examen abierto es el *portafolio*, en el que el evaluado presenta a evaluar un conjunto de trabajos que ha realizado y que considera buenos ejemplos del nivel de aprendizaje que ha alcanzado.

En las décadas 80-90 hubo mucho debate en contextos de evaluación educativa sobre si eran mejor los ítems de respuesta construida o seleccionada. Los partidarios de la respuesta construida decían que sólo este formato permite la evaluación de procesos superiores y que la respuesta seleccionada tiene el problema de los aciertos por azar. Los partidarios de la respuesta seleccionada enfatizaban que este formato muestrea mejor los contenidos, pues pueden hacerse más preguntas, y que la corrección es subjetiva y más costosa en los ítems de respuesta construida. Estudios posteriores han puesto de manifiesto que las respuestas a los ítems abiertos se pueden cuantificar de forma fiable, que con ambos tipos de ítems se puede evaluar procesos de aprendizaje de alto nivel y que ambos formatos proporcionan resultados altamente correlacionados cuando se mide el mismo dominio. Algunos autores enfatizan que no resultan formatos redundantes, pues se suelen medir destrezas distintas (Schmeiser y Welch, 2006). Por tanto, ambos tipos de ítems más que ser excluyentes son complementarios; unos son más apropiados que otros según sean los objetivos concretos del test (Martínez et al. 2005).

Una exposición más detallada de otros formatos alternativos para ítems de respuesta construida y para ítems de respuesta seleccionada puede consultarse en la página web http://www.uam.es/docencia/ace/ y en Martínez, Moreno y Muñiz (2005).

Formatos de ítems en tests de rendimiento típico

Los formatos de respuesta seleccionada más frecuentes en los tests de rendimiento típico son los de opción binaria y categorías ordenadas:

a) *Opción binaria***:** La persona debe elegir entre dos opciones antagónicas: por ejemplo, ante un determinado enunciado, manifestar si está de acuerdo o no, o decir si describe su

modo usual de comportarse. Un ítem de un cuestionario sobre la actitud de los padres hacia los profesores de sus hijos puede ser:

En realidad, los profesores hacen poco más que cuidar de nuestros hijos cuando trabajamos.
Desacuerdo ( ) Acuerdo ( )

b) *Categorías ordenadas*. El formato establece un continuo ordinal de más de dos categorías, que permite a la persona matizar mejor su respuesta. Puede o no incluir una categoría central para indicar la posición intermedia de la escala de respuesta. Por ejemplo, un ítem sobre la actitud de los adolescentes hacia el consumo de drogas podría ser el que sigue:

Las drogas pueden realmente resolver problemas de uno mismo.
( ) Muy en desacuerdo
( ) Bastante en desacuerdo
( ) Neutral
( ) Bastante de acuerdo
( ) Muy de acuerdo

En el ítem precedente *Muy en desacuerdo*, *Bastante en desacuerdo*… serían las etiquetas de las cinco categorías. A veces, se establecen sólo las dos etiquetas extremas del continuo, dejando señaladas las restantes categorías, como muestra la siguiente escala de respuesta:

*(Muy en desacuerdo)*__ __ __ __ __ __ __ *(Muy de acuerdo)*

Hay varios tipos de escalas de respuestas (Morales, Urosa y Blanco, 2003). Las más comunes son la de *grado de acuerdo* y la de *frecuencia*. En la primera, llamada también escala *tipo Likert*, se ha de manifestar el grado de acuerdo con la frase, mientras que en la segunda se ha de indicar la frecuencia del comportamiento descrito en el enunciado. En otras escalas de respuesta se ha de indicar la importancia que se da a lo que indica la frase o cómo de correcta es la descripción que la frase hace de quien responde. Los dos ítems que siguen utilizan la escala de grado de acuerdo y la de frecuencia, respectivamente.

Me encanta Madrid.
En desacuerdo
Indiferente
De acuerdo

Cuido mi alimentación.
Nunca
Algunas veces
Muchas veces
Siempre

Tres asuntos relevantes en relación a los ítems tipo Likert son el número de categorías de la escala de respuesta, la presencia o no de categoría central y la elección de las etiquetas.

Se suele recomendar que el número de categorías sea 5 o un valor próximo (Hernández, Muñiz y García-Cueto, 2000; Morales et al. 2003). No se obtienen mejores tests cuando se utilizan escalas de respuestas con muchas más categorías, pues se producen in-

consistencias en las respuestas. La probabilidad de que una persona elija la misma categoría ante una misma frase, supuesto que no haya cambiado su nivel de rasgo, será mayor si ha de responder con una escala de 5 categorías que con una de 20, de ahí que, cuando son muchas las categorías disponibles, se incremente la inconsistencia. Con sólo dos o tres categorías se puede dificultar la manifestación del auténtico nivel de rasgo. Por ejemplo, dos personas, una que esté muy de acuerdo y otra que esté sólo de acuerdo, tendrán que utilizar la misma categoría si la escala es *En desacuerdo/No sé/ De acuerdo*. En poblaciones especiales, como discapacitados o personas mayores, puede resultar más adecuada una escala de pocas categorías.

En las escalas de grado de acuerdo, no está del todo claro si es mejor fijar un número par o impar de categorías. Hay razones a favor y en contra de la categoría central (que puede etiquetarse como *indiferente*, *neutral*, *dudo*, *no sé*...). Su inclusión permite que alguien que realmente no está de acuerdo ni en desacuerdo con la frase pueda indicarlo. En un ítem sin categoría central tendría que manifestarse como ligeramente de acuerdo o en desacuerdo, cuando su posición ante el enunciado no es esa. Los partidarios de eliminar la categoría central argumentan que con demasiada frecuencia dicha categoría termina siendo la elegida por los que responden con poco cuidado o de forma poco sincera. Los partidarios de un número par de categorías suelen serlo también de un número más alto de categorías, de forma que se pueda entender que el ítem tiene en realidad dos categorías centrales (ligeramente de acuerdo y ligeramente en desacuerdo). La investigación muestra que los indicadores psicométricos de los ítems no dependen de la existencia o no de categoría central cuando el número de categorías es mayor de tres (Morales et al. 2003).

Por último, se han propuesto muchas tandas de etiquetas. Morales et al. (2003, p. 55-58) muestran varias. Las etiquetas han de abarcar todo el continuo (de acuerdo-desacuerdo, frecuencia, importancia…) y además se ha de procurar que el salto en el continuo entre cada dos etiquetas consecutivas sea de similar cuantía. La escala de respuesta *Muy en desacuerdo/En desacuerdo/Indeciso* incumpliría la primera exigencia, pues las personas que estén de acuerdo no tienen una categoría que les permita indicarlo. El ítem que sigue incumple la segunda exigencia, pues la distancia en el continuo entre las dos primeras categorías es menor que la que hay entre la tercera y cuarta.

La Educación está en crisis.
Muy en desacuerdo
En desacuerdo
De acuerdo
Muy de acuerdo

Redacción de ítems de opción múltiple

Se han propuesto conjuntos de recomendaciones para la correcta redacción de los ítems de opción múltiple. Se basan a veces, aunque no siempre, en estudios empíricos en los que se ha comprobado que su incumplimiento genera ítems de peor calidad. Haladyna, Downing y Rodríguez (2002) han propuesto 31 recomendaciones. Moreno, Martínez y Muñiz (2004) las han reelaborado y proponen las siguientes 12, clasificadas en 3 apartados, que reproducimos a continuación con ligeros cambios:

A. Elección del contenido que se desea evaluar.

1. *Cada ítem debe evaluar el contenido de una celdilla de la tabla de especificaciones, lo que garantiza que el test muestreará bien todo el contenido a evaluar. Hay que evitar los ítems triviales.*

2. *El ítem deberá ser sencillo o complejo, concreto o abstracto, memorístico o de razonamiento en función de las destrezas y contenidos que deba evaluar.*

Las dos primeras recomendaciones indican que la creación de los ítems ha de ceñirse a lo estudiado en el primer apartado sobre diseño del test y, en particular, a lo establecido en la tabla de especificaciones.

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,* pues la opción que no concuerda suele ser incorrecta.

4. *La sintaxis o estructura gramatical debe ser correcta. Conviene evitar ítems demasiado escuetos o profusos, ambiguos o confusos. Conviene cuidar especialmente las expresiones negativas* para evitar que puedan ser interpretadas incorrectamente.

5. *La semántica debe estar ajustada al contenido y a la comprensión lingüística de las personas evaluadas.* Si no es así, las respuestas al ítem dependerán del constructo que se pretende medir, como se pretende, pero también de la comprensión lingüística de los evaluados, que no se pretende.

C. Construcción de las opciones.

6. *La opción correcta debe ser sólo una y debe ir acompañada por distractores plausibles.* Si las opciones incorrectas no son plausibles, no sabremos cuantas opciones del ítem están actuando como auténticos distractores.

7. *La opción correcta debe estar repartida entre las distintas ubicaciones*, evitando la tendencia natural a ubicar la opción correcta en las posiciones centrales (Attali y Bar-Hillel, 2003).

8. *Las opciones deben ser preferiblemente tres.* Se han realizado trabajos que prueban que no suelen resultar mejores los ítems de 4 ó 5 opciones que los de 3 (p.ej, Abad, Olea y Ponsoda, 2001). Lo serían si la cuarta, quinta… opción fuesen de la misma calidad que las tres primeras, lo que no es frecuente. Por tanto, suele resultar más apropiado, por ejemplo, un test de 80 ítems de 3 opciones que uno de 40 ítems de 6 opciones, a pesar de que el tiempo dedicado al procesamiento de los ítems sea parecido en ambos casos.

9. *Las opciones deben presentarse usualmente en vertical.* Cuando se presentan en horizontal, una tras otra, es más fácil que alguna no se entienda correctamente.

10. *El conjunto de opciones de cada ítem debe aparecer estructurado*. Por ejemplo, si las opciones fuesen valores numéricos, se recomienda que aparezcan ordenados. La ordenación facilita la correcta comprensión del ítem.

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".* A veces se redactan dos opciones de forma que necesariamente una de las dos es correcta, de lo que se puede inferir que las restantes son incorrectas.

12. *Ninguna opción debe destacar del resto ni en contenido ni en apariencia.* Cuando una opción destaca en contenido o apariencia suele dar pistas sobre si es o no correcta. No es infrecuente encontrarse en un ítem varias opciones poco elaboradas y muy breves, que son incorrectas, y una más elaborada, más extensa, que es la correcta.

Hemos revisado (García, Ponsoda, Sierra, 2009) más de 50 exámenes de opción múltiple con los que se evalúa en la universidad y hemos comprobado que se suelen incumplir algunas de las recomendaciones expuestas. De hecho, hemos encontrado:

- Ítems con ninguna o más de una solución correcta.
- Ítems con demasiado texto. Con el loable propósito de que el estudiante vea el interés e importancia de lo que se pregunta, muchas veces se redactan ítems con mucho más texto del necesario, lo que puede dificultar su comprensión.
- Ítems que dan pistas de la solución correcta. A veces, la pista resulta de la falta de concordancia gramatical entre el enunciado y alguna opción. Otras veces, una opción es mucho más larga y está más elaborada que las demás. En otras ocasiones se ofrecen dos opciones que agotan las posibles respuestas. A veces se proponen ítems que aparecen resueltos en otros ítems del mismo test. En estas situaciones, el rendimiento en el ítem no depende solo del nivel de conocimiento, como debiera ser, sino de la capacidad del estudiante para captar estas pistas.
- Presencia de opciones del tipo "*Ninguna de las anteriores"* y "*Todas las anteriores"*. Muchas veces, por la necesidad de redactar el número de opciones preestablecido, se termina incluyendo una opción de este tipo, seguramente porque requiere menos esfuerzo que elaborar una opción plausible nueva.
- Opciones incorrectas (o *distractores*) poco plausibles. Las opciones incorrectas poco plausibles son poco elegidas y tenemos entonces la duda de cuantas opciones realmente funcionales tiene el ítem. Las opciones incorrectas no deberían descartarse utilizando sólo el sentido común. Las alternativas no ciertas deben ser elegidas entre los errores o confusiones que usualmente tienen las personas que no conocen la respuesta correcta a la pregunta. Otra posible estrategia para generar buenos distractores sería el uso de alternativas de respuesta que son verdaderas para otras preguntas, pero que son inciertas para el enunciado al que se asocian.

Redacción de ítems de categorías ordenadas

Respecto a la manera de formular las cuestiones en los tests de rendimiento típico, se han propuesto algunas recomendaciones que pueden ayudar a su correcta redacción:

1. *Utilizar el tiempo presente*.

2. *Deben ser relevantes*, en el sentido de que su contenido debe relacionarse claramente con el rasgo. Hay que redactar frases ante las que darían respuestas distintas los que tengan alto y bajo nivel en el rasgo que se pretenda evaluar.
3. *Se debe cuidar que el contenido sea claro* y evitar una excesiva generalidad. Resultan mejor los ítems formados por *frases cortas, simples e inteligibles*. Hay que evitar incluir dos contenidos en un ítem.
4. Para minimizar la aquiescencia conviene redactar *ítems de modo directo* e *inverso*.
5. Conviene *evitar el uso de negaciones*, pues dificultan la comprensión de la frase, *y de universales* (todo, siempre, nunca...), pues llevan a casi todos los evaluados a elegir la misma categoría de respuesta. Algunas escalas de Sinceridad utilizan precisamente estos universales para detectar el falseamiento de respuestas. Un enunciado de un posible ítem de una escala de Sinceridad sería *Nunca me ha apetecido hacer algo prohibido*, precisamente con la idea de que los evaluados que respondan sin falsear se habrán de manifestar en desacuerdo con el enunciado.

Se recomienda generar tantos ítems directos como inversos. Lo preferible es que los ítems inversos no lleven negaciones. En un ítem para medir el interés por el estudio, la frase "*Me gusta estudiar"* daría lugar a un ítem directo, y las frases "*No me gusta estudiar*" y "M*e aburre estudiar"* darían lugar a ítems inversos. La última sería preferible a la penúltima pues evita la negación. La presencia de ítems directos e inversos en un test tiene en ocasiones más trascendencia psicométrica de la que aparentemente cabría esperar. Se ha encontrado que la presencia de ítems directos e inversos termina afectando a la estructura interna del test, es decir, a las dimensiones que se miden. Por ejemplo, Tomás y Oliver (1999) comprueban que esto ocurre en el test de Autoestima de Rosenberg. De ahí que haya instrumentos que sólo contienen ítems directos.

Un error que suelen cometer los que tienen poca experiencia en la redacción de ítems es la introducción en la frase de más de un contenido. Por ejemplo, en un ítem hay que manifestarse de acuerdo o en desacuerdo ante el siguiente enunciado "*Pienso que es bueno premiar a los hijos cuando se portan bien y que da mejor resultado que castigarlos cuando hacen algo mal"*. La frase en realidad contiene dos afirmaciones y la respuesta dada a la frase original puede referirse a la primera, a la segunda o a ambas.

Otro error frecuente en la redacción de estos ítems es el uso inadecuado de la escala de respuesta. Por ejemplo, el ítem que sigue estaría mejor redactado con una escala de frecuencia que de grado de acuerdo, como se muestra en la redacción alternativa.

Juego al tenis al menos una vez por semana.

Muy en desacuerdo ( )
En desacuerdo ( )
Indeciso ( )
De acuerdo ( )
Muy de acuerdo ( )

Redacción alternativa:

Indique cuantas veces a la semana, en promedio, juega al tenis.
Ninguna ( )
Una ( )
Dos ( )
Tres o cuatro ( )
Cinco o más ( )

Revisión de los ítems

Una vez elaborados los ítems, resulta muy conveniente que algún experto en el contenido de la prueba y en construcción de tests los revise. Si no es posible recurrir a algún experto, no es mala idea, como sugiere Navas (2001), que sea el mismo redactor de ítems quien haga la revisión, dejando pasar algunos días entre la creación del ítem y su revisión.

Hay que comprobar que cada ítem evalúa los contenidos y destrezas que le corresponden, de acuerdo con la tabla de especificaciones. Se ha de comprobar que no es ambiguo, que gramaticalmente está bien redactado, que el lenguaje no resulta ofensivo y, en los ítems de opción múltiple, que la opción correcta lo es realmente y que todos los distractores son incorrectos. En realidad, se ha de comprobar que cada ítem cumple las recomendaciones que acabamos de ver.

Como vemos, los tests requieren un proceso sistemático de elaboración y una administración controlada. Esto significa, por ejemplo, que una persona deberá obtener la misma puntuación en un test de Responsabilidad independientemente del evaluador que se lo aplique. Con otros métodos de evaluación la puntuación obtenida puede depender más del evaluador. Por ejemplo, distintos psicólogos clínicos pueden llegar a una conclusión diferente respecto de la personalidad de un evaluado tras una entrevista clínica.

Un punto fuerte de los tests es que permiten evaluar a las personas, por ejemplo, únicamente por sus habilidades, conocimientos, competencias o capacidades; es decir, por sus méritos o cualidades y con escasa participación de la subjetividad del evaluador. Siendo esto importante, es si cabe más importante que los tests sean *justos*. Es decir, deben dar al evaluado la puntuación que corresponde a su nivel en el constructo, sea cual sea su edad, género, discapacidad, raza, grupo étnico, nacionalidad, religión, orientación sexual, lengua, y otras características personales. Los expertos deben analizar cada ítem para determinar que cumple lo anterior. En un ejercicio de acceso a la universidad se preguntó por el significado de *pucelana* (natural de Valladolid). Hicieron mejor el ejercicio los seguidores de las crónicas deportivas que los que sabían más Lengua. La revisión mediante expertos hubiese podido detectar que el ítem no era apropiado pues medía, además del conocimiento en Lengua, interés por el fútbol y por tanto resultaba injusto con los estudiantes que no eran aficionados al citado deporte. Existen procedimientos psicométricos para estudiar lo que se denomina como *Funcionamiento Diferencial*, que ayudan a determinar si los ítems y tests son o no justos. Los describiremos en el capítulo 13.

Cuantificación de las respuestas

Una vez decidido el tipo de ítem y el formato de respuesta que se consideran más apropiados, y de cara al estudio psicométrico de la prueba, es preciso decidir la manera de cuantificar las posibles respuestas a los ítems.

Tests de rendimiento óptimo

En general, los ítems de respuesta seleccionada en tests de rendimiento óptimo se cuantificarán con 1, el acierto, y con 0, el error. Se dice que un ítem es *dicotómico* cuando puede tomar sólo dos valores. La puntuación (directa) de un evaluado en el test, X_i, será la suma de las puntuaciones en los J ítems, e indicará su número de aciertos.

$$X_i = \sum_{j=1}^{J} X_{ij} \qquad [2.1]$$

Para la cuantificación de los ítems de respuesta construida breve se recomienda hacer una lista de respuestas aceptables y otra de no aceptables y puntuar con 1 ó 0, respectivamente. Se pueden hacer más de dos listas. Si se hicieran 4, una podría contener las respuestas *muy buenas;* otra, las *buenas;* una tercera, las *regulares;* y una cuarta, las respuestas *incorrectas*. Cada ítem sería cuantificado como 3, 2, 1 ó 0, respectivamente. Los ítems que admiten un número prefijado (mayor de 2) de posibles valores al ser cuantificados se llaman ítems *politómicos*. En este ejemplo, estaríamos ante ítems politómicos que pueden tomar cuatro valores.

En los ítems de respuesta construida extensa, conviene aplicar *rúbricas* (criterios definidos de corrección) para obtener una cuantificación adecuada. Las hay analíticas y holísticas. En las *rúbricas analíticas* se detalla los distintos elementos que hay que valorar en la respuesta, indicando cómo debe ser la respuesta que merezca cada una de las posibles cuantificaciones. Por ejemplo, en la evaluación de una redacción[2] se puede considerar que los elementos a evaluar son a) las ideas y el contenido, b) la organización, c) la fluidez y d) la corrección gramatical. Ante cada elemento, la rúbrica detallaría el rendimiento al que correspondería cada posible puntuación. Ante el elemento "*ideas y contenido"*, la peor calificación correspondería a redacciones que carezcan de idea central o que fuercen al lector a inferir la idea a partir de detalles sueltos. La máxima puntuación correspondería a una redacción clara, interesante y que aborde nítidamente el asunto central, que capture la atención de lector y que proporcione anécdotas enriquecedoras. La puntuación del estudiante en la redacción sería la suma de sus puntuaciones en las cuatro partes que forman la rúbrica. En las *rúbricas holísticas* no se establecen los distintos elementos a evaluar, sino que se evalúa el ítem como un todo. Una buena rúbrica debe proporcionar puntuaciones muy similares al mismo ejercicio cuando es aplicada correctamente por dos evaluadores distintos. Permite que el estudiante sea evaluado en forma objetiva y consistente. Al mis-

[2] Tomado y adaptado de http://web.ccsd.k12.wy.us/RBA/LA/SecSoph.html

mo tiempo, permite al profesor especificar claramente qué espera del estudiante y cuáles son los criterios con los que va a calificar cada respuesta. Livingston (2009) expone las ventajas e inconvenientes de los distintos tipos de rúbricas.

Sea cual sea el tipo de ítem de respuesta construida, la puntuación en el test se obtiene también aplicando la ecuación [2.1], es decir, sumando las puntuaciones obtenidas en los diferentes ítems.

Tests de rendimiento típico

La cuantificación de las respuestas a ítems de pruebas de rendimiento típico requiere ciertos matices. Dado un formato de respuesta determinado, es necesario cuantificar las posibles respuestas a un ítem teniendo en cuenta si es un ítem directo o inverso.

Por ejemplo, en un ítem con formato de respuesta de opción binaria (acuerdo/desacuerdo), cuantificaremos el acuerdo con 2 si el ítem está planteado para medir de manera directa el constructo de interés. Lo cuantificaremos con 1, si está redactado de manera inversa. Se muestran 2 ítems de un cuestionario de actitud ante al aborto voluntario:

Abortar es matar.
En desacuerdo ( ) De acuerdo ( )

La madre es la dueña de su cuerpo en asuntos de aborto.
En desacuerdo ( ) De acuerdo ( )

En el primero, que es inverso, la respuesta "*De acuerdo"* se puntuaría con 1 y "*En desacuerdo"* con 2; ya que estar en desacuerdo con esa afirmación indica una actitud más positiva hacia el aborto voluntario. En el segundo ítem, que es directo, "*De acuerdo"* se puntuaría con 2 y "*En desacuerdo"* con 1; ya que estar de acuerdo con esa afirmación indica una actitud más positiva hacia el aborto voluntario.

Si el formato de respuesta es de K categorías ordenadas, las diversas categorías se cuantificarán normalmente desde 1 hasta K, teniendo en consideración si el ítem es directo o inverso. Por ejemplo, en ítems de 5 categorías, las dos posibles cuantificaciones serán: 1 (*Muy en desacuerdo*), 2… 5 (*Muy de acuerdo*), en un ítem directo; y 5 (*Muy en desacuerdo*), 4… 1 (*Muy de acuerdo*), en un ítem inverso. El ítem podría también cuantificarse utilizando otras tandas de valores (por ejemplo, 0, 1, 2, 3 y 4, ó –2, –1, 0, 1 y 2). En realidad cualquier tanda de cinco valores enteros consecutivos es apropiada y proporciona los mismos resultados psicométricos. Además, la cuantificación de un ítem de opción binaria no ha de ser necesariamente 1 y 2 (podría ser, por ejemplo, 0 y 1). La puntuación de un evaluado en el test se obtiene sumando sus puntuaciones en los ítems (ecuación [2.1]).

Ejemplo 2.1. Cuantificación de ítems de categorías ordenadas

La Tabla 2.2 muestra dos ítems de un test de Calidad de vida, con tres categorías. Se indica la cuantificación apropiada de cada categoría según sea el ítem directo o inverso. El primer ítem es directo y el segundo, inverso.

Tabla 2.2. Cuantificación de dos ítems de categorías ordenadas

	En desacuerdo	Indeciso	De acuerdo
Me siento apoyado por mi familia	1	2	3
Mi vida carece de sentido	3	2	1

Análisis de ítems

Con *análisis de ítems*[3] nos referimos a los procedimientos dirigidos a extraer información sobre su calidad. Estudiaremos procedimientos que permiten seleccionar los ítems más apropiados a los objetivos específicos del test. Después del proceso de análisis de ítems se podrá determinar los ítems que formarán parte del test definitivo, o construir la versión breve o reducida de un instrumento ya en uso. En cualquier caso, vamos a obtener indicadores que no deben interpretarse de forma automática, sino inteligentemente, atendiendo al objetivo específico del test. En contextos de evaluación educativa, por ejemplo, el análisis de ítems permite ir mejorando las preguntas con las que evaluamos y el examen en su conjunto, y nos puede informar sobre qué han aprendido o aprendido mal los estudiantes (Morales, 2009).

Downing y Haladyna (1997) distinguen entre el *análisis cualitativo* de ítems y el *análisis cuantitativo*. El primero precede a la aplicación del test y requiere comprobar, por lo general mediante expertos, que se han realizado adecuadamente las actividades comentadas en los apartados previos. Aplicado el test, se recomienda hacer el análisis cuantitativo. Cuando hablamos de análisis de ítems sin más, nos referimos a éste último. Requiere la obtención para cada ítem de diversos indicadores, que pueden encuadrarse en tres categorías: los de dificultad, los de discriminación y el de validez.

Tras aplicar el test provisional a una muestra de evaluados representativa de la población a la que va dirigida la prueba (se aconseja entre 5 y 10 veces más evaluados[4] que ítems), y una vez cuantificadas las respuestas de cada individuo, se forma una matriz de datos de N filas (evaluados) x J columnas (ítems). El elemento X_{ij} de esta matriz indica el

[3] Tanto la TCT como la TRI proporcionan indicadores de las características psicométricas de los ítems. En este tema estudiaremos los indicadores que aporta la TCT. Los que aporta la TRI se verán al estudiar esta teoría. Un segundo comentario tiene que ver con la ubicación en el libro de este apartado. El análisis de ítems se ocupa del estudio de los ingredientes básicos de los tests. La calidad del todo (el test) depende, como cabe esperar, de la calidad de las partes (los ítems). Es, entonces, inevitable, que en el estudio de los ítems aparezcan conceptos de la calidad del test que estudiaremos en capítulos sucesivos. Por esta razón, en la mayoría de los manuales el análisis de los ítems más bien cierra los libros que los abre. Sin embargo, en el proceso de construcción de un test, el análisis de sus ítems precede a la determinación de los ítems que componen el test definitivo. Nuestra experiencia docente aconseja exponer a los estudiantes este tema al inicio de la materia y no al final, y siguiendo esta lógica hemos preferido mantener esa misma ordenación en el libro.

[4] Varios ejemplos incumplirán esta recomendación. En éste y siguientes capítulos expondremos ejemplos de tests con muy pocos ítems y muy pocos evaluados, muchos menos de los que necesitaría un test real. Son ejemplos pensados para facilitar la comprensión de lo expuesto, que requieren pocos cálculos y no demasiado espacio.

valor obtenido por el evaluado *i* en el ítem *j*. Según la ecuación [2.1], sumando por filas obtendremos las puntuaciones directas (*X*) de los evaluados en el test. La Tabla 2.3 muestra los datos obtenidos por cinco evaluados en un test de rendimiento óptimo de 3 ítems (X_1, X_2 *y* X_3). La columna más a la derecha muestra la puntuación de cada uno en el test (*X*) que es su número de aciertos si los 1 y 0 de la tabla indican acierto y error en el ítem.

Tabla 2.3. Resultados de 5 evaluados en 3 ítems y en el test *X*

X_1	X_2	X_3	X
1	1	0	2
1	0	0	1
0	1	1	2
1	1	0	2
0	1	1	2

En el caso de un test de rendimiento típico, tendríamos una tabla similar. Los datos de 4 evaluados en un test de Autoestima, con 5 ítems tipo Likert de 7 alternativas, podrían organizarse como se muestra en la Tabla 2.4. También en este caso, las puntuaciones en el test resultan de sumar las puntuaciones en los 5 ítems.

Tabla 2.4. Resultados de 4 evaluados en 5 ítems y en el test *X*

X_1	X_2	X_3	X_4	X_5	X
7	5	4	7	6	29
1	1	3	4	2	11
4	6	5	4	3	22
6	6	5	5	7	29

A estas tablas de datos se pueden aplicar los distintos indicadores que informarán de las características psicométricas de los ítems.

Índice de dificultad

Este indicador sirve para cuantificar el grado de dificultad de cada ítem. Se aplica a los ítems dicotómicos de los tests de rendimiento óptimo. El índice de dificultad de un ítem *j*, p_j, se define como la proporción de evaluados que ha acertado el ítem. Es el cociente entre el número de evaluados que lo han acertado (A_j) y el total de evaluados que lo han respondido (N_j).

$$p_j = \frac{A_j}{N_j} \qquad [2.2]$$

Ejemplo 2.2. Obtención del índice de dificultad

Supongamos que 5 evaluados responden a un test de 3 ítems. En la Tabla 2.5 se muestran sus puntuaciones. Nótese que los evaluados 4 y 5 han dejado ítems sin responder.

Tabla 2.5. Puntuaciones de 5 evaluados en 3 ítems y en el test X

Evaluado	X_1	X_2	X_3	X
1	1	1	0	2
2	1	0	0	1
3	0	1	1	2
4	1	-	-	1
5	0	-	1	1

Los índices de dificultad de los tres ítems serán:

$$p_1 = \frac{A_1}{N_1} = \frac{3}{5} = 0{,}60$$

$$p_2 = \frac{A_2}{N_2} = \frac{2}{3} = 0{,}67$$

$$p_3 = \frac{A_3}{N_3} = \frac{2}{4} = 0{,}50$$

En los tests de opción múltiple es posible obtener aciertos respondiendo al azar. En el último apartado veremos los procedimientos que permiten descontar del número de aciertos obtenidos por cada evaluado los que presumiblemente se deben a haber respondido al azar. Algo similar cabe plantearse en relación al índice de dificultad. En un test en el que no haya respuestas al azar tendremos presumiblemente menos aciertos de los que tendríamos en ese mismo test si las ha habido. Se han propuesto fórmulas que corrigen los aciertos debidos a respuestas al azar. El índice de dificultad corregido de un ítem de opción múltiple de K opciones, p_j^c, se obtiene aplicando la siguiente expresión (Schmeiser y Welch, 2006):

$$p_j^c = p_j - \frac{\frac{F_j}{N_j}}{K-1} \qquad [2.3]$$

Donde p_j es el índice de dificultad sin corregir y F_j es el número de personas que fallaron el ítem de los N_j que lo respondieron. Si los ítems del test del Ejemplo 2.2 tuviesen 4 opciones, los nuevos índices de dificultad corregidos serían:

$$p_1^c = p_1 - \frac{\frac{F_1}{N_1}}{K-1} = 0{,}60 - \frac{\frac{2}{5}}{4-1} = 0{,}60 - 0{,}13 = 0{,}47$$

$$p_2^c = p_2 - \frac{\frac{F_2}{N_2}}{K-1} = 0{,}67 - \frac{\frac{1}{3}}{4-1} = 0{,}67 - 0{,}11 = 0{,}56$$

$$p_3^c = p_3 - \frac{\frac{F_3}{N_3}}{K-1} = 0{,}50 - \frac{\frac{2}{4}}{4-1} = 0{,}50 - 0{,}17 = 0{,}33$$

Se observa que al aplicar la fórmula correctora los índices disminuyen cuando hay errores. Croker y Algina (1986) recomiendan que la dificultad media de los ítems sea mayor de 0,5 cuando haya en el test respuestas al azar. Proponen que la dificultad media sea 0,62, 0,67 y 0,75, si los ítems tienen 4, 3 y 2 opciones, respectivamente. Aplicando a estos valores la fórmula [2.3], con $F_j/N_j = 1 - p_j$, se obtiene que en los tres casos p_j^c es 0,50.

Propiedades del índice de dificultad

1. El valor mínimo que puede asumir p es 0 (cuando nadie acierta el ítem) y el valor máximo, 1 (todos los que lo intentan lo aciertan). A medida que p se acerca a 0, el ítem ha resultado más difícil; cuanto más se acerca a 1, ha resultado más fácil. Cuando el valor está cerca de 0,5, el ítem tiene una dificultad media, no ha resultado ni fácil ni difícil. En el Ejemplo 2.2 el ítem más fácil es el 2 y el más difícil, el 3. Nótese, por tanto, que valores altos en el índice de *dificultad*, indican mucha facilidad y no mucha dificultad, como se podría esperar. Algunos (p.e., McAlpine, 2002) prefieren llamar al indicador índice de *facilidad*, pero no termina de prosperar la propuesta.
2. El valor de p depende de la muestra. Un ítem aplicado a una muestra muy preparada (de alto nivel en el rasgo) será acertado por más evaluados que si es aplicado en una muestra poco preparada. Por tratarse del mismo ítem, lo deseable sería que el indicador de su dificultad no dependa de la muestra en la que es aplicado, pero el índice p no tiene esta propiedad. El indicador de la dificultad del ítem dentro de la TRI sí proporciona valores que no dependen del nivel de la muestra en la que se aplique.
3. El valor de p se relaciona con la varianza de los ítems: Si p es 0 ó 1, la varianza del ítem es igual a cero, pues sólo se han producido en el ítem fallos y aciertos, respectivamente. A medida que p se acerca a 0,5, la varianza del ítem aumenta. De hecho, la varianza de un ítem dicotómico puede obtenerse a partir de su índice de dificultad, pues

$S_j^2 = p_j(1-p_j)$. La máxima varianza de un ítem dicotómico (0,25) se alcanza cuando $p = 0{,}5$.

En un test, en el que la puntuación de la persona *i* es la suma de los *J* ítems (ecuación [2.1]), su varianza se puede obtener, a partir de las varianzas y covarianzas de los ítems, mediante la expresión

$$S_X^2 = \sum_{j=1}^{J} S_j^2 + 2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} S_{jj'} = \sum_{j=1}^{J} S_j^2 + 2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} r_{jj'} S_j S_{j'} \qquad [2.4]$$

Es decir, la varianza del test es la suma de las varianzas de los *J* ítems y la suma de las *J(J – 1)* covarianzas que resultan al formar todos los pares posibles con *J* ítems. En la expresión anterior, se ha sustituido la covarianza ($S_{jj'}$) entre cada dos ítems por su valor ($r_{jj'}S_jS_{j'}$), teniendo en cuenta la definición de la correlación de Pearson.

Por tanto, si queremos que el test tenga alta varianza conviene que contenga ítems también de alta varianza; es decir, ítems con índices de dificultad próximos a 0,5.

Al diseñar un test de rendimiento óptimo, se suele recomendar que se sitúen al inicio algunos ítems fáciles, por su efecto motivador (Navas, 2001; García-Cueto y Fidalgo, 2005); en la parte central, los de dificultad media (valores *p* entre 0,30 y 0,70); y al final, los más difíciles. El número de ítems de cada categoría de dificultad que deben incluirse en el test depende de sus objetivos.

En los tests referidos a norma, de poco sirve un ítem con $p = 0$ ó 1, ya que no ayudaría a establecer diferencias entre los evaluados, pues es un ítem que lo fallarían o acertarían todos los evaluados. En un test referido al criterio, puede tener interés saber si todos los evaluados tienen ciertos conocimientos elementales o muy básicos. Si es así, esos ítems tendrán necesariamente altos valores *p* y tendría sentido su inclusión y mantenimiento en el test.

Por tanto, en general, los mejores ítems son los que aportan más varianza al test y son los que tienen valores de *p* medios. De hecho, algunos programas para el análisis psicométrico de los ítems, como TAP (Brooks y Johanson, 2005), recomiendan el estudio detenido y eventual descarte de los ítems con valores *p* mayores de 0,9 y menores de 0,2. Sin embargo, como hemos señalado, puede tener sentido la inclusión y mantenimiento de algunos ítems fáciles o muy fáciles en ciertos contextos aplicados, por ejemplo en tests referidos al criterio.

El índice de dificultad en otros tipos de ítems

Lo visto hasta ahora sobre el índice de dificultad se aplica a ítems dicotómicos de tests de rendimiento óptimo. En este escenario se entiende muy bien que la proporción de personas que aciertan el ítem sea el indicador de su dificultad. El índice *p* es la media aritmética de las puntuaciones conseguidas en el ítem por los *N* evaluados que lo han respondido. Para el caso de ítems no dicotómicos de tests de rendimiento óptimo, la media en el ítem de los evaluados que han respondido sería también el indicador de su dificultad. Supongamos que la rúbrica para corregir un ítem de respuesta construida tiene como valores mínimo y máximo posibles, 0 y 12. Valores medios en el ítem próximos a 0 indicarán dificultad ex-

trema, y próximos a 12, facilidad extrema. Una alternativa al cálculo de la media consiste en dividir la suma de puntos en el ítem por la suma máxima de puntos posible. Procediendo así se obtiene un indicador de la dificultad comprendido entre 0 (máxima dificultad) y 1 (máxima facilidad). Supongamos, por ejemplo, una tarea que puede ser evaluada como 0, 1, 2 ó 3. Si las puntuaciones en la tarea de 5 estudiantes han sido 3, 1, 3, 0 y 3, la suma de puntos sería 10, la suma máxima posible sería 15 y el índice de dificultad será 10/15 = 0,67. Por tanto, la tarea es de dificultad media-baja.

En los tests de rendimiento típico la media de las puntuaciones en el ítem ofrece una información que guarda cierta similaridad con el concepto de dificultad del ítem, aunque no pueda hablarse propiamente de lo difícil que es el ítem. Por ejemplo, en un test de agresividad, un ítem podría ser *Participo en peleas*, con las opciones *Nunca*, *Alguna vez*, *De vez en cuando* y *Con frecuencia*, y recibiría una puntuación de 1 a 4. Un segundo ítem podría ser *Discuto con la gente*, con la misma escala de respuestas. La misma muestra responde a ambos ítems. Supongamos que la media de la muestra en el primero es 1,75 y en el segundo, 2,81. La menor media del ítem 1 indica que hay que tener más agresividad para obtener una puntuación concreta (por ejemplo, 3) en el ítem 1 que en el 2. Cuanto más baja es la media, más nivel de rasgo hace falta para alcanzar una cierta puntuación en el ítem.

Índices de discriminación

Un ítem que mida el constructo de interés debe discriminar entre los que tienen altos y bajos valores en el constructo. Las personas con alta y baja Responsabilidad deberán puntuar de forma diferente en un ítem que realmente mida este constructo, aunque podrán puntuar de forma similar en un ítem que mida otro constructo. Se han propuesto varios indicadores de la discriminación del ítem. Todos ellos requieren una medida apropiada del constructo, que muchas veces, aunque no necesariamente, es la puntuación obtenida en el test completo. Otras veces es un subconjunto de los ítems del test y otras, incluso, una medida del constructo externa al test.

El índice de discriminación

Este indicador se obtiene exclusivamente para ítems dicotómicos. Requiere establecer dos subgrupos de evaluados a partir de sus puntuaciones en el test: el de los que tienen altas y el de los que tienen bajas puntuaciones. Los subgrupos pueden estar compuestos por la mitad de la muestra o, más frecuentemente, por un porcentaje menor (27%, 33%, por lo general) si la muestra tiene suficiente tamaño.

Sea p_s la proporción de personas del subgrupo superior que ha acertado el ítem. Sea p_i la correspondiente proporción en el subgrupo inferior. El índice de discriminación de ítem j, D_j, se define como la diferencia entre ambas proporciones.

$$D_j = p_s - p_i \qquad [2.5]$$

El indicador D toma valores entre –1 y 1. Cuando $D = 1$, todos los evaluados del subgrupo superior han acertado el ítem y ninguno del subgrupo inferior lo ha hecho. Cuando $D = 0$, la proporción de los que han acertado el ítem es la misma en ambos subgrupos. Cuando $D = -1$, ninguno del subgrupo superior ha acertado el ítem y todos los del subgrupo inferior lo han hecho. Valores próximos a cero indican que el ítem no discrimina. Cuanto D más se acerca a uno, mayor es la capacidad discriminativa del ítem. Valores inferiores a 0,20 se consideran valores inaceptables e indican que el ítem ha de ser eliminado (Croker y Algina, 1986). Los valores que puede tomar D dependen del valor del índice de dificultad p (Oosterhof, 1976). En el caso de valores extremos de p, no es posible que D tome valores altos. Por ejemplo, si el valor p de un ítem es 0,98, es evidente que ha debido ser acertado por prácticamente todos los del grupo superior y también por prácticamente todos los del grupo inferior, no pudiendo D tomar un valor alto. Un razonamiento similar puede aplicarse cuando el valor de p es muy bajo. Cuando p toma un valor central es cuando D puede tomar un valor próximo o alejado de cero.

Índices basados en la correlación entre el ítem y el test

Otra estrategia para determinar si un ítem discrimina entre los evaluados que tienen altas y bajas puntuaciones en el constructo consiste en correlacionar las puntuaciones en el ítem con una medida del constructo, que por lo general es el test. Esta estrategia da lugar a los indicadores de discriminación basados en la correlación ítem-test. El indicador D es muy fácil de aplicar, pero normalmente no utiliza toda la información de la muestra, pues sólo entran en su cálculo los evaluados que pertenecen al subgrupo superior o inferior y se aplica sólo a ítems dicotómicos. Los indicadores de discriminación basados en la correlación ítem-test pueden aplicarse a ítems dicotómicos y no dicotómicos, a tests de rendimiento óptimo y típico, y la muestra completa participa en su cómputo.

El índice de discriminación del ítem j basado en la correlación ítem-test, r_{jX}, se define como la correlación de Pearson entre las puntuaciones en el ítem y en el test. Se le suele llamar *correlación ítem-test*.

Ejemplo 2.3. Obtención de la correlación ítem-test en ítems politómicos

Hemos aplicado un test de Satisfacción con los estudios universitarios. En la Tabla 2.6 se muestran las puntuaciones de 4 estudiantes en dos ítems del test y en el test completo X. El ítem 1 es *Organizo actividades extracurriculares* y el 2 es *Asisto a clase regularmente*. Ambos tienen cinco posibles respuestas, puntuadas de 1 a 5: *Muy en desacuerdo* (1), *En desacuerdo*, *Indeciso*, *De acuerdo*, y *Muy de acuerdo* (5).

La correlación ítem-test del ítem 1, que se obtiene calculando la correlación de Pearson entre las columnas 1 y 3, es $r_{1X} = 0{,}638$. La del ítem 2, que resulta de correlacionar las columnas 2 y 3, es $r_{2X} = 0{,}348$.

Tabla 2.6. Puntuaciones en 2 ítems

X_1	X_2	X
3	2	40
2	3	35
5	5	37
1	1	32

En el caso de un ítem dicotómico, podemos obtener la correlación ítem-test por tres procedimientos:

1. El primero consiste en obtener la correlación de Pearson entre la columna de puntuaciones en el ítem y la de puntuaciones en el test, como en el Ejemplo 2.3.
2. La correlación de Pearson entre una variable dicotómica y una continua recibe el nombre de correlación *biserial puntual* (Amón, 1984). Por tanto, un segundo procedimiento consiste en hallar la correlación biserial puntual, r_{bp}, entre el ítem y el test:

$$r_{bp} = \frac{\overline{X}_p - \overline{X}_q}{S_X}\sqrt{pq} \qquad [2.6]$$

   Donde $\overline{X}_p$ y $\overline{X}_q$ son las medias en el test de los que acertaron el ítem y de los que no lo acertaron, S_X es la desviación típica en el test y p es la proporción de evaluados que acertó el ítem. Por último, $q = 1 - p$.
3. Un tercer procedimiento a aplicar cuando el ítem es dicotómico es la correlación *biserial*, r_b. Se puede aplicar cuando una variable es continua (puntuaciones en el test) y otra es dicotómica (el ítem), pero la variable dicotómica se considera como el resultado de dicotomizar una variable continua. La correlación biserial es una estimación de lo que sería la correlación de Pearson entre ambas variables continuas (Amón, 1984).

$$r_b = \frac{\overline{X}_p - \overline{X}_q}{S_X}\frac{pq}{y} \qquad [2.7]$$

   El único elemento nuevo, y, es la ordenada que corresponde en la distribución normal a la puntuación que deja a su izquierda la probabilidad p.
   La correlación biserial puntual, en valor absoluto, es menor que la biserial para unos mismos datos. De hecho, $r_{bp} < 0{,}8r_b$ (Lord y Novick, 1968, p. 340).

Cuando un test tiene un número pequeño de ítems, resulta más apropiado obtener la *correlación ítem-test corregida,* r_{jX}^{c}, o *correlación del ítem con el resto del test*. Consiste en correlacionar las puntuaciones en un ítem con las puntuaciones en el total del test después de restarle las puntuaciones del ítem cuyo indicador queremos obtener. La correlación entre un ítem y el resto del test suele ser inferior a su correlación ítem-test, pues en este caso

se correlaciona una variable (el ítem) con otra (el test) en la que la primera variable está contenida. La correlación entre el ítem y el test puede ser artificialmente alta, por lo indicado, especialmente cuando el test tiene pocos ítems. Izard (2005) considera que el efecto es depreciable cuando el test tiene más de 20 ítems.

Ejemplo 2.4. Obtención de las correlaciones ítem-test e ítem-resto del test[5]

Hemos aplicado un test de cuatro ítems a cinco estudiantes. Sus puntuaciones se muestran en la Tabla 2.7.

Tabla 2.7. Puntuaciones en 4 ítems de un test

X_1	X_2	X_3	X_4	X
0	1	1	0	2
1	1	1	1	4
1	0	1	1	3
0	1	1	1	3
1	1	0	1	3

La correlación de Pearson entre el ítem X_1 y el test X es 0,645. Aplicando la fórmula [2.6] se llega al mismo resultado:

$$r_{bp} = \frac{\overline{X}_p - \overline{X}_q}{S_X}\sqrt{pq} = \frac{(10/3)-(5/2)}{\sqrt{0,4}}\sqrt{(3/5)(2/5)} = 0,645$$

Para X_1, la correlación biserial es:

$$r_b = \frac{\overline{X}_p - \overline{X}_q}{S_X}\frac{pq}{y} = \frac{(10/3)-(5/2)}{\sqrt{0,4}}\ \frac{(3/5)(2/5)}{0,3863} = 0,819.$$

Comprobamos que $r_{bp} < 0,8r_b = (0,8)(0,819) = 0,655$.

Sumando las puntuaciones en los ítems 2, 3 y 4, obtenemos las puntuaciones en el resto del test para el ítem 1. Al correlacionar el ítem 1 con el resto del test para ese ítem (las puntuaciones de los cinco evaluados serían, respectivamente, 2, 3, 2, 3 y 2) se obtiene la correlación ítem-test corregida o correlación ítem-resto del test para el ítem 1, que es $r^c_{1X} = -0,167$. Nótese el fuerte descenso en el valor de la correlación (de 0,645 a –0,167), pues el test tiene sólo 4 ítems.

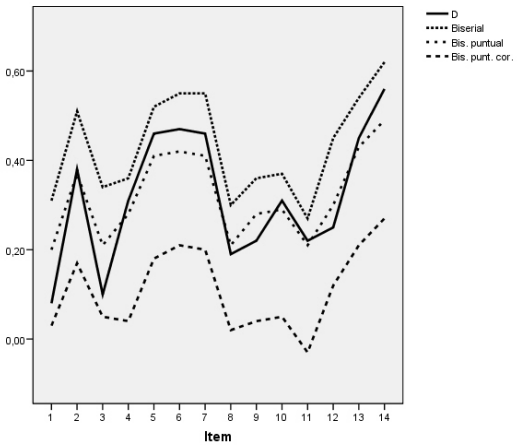
[5] Los indicadores propuestos se obtienen mediante los programas de ordenador que se describen en el Apéndice. En el ejemplo 2.4 se detalla el cálculo de los indicadores para facilitar la comprensión de las fórmulas. El valor de la ordenada *y* puede extraerse de las tablas de la curva normal (p.ej. Amón, 1984) o calcularse directamente.

Se han propuesto otros muchos indicadores de discriminación. Oosterhof (1976) comparó 19 de ellos tras aplicarlos a 50 ítems. Comprobó que la ordenación (en discriminación) que hacían los 19 indicadores de los 50 items básicamente coincidía. De hecho, obtuvo que la mayoría de las correlaciones entre los órdenes superaron el valor 0,90 (la menor correlación fue 0,616). Veamos en el siguiente ejemplo la relación entre los cuatro indicadores de discriminación que hemos descrito.

Ejemplo 2.5. Comparación entre los índices de discriminación

Hemos aplicado los cuatro índices de discriminación (D, correlación biserial, correlación biserial puntual y correlación biserial puntual corregida) a los 14 ítems dicotómicos de un test. La Figura 2.1 muestra los valores obtenidos.

Figura 2.1. Indicadores de discriminación de 14 ítems



La gráfica muestra que los tres indicadores de discriminación que se basan en la correlación entre el ítem y el test (las líneas de trazo discontinuo) dan valores ordenados: los valores más altos corresponden a la correlación biserial, los medios a la biserial puntual y los menores a la biserial puntual corregida. El indicador *D* (trazo continuo) en estos datos da valores similares a la biserial puntual. Al correlacionar entre sí los valores de los 3 indicadores basados en la correlación, la menor correlación es 0,96 (entre la biserial puntual y la biserial puntual corregida), lo que muestra que la ordenacion de los 14 ítems sería básicamente la misma con cualquiera de ellos. Las correlaciones de estos 3 indicadores con *D* son 0,88 (con la biserial puntual corregida), 0,91 (biserial) y 0,96 (biserial puntual). Por tanto, en este ejemplo, se confirma la conclusión alcanzada por Oosterhof (1976) en el sentido de que los distintos indicadores aplicados a unos mismos ítems producen una ordenación similar de sus capacidades de discriminación.

Propiedades de las correlaciones ítem-test e ítem-resto del test

1. La correlación de Pearson (y, por tanto, la correlación biserial puntual) toma valores entre –1 y 1. La correlación biserial puede valer más de 1 o menos de –1. Los indicadores de la discriminación basados en la correlaciones ítem-test nos informan de si el ítem está midiendo lo mismo que la prueba globalmente; es decir, del grado en que el ítem contribuye a medir lo mismo que mide el test. Los ítems con correlaciones nulas miden algo diferente a lo que refleja la prueba en su conjunto. Si con el test se pretende evaluar un rasgo o constructo unitario, debiera considerarse la posible eliminación de los ítems con correlaciones próximas a cero.
   Cuanto más se acerque a 1 el índice, mejor discrimina el ítem entre los que tienen puntuaciones altas y bajas en el test. En el Ejemplo 2.3 vimos que la correlación ítem-test del ítem *Organizo actividades extracurriculares* es mayor que la del ítem *Asisto a clase regularmente*, por lo que el primer ítem discrimina mejor entre los que tienen alta y baja satisfacción con los estudios que el segundo. Es decir, la satisfacción con los estudios se relaciona más con la organización de actividades extracurriculares que con la asistencia a clase.
2. Cuando la correlación ítem-test es negativa y de entidad, debemos cuestionar la cuantificación que se ha aplicado al ítem. Se ha podido proponer como opción correcta una que no lo es, o se ha podido tomar el ítem como directo cuando es inverso, o viceversa.
3. Crocker y Algina (1986) proponen que se tome como criterio mínimo de retención del ítem que la correlación sea significativamente distinta de cero. Dado que una correlación de Pearson puede considerarse significativa cuando $\left|r_{xy}\right|\sqrt{N} > z_{1-\alpha/2}$, (Pardo, Ruiz y San Martín, 2009, p. 350), con una muestra de 100 personas y nivel de confianza del 95%, se llega a la referencia 0,20[6]. Valores de *r* menores de 0,2 nos llevarían a descartar el ítem. Schmeiser y Welch (2006) coinciden en que para un test normativo son deseables índices de discriminación superiores a 0,2. Kehoe (1995) fija la referencia en 0,15.

Ejemplo 2.6. Ejemplo de posible cuantificación incorrecta

En una escala de Romanticismo, que hicieron nuestros estudiantes como práctica para aprender a construir un test de rendimiento típico, las correlaciones ítem-test corregidas de varios ítems se muestran en la Tabla 2.8. Se indica también qué ítems consideraron directos e inversos (D e I). Las correlaciones que se exponen son las halladas tras recodificar los ítems inversos, como se indicó en el apartado sobre cuantificación de las respuestas.

[6] De $r_{xy}\sqrt{100} > z_{0.975} = 1{,}96$, se sigue que $r_{xy} > 0{,}196 \approx 0.20$.

Tabla 2.8. Correlaciones ítem-test corregidas de 5 ítems

	Escala de Romanticismo	r^c
D	*El amor es la razón de mi vida*	0,60
I	*Preferiría que el/ella se me declarara por teléfono y sin rodeos*	0,10
D	*Siempre que puedo, suelo sorprenderle con detalles inesperados*	0,58
D	*Si me invita la primera noche a su casa, no vuelvo a mirarle a la cara*	–0,27
I	*Nunca me identifico con personajes de películas o cuentos*	0,47

Tres ítems, dos directos y uno inverso, tienen valores altos de la correlación entre el ítem y el resto del test (mayores de 0,47). Según el enunciado de los ítems, vemos que los muy románticos están de acuerdo en que el amor es la razón de sus vidas y que siempre que pueden sorprenden con regalos. Los muy románticos están en desacuerdo con el último ítem, pues es inverso. Hay que entender entonces que se identifican con personajes de películas y cuentos. El índice ítem-resto del test es cercano a cero (0,1) en el segundo ítem, lo que indica que los muy románticos no estarían especialmente de acuerdo ni en desacuerdo con ese ítem. De hecho, ¿por qué habrían de preferir los muy o poco románticos la declaración por teléfono? Por último, en un ítem, considerado por los estudiantes directo, se obtiene una correlación ítem-resto del test negativa y de cierta entidad (–0,27). Los estudiantes consideraron, al etiquetar el ítem como directo, que las personas muy románticas debían estar de acuerdo con el enunciado y recibir con disgusto una invitación a subir a casa al poco de conocerse. El análisis psicométrico revela que en la muestra en la que se aplicó el test (estudiantes universitarios) no es así, sino al contrario. En este caso convendría plantearse considerar el ítem como inverso y repetir el análisis psicométrico de todos los ítems tras recodificarlo como inverso.

Índice de validez

A veces aplicamos tests no tanto porque estemos interesados en evaluar directamente el constructo que el test mide, sino porque sabemos que sus puntuaciones predicen bien una variable que interesa pronosticar. En un proceso de selección de personal, podemos aplicar un test de Responsabilidad no porque estemos directamente interesados en conocer las puntuaciones de los candidatos, sino porque se sabe (Salgado y Moscoso, 2008) que las puntuaciones en Responsabilidad ayudan a predecir el desempeño laboral. En el tema 5 estudiaremos los detalles dentro del apartado sobre evidencias de validez referida al criterio. Se suele llamar *criterio* a la variable que queremos predecir y nos solemos referir a ella con la letra *Y*.

Se llama índice de *validez* de un ítem *j*, r_{jY}, a la correlación[7] entre las puntuaciones en el ítem y el criterio externo *Y*. Por ser r_{jY} un coeficiente de correlación, toma valores entre –1 y 1, y elevado al cuadrado indica la proporción de la varianza de *Y* que puede explicar-

[7] Lo ordinario es aplicar la correlación de Pearson, pero en ocasiones otras correlaciones pueden resultar más apropiadas para indicar la relación entre el ítem y el criterio. Si no se especifica nada más, se entiende que hablamos de la correlación de Pearson.

se por el ítem[8]. Cuanto más alejado de cero esté, más fuerte es la relación y mayor la capacidad predictora del ítem en relación al criterio *Y*. La capacidad predictora del ítem no depende del signo de la correlación. Si el índice de validez de un ítem con un criterio de Puntualidad fuese positivo (de 0,25, por ejemplo), es muy posible que el índice de validez de ese mismo ítem con otro criterio, como Absentismo laboral, sea negativo; dada la relación inversa que cabe esperar entre Puntualidad y Absentismo.

Ejemplo 2.7. Cálculo del índice de validez

Supongamos que las puntuaciones de 5 personas en Desempeño laboral son las que aparecen en la columna *Y* de la Tabla 2.9. Queremos construir un test de Responsabilidad que pronostique las puntuaciones en el criterio *Y*. La tabla muestra además las puntuaciones de las 5 personas en los tres ítems del test y en el test completo *X*.

Tabla 2.9. Puntuaciones de 5 evaluados en 3 ítems, el test *X*, y un criterio *Y*

X_1	X_2	X_3	X	Y
2	3	5	10	8
3	1	0	4	2
0	4	5	9	2
5	1	0	6	4
4	3	0	7	5

Calculando la correlación de Pearson entre cada ítem y la columna *Y* se obtienen los índices de validez, que son 0,167 (ítem 1), 0,195 (ítem 2) y 0,293 (ítem 3). El ítem 3 tiene una relación más fuerte con el criterio que los otros dos.

El índice de validez informa de la relación entre el ítem y el criterio *Y*. El concepto análogo, pero referido al test, es el *coeficiente de validez*, que estudiaremos en el tema 5. El coeficiente de validez de un test *X* en relación a un criterio *Y*, r_{XY}, se puede obtener mediante la expresión (Lord y Novick, 1968, p. 332):

$$r_{XY} = \frac{\sum_{j=1}^{J} S_j r_{jY}}{\sum_{j=1}^{J} S_j r_{jX}} \qquad [2.8]$$

[8] Lo habitual es que el índice de validez de un ítem sea menor que sus índices de discriminación basados en la correlación ítem-test, pues lo normal es que el ítem correlacione más con el test para el que se ha construido que con un criterio externo. Los índices de validez suelen ser especialmente bajos (próximos a cero) cuando los ítems son dicotómicos.

La expresión anterior permite obtener la capacidad predictora del test respecto al criterio *Y* a partir de las propiedades (la desviación típica, la correlación ítem-test y el índice de validez) de los *J* ítems que forman el test. Nos puede facilitar la selección de los ítems que más ayuden a construir un test con máxima capacidad predictiva del criterio *Y*.

Ejemplo 2.8. Relación entre el coeficiente de validez y los índices de validez

En el Ejemplo 2.7, si calculamos la correlación de Pearson entre las columnas *X* e *Y*, se obtiene el coeficiente de validez del test formado por los tres ítems, que es $r_{XY} = 0{,}580$.

Tabla 2.10. Datos descriptivos para 3 ítems

	S_j	r_{jX}	r_{jY}	$S_j r_{jX}$	$S_j r_{jY}$
X_1	1,924	–0,588	0,167	–1,131	0,321
X_2	1,342	0,827	0,195	1,110	0,262
X_3	2,739	0,879	0,293	2,408	0,802

A partir de los datos de la tabla podemos comprobar que la fórmula [2.8] proporciona ese mismo resultado:

$$r_{XY} = \frac{\sum_{j=1}^{3} S_j r_{jY}}{\sum_{j=1}^{3} S_j r_{jX}} = \frac{0{,}321 + 0{,}262 + 0{,}802}{-1{,}131 + 1{,}110 + 2{,}408} = 0{,}580.$$

Siguiendo a Lord y Novick (1968) y a Muñiz (1992), entre otros, hemos definido el índice de validez de un ítem como la correlación de Pearson entre el ítem y el criterio *Y*. otros autores, por ejemplo Crocker y Algina (1986) y Gulliksen (1987) definen el índice de validez como dicha correlación multiplicada por la desviación típica del ítem. Análogamente, estos autores definen el *índice de fiabilidad* del ítem como la correlación ítem-test multiplicada por la desviación típica del ítem. La fiabilidad de un test es un concepto psicométrico que se estudiará en el tema siguiente y que nos indica su capacidad para dar puntuaciones similares a personas con el mismo nivel en el rasgo. El índice de fiabilidad de un ítem informa de la aportación del ítem a la fiabilidad del test. Por tanto, siguiendo estas definiciones, el coeficiente de validez del test tiene en el numerador la suma de los índices de validez de los ítems que forman el test y en el denominador la suma de los índices de fiabilidad. Es, por tanto, evidente que si queremos un test que pronostique bien el criterio debemos seleccionar los ítems con altos índices de validez y/o bajos índices de fiabilidad. La situación es paradójica (Muñiz, 1992), pues nos indica que podríamos conseguir mejorar la capacidad predictora de un test por la vía de seleccionar ítems que correlacionen menos con el test total (es decir, disminuyendo una propiedad positiva de un test,

como es su fiabilidad). Lo expuesto muestra que no siempre los ítems con mayores índices de discriminación resultan los más apropiados a los objetivos específicos del test. Visto de otro modo, al eliminar ítems con bajas correlaciones ítem-test, con el propósito de maximizar la fiabilidad del test, seguramente afectaremos negativamente a su coeficiente de validez (Izard, 2005).

Consideraciones adicionales sobre el análisis de ítems

Livingston (2006) y Schmeiser y Welch (2006) señalan otros asuntos a tener en cuenta para un correcto análisis de ítems. El análisis de los ítems se complica en los tests de velocidad. En los tests de rendimiento óptimo, si los evaluados no han tenido tiempo para dar una respuesta meditada a todos los ítems, los que estén al final serán los que resulten más afectados. En estos ítems tendremos respuestas meditadas y respuestas casi aleatorias, lo que no ocurrirá en los que se encuentren al principio. El índice de dificultad por tanto resultará afectado por la posición que ocupa el ítem en el test. En el apartado sobre formatos y tipos de ítems se ha expuesto la norma a seguir para convertir los ítems sin respuesta en ítems no alcanzados o en omisiones. El Ejemplo 2.9 muestra su impacto en los índices de dificultad de los ítems.

Ejemplo 2.9. Índices de dificultad e ítems sin respuestas

La aplicación de la regla para considerar un ítem sin respuesta como omisión o como valor perdido a los datos del Ejemplo 2.2 daría lugar a la Tabla 2.11. Hay tres ítems dejados sin responder. En el caso del evaluado 4 sus dos ítems dejados sin responder siguen a su única respuesta, luego habrían de clasificarse como ítems no alcanzados o valores perdidos (y no se convertirían en errores). En el caso del evaluado 5, el ítem dejado sin responder tiene detrás un ítem con respuesta y por tanto debe ser clasificado como omisión (y convertido en error). Se indica en la tabla con la cuantificación de 0 entre paréntesis.

Tabla 2.11. Puntuaciones en 3 ítems

X_1	X_2	X_3	X
1	1	0	2
1	0	0	1
0	1	1	2
1	-	-	1
0	(0)	1	1

Los índices de dificultad de los ítems 1 y 3 no cambian, pero sí el del ítem 2, que pasará a ser $p_2 = A_2/N_2 = 2/4 = 0{,}5$, en vez de 0,67.

Otro asunto a considerar es el de la posible multidimensionalidad del test. Vamos a ver en temas posteriores procedimientos para detectar si tras las puntuaciones en el test hay sólo una dimensión (lo responsable que una persona es, por ejemplo), dos dimensiones (lo responsable y lo emocionalmente estable, por ejemplo) o más. En el caso de tests multidimensionales tiene más sentido analizar conjuntamente los ítems que se relacionan con cada dimensión, que un análisis conjunto de todos ellos. En el caso de tests educativos, Kehoe (1995) recomienda explícitamente que sólo se haga el análisis conjunto de los ítems que evalúen un material homogéneo (es decir, un material en el que es poco probable que un estudiante lo haga bien en una parte y mal en otra). Si el material a evaluar no fuese homogéneo, habría que hacer un análisis conjunto de los ítems de cada bloque homogéneo de contenidos.

La estrategia anterior puede llevar a tener que hacer el análisis de un conjunto muy reducido de ítems, lo que también plantea problemas. Para Livingston (2006) un análisis de 20 ítems puede ser adecuado; pero de 10, quizás no. Cuando hay pocos ítems el impacto de uno en el test puede ser fuerte. Hemos visto procedimientos para corregir ese impacto, como la correlación ítem-test corregida, pero este indicador tiene el inconveniente de que se correlaciona cada ítem con un test diferente (el test menos el ítem del que estamos hallando el indicador), lo que dificulta la comparación de los índices de los distintos ítems.

Otro asunto a tener en cuenta es la presencia de ítems de baja calidad en el test. Si un test tiene sólo algún ítem deficiente, la correlación del ítem deficiente con el test nos dirá que efectivamente lo es. Si el test tuviese muchos ítems deficientes, la correlación podría no decir demasiado, ¡podría incluso informar erróneamente de la calidad de los buenos ítems! En los procedimientos para el estudio del funcionamiento diferencial de los ítems es habitual generar una medida del constructo de interés que se va progresivamente depurando; es decir, de la que se van eliminando los ítems que parecen no medir lo que miden los demás. Algo similar cabría hacer en el análisis de ítems, para que la medida del constructo no esté contaminada por los ítems deficientes.

Un último asunto tiene que ver con las características de la muestra de evaluados en la que obtenemos los indicadores. Preparamos un examen, lo aplicamos y hacemos el correspondiente análisis de ítems. ¿Estamos seguros de que un ítem que resulte fácil (al corresponderle, por ejemplo, un valor $p = 0{,}80$) volverá a ser fácil si lo volviésemos a aplicar? ¿Estamos seguros de que un ítem con una correlación ítem-test negativa volverá a dar un índice negativo en otra aplicación? La respuesta a estas preguntas requiere, al menos, dos consideraciones. La primera es que cabe sólo esperar valores similares cuando las dos muestras de estudiantes tengan características similares. Si una muestra tuviese un nivel alto de conocimiento y otra un nivel bajo, evidentemente, no cabe esperar que el índice de dificultad de un ítem sea igual en ambas aplicaciones. Aceptando que las dos muestras tengan similares características, hay que tener en cuenta el tamaño de la muestra. El índice de dificultad, el de discriminación, etc. son indicadores que fluctúan muestralmente. Supongamos, por ejemplo, que un ítem de Matemáticas tiene un índice de dificultad de 0,6 al ser aplicado a todos los estudiantes de la Comunidad de Madrid. Si lo aplicásemos a dos muestras de 100 estudiantes extraídos al azar de la citada población, muy probablemente no obtendremos que sea acertado por un mismo número de estudiantes en ambas muestras. Es probable que en ninguna de las dos sea acertado exactamente por 60 estudiantes. Los posibles valores del índice de dificultad vendrían determinados por la distribución muestral de la proporción. Por lo tanto, cuanto menor sea el tamaño de la muestra en la

que se aplica el test, menos debemos fiarnos de los particulares valores de los indicadores, y tanto más probable es que, de haber aplicado el test a otra muestra idéntica, obtengamos resultados diferentes.

En un estudio de simulación[9] hemos comprobado que cuando se aplica un test de 20 ítems a muestras de 50 personas simuladas extraídas de la misma población los índices de dificultad de los ítems de una muestra difieren poco de los obtenidos en las demás. Esto no ocurre, sin embargo, con las correlaciones ítem-test corregidas. De hecho, para que las correlaciones ítem-test corregidas sean similares en distintas muestras, deben estar formadas al menos por 400 personas simuladas.

Conviene, por tanto, que la muestra en la que aplicamos el test tenga un tamaño razonable si se quiere extrapolar a otras aplicaciones los resultados obtenidos en un análisis de ítems. Morales (2009) recomienda muestras de 400 estudiantes o más. Crocker y Algina (1986) sugieren que no tengan menos de 200 evaluados y recomiendan, si el tamaño muestral lo permite, que se haga el análisis de ítems sobre una mitad de la muestra y se informe de los indicadores de los ítems y del test con los datos de la otra mitad[10].

Burton (2001a) concluye que tanto el índice *D* como las correlaciones ítem-test son muy poco estables, a no ser que se obtengan en muestras mucho mayores de las habituales en los contextos educativos. Su utilidad debería limitarse a comprobar las características de los ítems que resultan diagnosticados como muy buenos o muy malos. Éstos últimos son los más interesantes porque pueden revelar que hay algún error en la clave de respuestas. Concluye que hay que quitar importancia a la discriminación de los ítems en la evaluación de la calidad de los exámenes.

Análisis de las opciones incorrectas de respuesta

En relación con el análisis de los ítems se encuentra también el estudio de los patrones de respuesta que se dan a las diferentes opciones de los ítems de opción múltiple. Un modelo muy simple, que desarrollaremos más extensamente en el apartado final de este tema, de cómo una persona responde a un ítem de opción múltiple, supone que:

1. La persona conoce la opción correcta o no la conoce. Si la conoce, responde y acierta necesariamente. Es decir, no se contempla la posibilidad de que conociendo la respuesta, por despiste u otras razones, pueda seleccionar una opción incorrecta.
2. Si no la conoce, tiene dos opciones: puede no responder o puede responder al azar entre las *K* opciones disponibles. Es este caso, se supone que elige las opciones con equiprobabilidad y por tanto la probabilidad de acierto es $1/K$ y la de fallo es $(K - 1)/K$.

Supongamos que 300 personas responden a un ítem de opción múltiple con 4 opciones (A, B, C y D) siguiendo el modelo anterior. Supongamos que ninguno sabe la respuesta correcta (la B, en nuestro caso marcada con un asterisco). Según el modelo, cada evaluado tendrá que responder al azar y la probabilidad de elegir cada opción es 1/4. Por tanto, el

[9] Los detalles pueden solicitarse a los autores.

[10] Esta es una estrategia común de control de lo que se viene llamando "capitalización en el azar" o "sobreaprendizaje". Tal estrategia reduce el efecto de las singularidades de la muestra en los valores de los indicadores.

número esperado de personas que deberá elegir cada opción es 300(1/4) = 75, como muestra la siguiente tabla:

	A	B*	C	D
Frecuencia esperada	75	75	75	75

Supongamos que 100 de los 300 saben la respuesta. Según el modelo, esos 100 elegirán la opción correcta, B. Los restantes 200, al no saber la respuesta, elegirán al azar una de las cuatro opciones con equiprobabilidad. Las frecuencias esperadas de las 4 opciones se muestran en la siguiente tabla:

	A	B*	C	D
Frecuencia esperada	50	100 + 50	50	50

Si supiesen 200 la respuesta correcta, la correspondiente tabla sería:

	A	B*	C	D
Frecuencia esperada	25	200 + 25	25	25

Por último, si los 300 saben la respuesta, la tabla resultante sería:

	A	B*	C	D
Frecuencia esperada	0	300	0	0

Por tanto, en un ítem en el que se responde según el modelo expuesto, debe ocurrir que: 1) la alternativa correcta sea la más seleccionada, y 2) que las alternativas incorrectas lo sean por un número similar de personas. Estas dos circunstancias se cumplen exactamente en las tablas precedentes. En la aplicación real de un ítem no cabe esperar que la frecuencia de elección de las alternativas incorrectas coincida exactamente. Lo que sí debiera ocurrir es que se dé aproximadamente el patrón descrito.

Ejemplo 2.10. Estudio de las opciones incorrectas de respuesta

Observemos los porcentajes de elección en las cinco opciones de tres ítems que se presentan en la Tabla 12.2. El patrón de respuestas obtenido para el ítem 1 es adecuado, pues la mayor parte de la muestra selecciona la alternativa correcta, mientras que las incorrectas son seleccionadas por un porcentaje parecido de personas. El ítem 2 no sería muy adecuado, pues la muestra selecciona en mayor grado una alternativa incorrecta (la *A*) como correcta; al menos, debería pensarse en reformular esa alternativa incorrecta. En el ítem 3, dos alternativas incorrectas apenas son seleccionadas, con lo que se consideran como alternativas no funcionales. Habría que reformular esas dos opciones de respuesta.

Tabla 2.12. Porcentajes de elección de las opciones en 3 ítems

	Opción correcta	*Porcentaje elección de las opciones*				
		A	*B*	*C*	*D*	*E*
X_1	*B*	17	40	14	13	16
X_2	*C*	35	15	21	17	12
X_3	*A*	60	1	21	18	0

Las opciones que no son elegidas tienen especial importancia, pues esos ítems tienen *K* opciones, pero funcionalmente tienen menos. Esta situación plantea dudas sobre el proceder adecuado cuando hay que aplicar fórmulas que requieren especificar el número de opciones; por ejemplo, a la hora de obtener el valor de *p* corregido por azar, aplicando la expresión [2.3].

Un comentario sobre la adecuación al modelo expuesto. Hemos propuesto que hay que elegir con cuidado los distractores. Hemos propuesto incluso que una buena estrategia es proponer como distractores los errores que comenten los estudiantes. Por ejemplo, un ítem de Matemáticas podría ser este:

¿Cuál es el resultado de la operación $6 + (2–3)^3$*?*
a) –13
b) 5
c) 7

La opción correcta es la *b*. El distractor *a* resulta de la operación $6 + (2^3–3^3)$ y el distractor *c,* de $6 + (1)^3$. Un test construido con ítems así permitiría conocer el nivel de cada estudiante en Matemáticas, pero no sólo eso. Los distractores elegidos darían pistas de qué no saben, qué tienen los estudiantes mal aprendido. Pero, ¿no es esto contradictorio con el modelo propuesto? Según el modelo, los evaluados que han elegido las opciones incorrectas lo han hecho porque no sabían la respuesta correcta, han decidido no omitir y han respondido al azar entre todas las opciones disponibles con equiprobabilidad. ¿Qué podemos concluir cuando una opción no es elegida, o una lo es más que la opción correcta? Una primera conclusión es que el modelo no se ha cumplido. Si los errores, como plantea el modelo, son exclusivamente resultado de las respuestas al azar, no se puede explicar que una opción no sea elegida por nadie y otra, por muchos.

Un modelo alternativo es que los estudiantes, cuando se penalizan los errores, no responden al azar sino que eligen la opción que creen correcta. Algunos eligen la realmente correcta, y otros, que saben menos, eligen la opción incorrecta que consideran correcta. Según este modelo, es posible que en una pregunta difícil sólo unos pocos elijan la opción correcta y la mayoría se decante por las distintas opciones incorrectas, que no necesariamente habrían de ser igual de atractivas. Cada distractor plantea una solución considerada correcta por los que saben poco y la frecuencia de elección de cada una indicaría qué proporción de estudiantes tiene el correspondiente aprendizaje incorrecto. Por tanto, no cabe esperar que la proporción de evaluados que tengan el conocimiento erróneo que les lleva al distractor *a* tenga que ser similar que la proporción de los que tengan el aprendizaje incorrecto que lleva al *c*, y tampoco que tenga que ser menor que la proporción de estudian-

tes que saben la respuesta correcta. Kehoe (1995) realiza las siguientes recomendaciones en relación a cómo se ha de proceder tras el estudio de las opciones incorrectas: a) Hay que reemplazar o eliminar los distractores que no son elegidos. b) No debiera preocuparnos que los distractores no sean elegidos por el mismo número de estudiantes, pues diferentes tipos de errores pueden ser cometidos por distinto número de estudiantes. c) Que la mayoría de los estudiantes falle un ítem no implica que deba ser cambiado, aunque los ítems en los que ocurre esto debieran analizarse detenidamente. d) Hay que sospechar de un ítem en el que un distractor es más elegido que todas las demás opciones juntas, en especial si la elección del distractor correlaciona positivamente con la puntuación en el test.

Los indicadores de discriminación vistos se pueden aplicar también a las opciones incorrectas. El índice de discriminación *D* aplicado a cada distractor nos diría si hay diferencia o no en la tasa de elección del distractor entre los subgrupos superior e inferior. Algo similar puede hacerse con los índices basados en la correlación ítem-test o ítem-resto del test.

Ejemplo 2.11. Correlación ítem-test en el estudio de las opciones incorrectas

Los autores generamos el siguiente ítem de Razonamiento:

Descubra el elemento que sigue en la serie 0, 1, 10, 11, 100, 101, ¿?
a) 102
b) 200
c) 110
d) 1000

Aplicado el ítem a una muestra de *N* evaluados, conocemos la opción que cada uno ha elegido y la puntuación en el test. Con estos resultados, podemos generar la Tabla 2.13. La segunda columna contiene la opción elegida por cada evaluado.

Tabla 2.13. Opción elegida por cada evaluado y puntuación en el test

Evaluado	*Opción elegida*	*a*	*b*	*c*	*d*	*X*
1	a	1	0	0	0	30
2	d	0	0	0	1	23
3	b	0	1	0	0	32
4	a	1	0	0	0	25
5	c	0	0	1	0	37
6	c	0	0	1	0	12
7	b	0	1	0	0	19
.	.	.	.	.	.	.
.	.	.	.	.	.	.
N	d	0	0	0	1	23

Las columnas *a*, *b*, *c* y *d* muestran un 1 y 3 ceros (1 en la columna que corresponde a la opción elegida). La proporción de evaluados que eligió cada opción fue: 0,01 (*a*), 0,04 (*b*), 0,29 (*c*) y 0,56 (*d*). La correlación de las columnas 3, 4, 5 y 6 de la tabla con la puntuación

en el test, *X*, fue –0,09 (*a*), –0,12 (*b*), 0,07 (*c*) y 0,13 (*d*). Nótese que las proporciones y las correlaciones serían los índices de dificultad y discriminación, respectivamente, si consideramos cada opción como la opción correcta. El estudio de estos valores da pistas sobre si la opción propuesta como correcta efectivamente lo es.

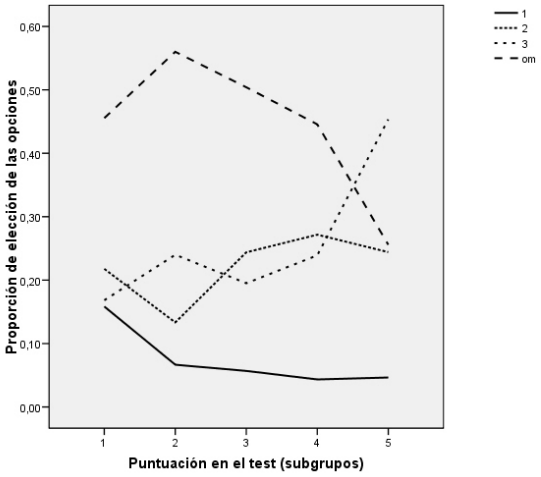
¿Qué nos dicen los anteriores resultados de la calidad del ítem? Los creadores del ítem propusimos como opción correcta la *d* (la serie sería: *0, 1, 10, 11, 100, 101, 1000, 1001,* etc.). La correlación ítem-test de esa opción es positiva (0,13), aunque baja. Lo sorprendente es que otra opción, en principio falsa, dio una correlación también positiva con la puntuación en el test. Volvimos a leer el ítem y nos dimos cuenta de que la serie, si se entiende expresada en código binario[11], debe continuar con la opción *c* (110). Por tanto, la opción *c* es también una opción correcta posible. El estudio de la discriminación de las opciones del ítem nos ha indicado que tiene de hecho dos soluciones. Otro resultado de interés es la escasa frecuencia de elección de los otros dos distractores. El ítem se aplicó en una oposición, donde los candidatos se juegan un puesto de trabajo y muchos años de preparación, y se penalizaban los errores. En estos contextos, los que no saben la respuesta prefieren dejar el ítem en blanco a responder al azar. La proporción de omisión en este ítem fue del 10%. Esta proporción pudo también haberse incrementado por los opositores que se dieran cuenta de las dos soluciones posibles. La opción 1 no ha sido elegida casi por nadie. En resumen, es un ítem manifiestamente mejorable o directamente descartable.

Análisis gráfico de ítems de opción múltiple

Una estrategia complementaria, más que alternativa, de hacer el análisis de ítems consiste en recurrir a gráficos. Livingston (2006) y Dowing y Haladyna (1997) recomiendan esta estrategia. Supongamos que tenemos un test formado por ítems de 3 opciones. Se puede fácilmente obtener la gráfica que se muestra a continuación (Figura 2.2). Lo primero que hacemos es dividir la muestra en varios subgrupos (por lo general, 5) con un número de evaluados similar. En el ejemplo que sigue el primer subgrupo está formado por las personas que tienen las peores puntuaciones en el test (menores de 12); el segundo subgrupo, por los que tienen las puntuaciones 13 ó 14; el tercero, por los que tienen puntuaciones entre 15 y 17; el cuarto por los que tienen puntuaciones entre 18 y 20; y el quinto por las mejores puntuaciones (superiores a 20). Se ha procurado que en cada subgrupo haya alrededor de un 20% de la muestra. En el eje de ordenadas se muestra la proporción de evaluados del correspondiente subgrupo que ha elegido cada una de las tres alternativas y la omisión. En la gráfica puede comprobarse que en el subgrupo con peor rendimiento en el test, alrededor de un 46% ha dejado el ítem sin responder, alrededor de un 22% ha elegido la opción 2, un 16% ha elegido la opción 1 y el restante 16% la opción 3. Similar información se ofrece para cada uno de los cinco subgrupos.

[11] Pues 0 en binario es, en decimal, 0; 1 es 1; 10 es 2; 11 es 3; 100 es 4; y 101 es 5. Por tanto, el término que sigue a 101 podría ser 110 (en decimal, 6), que aparece como opción c.

Figura 2.2. Elección de las opciones de un ítem en función de la puntuación en el test



Cada curva muestra cómo funciona la opción en los distintos subgrupos. En el caso de la opción correcta, cabe esperar que sea tanto más elegida cuanto mayor sea la puntuación en el test. Es decir, a la opción correcta deberá corresponder una curva creciente. La opción especificada como correcta en el ítem es la opción 3. En las opciones incorrectas o distractores debe ocurrir lo contrario: la proporción de personas que elige el distractor debe ser menor cuanto mayor es la puntuación en el test. Por lo tanto, cabe esperar curvas decrecientes. En la gráfica vemos que la curva es decreciente, aunque muy ligeramente, para el distractor 1 y creciente para el distractor 2. Parece, por tanto, que el distractor 2 no está funcionando bien y el 1 tampoco discrimina demasiado entre los que tienen puntuaciones altas y bajas en el test. Discrimina mejor la omisión. Otra información útil que nos da la gráfica es la proporción de elección de cada opción. Vemos que la opción 1 es muy poco elegida (sólo pasa, y ligeramente, del 10% en el subgrupo de los que menos puntuación han tenido en el test). Sin embargo, la proporción de omisión es la más alta en todos los subgrupos menos el último (que es sobrepasada por la opción correcta).

En el análisis cuantitativo, los indicadores psicométricos de este ítem se presentan en la Tabla 12.4. Se aprecia la alta proporción de omisiones y la baja tasa de elección del distractor 1. Los índices de discriminación de la opción correcta (marcada con un asterisco) son más bien bajos (sólo uno de los dos supera y por poco el valor 0,2). El distractor 2 da indicadores de discriminación positivos, aunque muy bajos, cuando los debiera dar negativos, como los da el distractor 1. En conjunto, puede decirse que es un ítem de calidad baja.

Tabla 2.14. Indicadores de las 3 opciones y de la omisión

	1	2	3*	*Omisión*
Proporción de elección (*p*)	0,075	0,226	0,252	0,447
Correlación ítem-test (r_{jX})	–0,127	0,032	0,224	–0,154
Índice de discriminación (*D*)	–0,091	0,067	0,180	–0,156

Ejemplos de análisis de ítems

Se muestran tres ejemplos. El primero corresponde a un examen con preguntas de opción múltiple; el segundo, a un test de rendimiento óptimo con preguntas abiertas; y el tercero, a un test de rendimiento típico con ítems de categorías ordenadas.

Ejemplo 2.12. Análisis de un examen de opción múltiple

Hemos aplicado un examen de 14 ítems de opción múltiple (3 opciones) sobre los contenidos de este tema a 87 estudiantes que cursaban la asignatura de Introducción a la Psicometría. Los estudiantes respondieron sabiendo que la calificación obtenida no tendría repercusión alguna en su nota final y con la instrucción de no dejar ítems sin responder.

El análisis psicométrico comienza con la creación del archivo de datos, que consta de tantas filas como evaluados y tantas columnas como ítems. Para la obtención de los resultados que siguen hemos utilizado los programas TAP (Brooks y Johanson, 2005) y SPSS.

Unos primeros datos de interés tienen que ver con la distribución de frecuencias de las puntuaciones en el test de los 87 estudiantes. El número medio de aciertos ha sido 9,149, el 65,4% de los 14 aciertos posibles. En proporción, 0,65 es también la media de los índices de dificultad p de los 14 ítems. Este valor incluye los aciertos que puedan haberse obtenido respondiendo al azar. En este test la consideración es relevante pues pedimos a los estudiantes que no dejasen respuestas sin contestar. Aplicando la fórmula [2.3], obtenemos la proporción media de acierto corregida, $p^c = 0{,}65 - (1 - 0{,}65)/2 = 0{,}48$ que queda muy cerca del valor 0,5 de referencia. Una primera conclusión del examen es que su nivel de dificultad medio es apropiado. Por tanto, los ítems facilitan que el test tenga variabilidad. La varianza de las puntuaciones en el examen resultó ser 4,15.

La Tabla 2.15 muestra para cada ítem el índice de dificultad (p) y cuatro indicadores de la discriminación: el índice de discriminación (D), la correlación biserial (r_b), la correlación biserial puntual (r_{bp}) y la correlación biserial puntual corregida (r^c_{bp}).

El primer ítem ha sido acertado por 74 de los 87 estudiantes. El índice de dificultad p es 0,85 (=74/87), el índice de discriminación D es 0,08. La correlación biserial es 0,31, la biserial puntual ítem-test es 0,20 y la correlación biserial puntual corregida, o ítem-resto del test es 0,03. La tabla anterior proporciona similar información de los restantes 13 ítems. No se han obtenido índices de discriminación negativos, excepto la correlación biserial puntual corregida del ítem 11. En todos los ítems la correlación biserial puntual está por encima de 0,20. En general, los ítems no plantean problemas de discriminación, aunque la correlación ítem-test corregida está muy cerca de cero en varios ítems.

Tabla 2.15. Resultados del análisis de 14 ítems de opción múltiple

Ítem	*Dificultad*	*Discriminación*			
	p	*D*	r_b	r_{bp}	r_{bp}^c
1	0,85	0,08	0,31	0,20	0,03
2	0,78	0,38	0,51	0,37	0,17
3	0,87	0,10	0,34	0,21	0,05
4	0,63	0,31	0,36	0,28	0,05
5	0,48	0,46	0,52	0,41	0,18
6	0,70	0,47	0,55	0,42	0,21
7	0,28	0,46	0,55	0,41	0,20
8	0,82	0,19	0,30	0,21	0,02
9	0,62	0,22	0,36	0,28	0,04
10	0,60	0,31	0,37	0,29	0,05
11	0,63	0,22	0,27	0,21	–0,03
12	0,83	0,25	0,45	0,30	0,12
13	0,61	0,45	0,54	0,43	0,21
14	0,45	0,56	0,62	0,49	0,27

Analicemos las tasas de elección de las opciones incorrectas del ítem 4, que era:

"La Comunidad Valenciana tiene
a) más de 3 millones de habitantes, b) cinco aeropuertos, c) menos de tres millones de habitantes."

¿Cuál es el principal fallo del ítem anterior?
1) Las opciones no están dispuestas verticalmente.
2) Da pistas sobre la respuesta correcta.
3) Evalúa sólo el recuerdo.

La Tabla 2.16 muestra la proporción de la muestra total que ha elegido cada una de las tres opciones (primera fila), la proporción que ha elegido cada opción del subgrupo superior (segunda fila) y del subgrupo inferior (tercera fila). La cuarta fila contiene la diferencia entre las proporciones que aparecen en las filas segunda y tercera (es decir, el índice de discriminación *D* de cada opción). Las dos últimas filas muestran las correlaciones ítem-test e ítem-resto del test si se toma cada opción como la opción correcta.

En la Tabla 2.16 comprobamos que la diferencia entre las proporción de acierto del grupo superior e inferior (0,31) coincide con el valor del índice de discriminación *D* para el ítem 4 en la Tabla 2.15. En el grupo completo la opción más elegida es la correcta (opción 2). De las dos opciones incorrectas, la opción 1 es elegida por un 31% de los estudiantes, mientras que la 3 lo es sólo por el 6%. La opción 1 está funcionando como un buen distractor, pues efectivamente en una de las recomendaciones expuestas en el apartado *Redacción de ítems de opción múltiple* se afirma que hay que disponer las opciones verticalmente, y ciertamente el ítem incumple esta recomendación. Los estudiantes de mayor conocimiento seguramente saben que el ítem incumple esa recomendación, pero se dan cuenta de que incumple otra más importante. Como está redactado el ítem, las opciones *a* y *c* son exhaustivas, pues la Comunidad Valenciana ha de tener más o menos de 3 millones de habitantes, por lo que la opción correcta no puede ser la opción 1. Por tanto, el

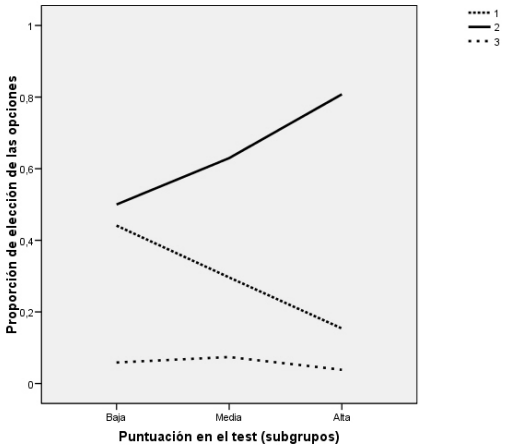
principal fallo del ítem es que da pistas sobre la opción correcta. El distractor 3 ha sido muy poco elegido. Habría que cambiarlo por otro. La presencia de la palabra *sólo* ayuda quizás a hacer poco plausible el distractor, pues es difícil que un ítem evalúe *sólo* algo. Se podría cambiar ese distractor por *La idea principal no está en el enunciado*, que se refiere a otra recomendación que el ítem incumple pero que es también menos importante que lo indicado por la opción 2.

Tabla 2.16. Indicadores de las 3 opciones

	1	2*	3
Completo	0,31	0,63	0,06
27% Superior (p_s)	0,15	0,81	0,04
27% Inferior (p_i)	0,44	0,50	0,06
Diferencia (p_s–p_i)	–0,29	0,31	–0,02
r_{bp}	–0,25	0,28	–0,09
r^c_{bp}	–0,04	0,05	–0,02

Al aplicar el indicador *D* a los dos distractores, vemos que al 1 corresponde un indicador negativo de –0,29, mostrando que ha sido elegido preferentemente por los estudiantes del subgrupo inferior. El valor de *D* en el otro distractor, el 3, está muy próximo a cero. En cualquier caso, sólo 5 personas de la muestra total han elegido esa opción. De esas 5, una pertenece al subgrupo superior y dos al inferior. Cuando la frecuencia total de elección del distractor es tan baja no es posible obtener diferencias de entidad entre los subgrupos.

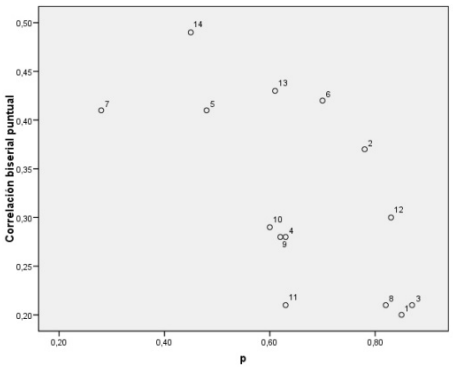
Figura 2.3. Elección de las opciones de un ítem en función de la puntuación en el test



La Figura 2.3 muestra la proporción de estudiantes que ha elegido cada alternativa dentro del subgrupo con puntuaciones bajas (33 % de peores calificaciones en el test), medias (33% de puntuaciones centrales) y altas (34% de puntuaciones mejores). Se han formado sólo tres subgrupos por tener la muestra sólo 87 estudiantes. Se aprecia el adecuado funcionamiento del distractor 1 y de la opción correcta 2. El distractor 3 apenas ha sido elegido en ninguno de los subgrupos.

En todos los items la correlación biserial puntual y la biserial están por encima de la referencia 0,20, lo que indica que no parece que ningún item requiera una revisión profunda. La Figura 2.4 muestra que de los cuatro ítems con menores valores de la biserial puntual, tres (ítems 1, 3 y 8) de ellos resultaron muy fáciles, con valores *p* superiores a 0,80. Como se ha comentado anteriormente, resulta complicado que ítems muy fáciles o muy difíciles sean a la vez discriminativos. Oosterhof (1976) encontró que cuanto más se aleja de 0,5 el índice de dificultad *p* del ítem, menores suelen ser los índices discriminación.

Figura 2.4. Relación entre los índices de dificultad y de discriminación



El análisis visto puede extenderse a un examen compuesto por *J* preguntas abiertas. En ese caso, obtendríamos la media como indicador de la dificultad. Prestaríamos atención a la varianza de cada pregunta. En principio, como ocurre en el caso de ítems dicotómicos, los ítems que tienen más varianza son los que más ayudan a que el test tenga varianza. Por tanto, una pregunta con varianza nula o casi nula, en la que la mayoría de los estudiantes hayan obtenido la misma puntuación, no parece en principio una buena pregunta, aunque también aquí cabe hacer la salvedad de que puede tener sentido mantener algunas preguntas muy fáciles si se introducen para constatar el dominio de conocimientos fundamentales. El indicador de la capacidad discriminativa de cada pregunta sería la correlación de Pearson entre las puntuaciones en cada ítem y la puntuación en el test. Cabe también obtener la correlación de Pearson entre las puntuaciones en la pregunta y en el resto del test, si son pocas las preguntas. Si tenemos una medida en un criterio externo que nos interese predecir, podríamos obtener el índice de validez de los ítems. En los dos ejemplos siguientes se obtienen e interpretan todos estos indicadores con datos reales.

Ejemplo 2.13. Análisis de ítems abiertos

En la parte práctica de un examen el estudiante ha de responder a 8 preguntas abiertas, puntuadas cada una entre 0 y 1. La nota en el examen práctico es la suma de las califica-

ciones en sus 8 ítems. La Tabla 2.17 muestra la media, la varianza y la correlación ítem-resto del test de cada ítem. El número de estudiantes del examen ha sido 68.

Tabla 2.17. Resultados del análisis de 8 ítems

Ítem	*Media*	*Varianza*	*Correlación ítem-resto del test*
1	0,79	0,10	0,06
2	0,37	0,11	0,26
3	0,12	0,09	–0,02
4	0,92	0,07	0,16
5	0,77	0,12	0,26
6	0,69	0,09	0,40
7	0,64	0,20	0,31
8	0,36	0,16	0,21

Se aprecia en la tabla que las preguntas han resultado muy diferentes en dificultad. La 4 ha resultado muy fácil (su media, 0,92, está muy cerca de la máxima puntuación posible, 1). La 3 ha resultado muy difícil (su media, 0,12, está cerca de cero). Los ítems 6 y 7 difieren poco en dificultad (sus medias son 0,69 y 0,64), pero más en varianza. El ítem 7, en principio, ayuda más que el ítem 6 a que la nota en el examen tenga variabilidad. La última columna muestra que cinco de los ocho ítems correlacionan más de 0,20 con el resto del test. En el ítem 4 la correlación está ligeramente por debajo de ese valor. En dos ítems (1 y 3) la correlación es muy próxima a cero y esos ítems no parecen relacionarse con el examen práctico en su totalidad.

Ejemplo 2.14. Análisis de ítems de categorías ordenadas

Los autores hemos elaborado una escala de 12 ítems para medir Estabilidad Emocional. Cada ítem es un adjetivo y el evaluado debe indicar cómo de bien le describe, seleccionando una de las 5 categorías disponibles (*Muy mal, Mal, Ni bien ni mal, Bien, Muy bien*). Los principales resultados del análisis de ítems (media, desviación típica y correlación ítem-test corregida) se muestran en la Tabla 2.18. Lo primero que llama la atención son los valores tan elevados de las medias[12]. Cada ítem se puntuó entre 1 y 5 (ítems directos) o entre 5 y 1 (ítems inversos). Por tanto, en los ítems directos, al obtenerse medias por encima de 4, prácticamente todos los evaluados consideran que ser feliz, ser una persona madura… una persona equilibrada les describe bien o muy bien. Igualmente, en los ítems inversos, por superar las medias el valor 4, consideran que ser irritable, malhumorada… y ser una persona con sentimientos de culpa les describe mal o muy mal. Las desviaciones típicas son pequeñas, como cabe esperar cuando las medias son tan altas. Las correlaciones de cada ítem con el resto del test son todas positivas, estadísticamente significativas

[12] Los datos se han obtenido en un proceso selectivo y muy probablemente las respuestas han sido parcialmente falseadas (deseabilidad social) para acomodarse al perfil psicológico que demanda el puesto.

distintas de cero, y mayores de la referencia 0,2. Por tanto, todos los ítems tienen una adecuada discriminación y están contribuyendo a medir lo que se pretende medir con el test. No parece, por tanto, que haya que reconsiderar o anular ninguno de los 12 ítems.

Tabla 2.18. Resultados del análisis de 12 ítems de categorías ordenadas

Soy una persona…	*Media*	*Desviación típica*	*Correlación ítem-test corregida*
Feliz	4,39	0,583	0,423
Estable	4,43	0,559	0,586
Madura	4,28	0,537	0,521
Optimista	4,32	0,577	0,482
Equilibrada	4,43	0,576	0,571
Coherente	4,26	0,578	0,486
Irritable	4,33	0,614	0,542
Malhumorada	4,34	0,568	0,594
Miedosa	4,13	0,564	0,438
Envidiosa	4,31	0,611	0,491
Desanimada	4,35	0,596	0,574
Con sentimientos de culpa	4,25	0,794	0,381

Corrección de los efectos del azar

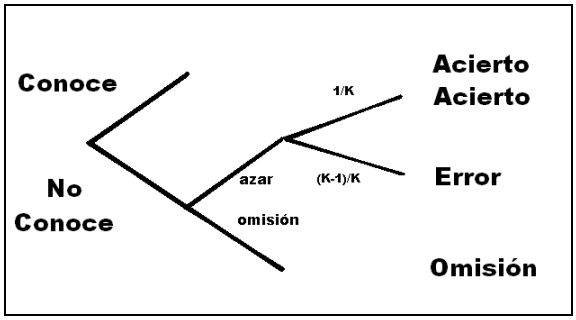
En los tests formados por ítems de opción múltiple podemos sobrestimar el nivel de rendimiento de algunas personas, dado que algunos de sus aciertos han podido producirse por haber respondido al azar, y no por saber la opción correcta. El problema entonces consiste en establecer un procedimiento para descontar del número total de aciertos (X) los que supuestamente se han producido por haber respondido al azar (X_a).

Supongamos que dos estudiantes saben lo mismo y responden al mismo test, que consiste en 100 preguntas con formato verdadero-falso. Los dos saben 60 preguntas. El primero responde las 60 preguntas que sabe y omite las 40 que no sabe. Su puntuación en el test, su número de aciertos, será 60. El segundo estudiante responde las 60 preguntas que sabe y decide responder estrictamente al azar las otras 40. Como cada una tiene dos opciones, supongamos que acierta 20 de las 40. Pues bien, mientras que el primer estudiante tiene 60 aciertos (las preguntas que sabe), el segundo tiene 80 (las 60 que sabe y las 20 que ha acertado por haber respondido al azar). En este apartado vamos a proponer un procedimiento que pretende eliminar del número total de aciertos los que presumiblemente se deben al azar.

Ante un ítem, supondremos que la persona se encuentra en uno de dos estados de conocimiento: en el estado *conoce* la respuesta o en el estado *no conoce* la respuesta. En el primer estado supondremos que conoce la respuesta y acierta con probabilidad 1. Si se encuentra en el segundo estado, tiene dos posibilidades: no responder o elegir al azar una de las K opciones. Dado que el ítem tiene una sola opción correcta y que suponemos que todas las opciones son equiprobables cuando se responde al azar, la probabilidad de acierto

será *1/K* y la de fallo será *1 – (1/K) = (K – 1)/K*. La Figura 2.5 muestra las diferentes posibilidades.

Figura 2.5. Estados de conocimiento y resultados en el ítem



Llamemos R_a al número de respuestas al azar que la persona da (es decir, al número de ítems que ha contestado sin saber la respuesta). De las R_a respuestas, algunas serán aciertos aleatorios (X_a) y otras serán errores (E). Nuestro objetivo es obtener el valor de X_a para descontarlo del número total de aciertos (X) que ha obtenido.

Respondiendo al azar, la probabilidad de fallar un ítem vimos que es *(K – 1)/K*. Si se responde al azar a R_a ítems, el número esperado de errores (E) será:

$$E = R_a \frac{K-1}{K} \qquad [2.9]$$

Si despejamos R_a de esta expresión, se obtiene:

$$R_a = \frac{K}{K-1} E \qquad [2.10]$$

Siguiendo el mismo razonamiento, el número esperado de aciertos aleatorios cuando se se dan R_a respuestas al azar será:

$$X_a = R_a \frac{1}{K} \qquad [2.11]$$

Si realizamos las sustituciones oportunas, se obtiene:

$$X_a = \frac{K}{K-1} E \frac{1}{K} = \frac{1}{K-1} E \qquad [2.12]$$

La expresión anterior permite obtener X_a, a partir de los errores cometidos (E) y del número de alternativas que tienen los ítems (K). Podemos observar que cada error se pondera por la expresión *1/(K – 1)*, lo que significa que por cada error hay que descontar tantos aciertos como indica ese cociente: en tests de 2 alternativas de respuesta, hay que descon-

tar 1 acierto por cada error; en ítems de 3 alternativas, 0,5 aciertos por cada error; en ítems de 4 alternativas, 0,33 aciertos por cada error; y así sucesivamente.

La puntuación corregida de una persona en el test será:

$$X^c = X - X_a = X - \frac{E}{K-1} \qquad [2.13]$$

Si aplicásemos esta fórmula al ejemplo que planteamos al comienzo, tendríamos que, para el primer estudiante,

$$X^c = X - \frac{E}{K-1} = 60 - \frac{0}{2-1} = 60$$

Para el segundo,

$$X^c = X - \frac{E}{K-1} = 80 - \frac{20}{2-1} = 60$$

La fórmula correctora deja a ambos estudiantes, que sabían lo mismo, con la misma puntuación (60), que son por cierto los ítems que sabían.

La fórmula anterior se aplica cuando todos los ítems tienen igual número de opciones. Si el número varía, un error en un ítem *j* de K_j opciones quitaría $1/(K_j - 1)$ aciertos (Frary, 1988). Por tanto, X_a sería la suma de los valores $1/(K_j - 1)$ de los ítems en los que se obtuvo un error.

Ejemplo 2.15. Obtención de las puntuaciones corregidas

Un test de conocimientos del nivel de inglés está formado por 140 ítems con 5 opciones de respuesta cada uno. En la Tabla 2.19 se detallan el número de aciertos (*X*), errores (*E*) y omisiones (*O*) que obtuvieron 3 evaluados. La última columna contiene sus puntuaciones corregidas. Si atendemos únicamente al número de aciertos obtenidos, quien más inglés parece saber es el evaluado 1, seguido del 2 y en último lugar el 3. Sin embargo, tras corregir los efectos del azar, comprobamos que la corrección afecta al orden que establecimos a partir de las puntuaciones sin corregir. Similarmente, si nos fijamos en la corrección hecha para el evaluado 3, vemos que no se le ha descontado nada, pues no cometió ningún error.

Tabla 2.19. Aciertos, errores, omisiones y puntuaciones corregidas

Evaluado	*X*	*E*	*O*	X^c
1	112	28	0	112 – 28/4 = 105
2	110	12	18	110 – 12/4 = 107
3	109	0	31	109 – 0/4 = 109

Haciendo así las cosas se está asumiendo que sólo se puede obtener un error cuando se responde al azar. El modelo no contempla la posibilidad de error por descuido o por haber aprendido algo mal, sino exclusivamente como resultado de una respuesta completamente al azar entre las *K* opciones. Por tanto, si hay errores es que ha habido respuestas al azar. Según la fórmula [2.13], a partir del número observado de errores puede obtenerse el número de aciertos que han debido producirse por azar y ese valor se resta del total de aciertos. Se pueden plantear otros modelos alternativos al expuesto en la Figura 2.5, de cómo los evaluados responden los ítems de opción múltiple. De hecho, no es infrecuente que los alumnos salgan de un examen diciendo que no han dado una sola respuesta al azar y sin embargo obtienen errores. Esto ocurre porque consideran correctas opciones que no lo son. En cualquier caso, lo que es evidente es que inferir el número de aciertos debidos al azar a partir de estos errores es incorrecto, pues no se han generado por haber respondido al azar. En el apéndice de este tema se describe otra fórmula para la corrección de los efectos del azar.

¿Hay que aplicar o no las fórmulas correctoras? No hay duda de que se ha de avisar al evaluado de si se va a aplicar o no alguna fórmula y de sus detalles, en su caso. No hay tanto acuerdo en relación a si es adecuado aplicarlas o no. Conviene tener en cuenta las consideraciones que se exponen a continuación.

Lo que hace la fórmula correctora es eliminar los aciertos que se obtienen al responder completamente al azar. En ese sentido, quien responde sólo a lo que sabe y quien responde a lo que sabe y a lo que no (y a estas preguntas completamente al azar) deberá esperar, tras la aplicación de la fórmula correctora, la misma puntuación. Por lo tanto, si se aplica la fórmula descrita, debiera no importar dar respuestas al azar, pues se espera obtener el mismo número de aciertos. Supongamos que estamos ante un ítem de cinco opciones. Si alguien responde completamente al azar, la probabilidad de acierto es 0,20 y de fallo 0,80. Al aplicar la fórmula, por cada error el número de aciertos queda reducido en ¼ = 0,25. Si esto lo hace en los 20 ítems de un examen, su número esperado de aciertos y de errores es 20(0,20) = 4 y 20(0,8) = 16, respectivamente. Al aplicar la fórmula correctora [2.13], tendríamos que $X^c = 4 - 16/4 = 0$. Supongamos que alguien sabe que una de las opciones no es correcta. En ese caso, si responde completamente al azar entre las demás, la probabilidad de acierto es 0,25 y la de fallo es 0,75. Si, por ejemplo, en 20 preguntas responde al azar entre cuatro opciones, pues tiene la seguridad de que una de las opciones no es correcta, el número esperado de aciertos por azar en esas 20 preguntas será (20)(0,25) = 5 y el de errores (20)(0,75)= 15. Sin embargo, al aplicarle la fórmula correctora, el número esperado de aciertos que se le quitaran serán (15)(0,25) = 3,75. Es decir, se le quitarían menos aciertos (3,75) de los que esperaría (5). Supongamos que puede descartar dos opciones en cada ítem. En ese caso, si responde completamente al azar entre las demás, la probabilidad de acierto es 1/3 y la de fallo es 2/3. Si, por ejemplo, en 20 preguntas responde al azar entre las tres opciones, pues tiene la seguridad de que dos de las opciones no son correctas, el número esperado de aciertos por azar en esas 20 preguntas será (20)(1/3) = 6,7, y el de errores será (20)(2/3)=13,3. Al aplicarle la formula, el numero de aciertos que se le quitarían es (13,3)(0,25) = 3,32, que es inferior al número esperado de aciertos (6,7). Vuelve a resultar interesante responder al azar entre las tres opciones.

En conclusión, si no se puede descartar ninguna opción, la fórmula te va quitar, en promedio, lo que ganes por haber respondido al azar. Si se tiene seguridad de que alguna opción es incorrecta, el número de aciertos esperado es mayor que el número de aciertos que la fórmula resta si se responde al azar entre las opciones no descartadas. Este resulta-

do es importante, pues muestra que la aplicación de la fórmula correctora NO elimina todos los aciertos que puedan haberse producido por responder al azar. Elimina todos los aciertos cuando se responde al azar *entre todas las opciones*, pero no cuando se elimina alguna porque se conoce que es falsa.

Entre los especialistas, no existe acuerdo sobre el tipo de instrucciones a dar, por ejemplo, en un examen con preguntas de opción múltiple. Cuando un estudiante no sabe lo suficiente para aprobar, la mejor estrategia que puede seguir es responder al azar las preguntas que no sabe, por si pudiera, por puro azar, obtener el número de aciertos requerido para aprobar. La recomendación general de "no responder al azar" no es la adecuada para estos estudiantes y cabe plantearse si puede darse como instrucción general cuando no es apropiada en algunas situaciones (Frary, 2008). Este autor concluye que: "…*es difícil recomendar una fórmula correctora de los aciertos obtenidos por las respuestas dadas al azar en los exámenes de opción múltiple habituales en la universidad… Lo más justo es recomendar a todos los estudiantes que lo mejor para ellos es contestar a todas las preguntas sea cual sea su nivel de conocimientos*".

Otros autores (Burton y Miller, 1996; Burton 2001, 2004) están a favor de aplicar las fórmulas correctoras, porque son eficaces en la reducción de las respuestas al azar, indicando a los evaluados la reducción que se va a aplicar por cada error. Burton (2001) propone que se apliquen la Fórmula [2.13] a pesar de que no corrija adecuadamente los aciertos atribuibles a las respuestas al azar, precisamente porque reduce o elimina dichas respuestas y porque considera deshonesto instruir a los evaluados para que respondan a lo que no saben. Otra ventaja de intentar evitar las respuestas al azar es que, desde un punto de vista instruccional, las respuestas erróneas son informativas de lo que un estudiante no ha llegado a aprender. Cuando fomentamos las respuestas al azar, se pierde esta valiosa información (Burton, 2004).

Apéndice

Segunda fórmula correctora

Traub, Hambleton y Singh (1969) propusieron una segunda fórmula que premia las omisiones en vez de penalizar los errores. En un test de *J* ítems de opción múltiple de *K* opciones, con sólo una opción correcta, una persona obtiene *X* aciertos, *O* omisiones y *E* errores. En un ítem, si en vez de omitir se hubiese respondido al azar, la probabilidad de acierto sería *1/K*. De haber hecho esto mismo en los O ítems omitidos, el valor esperado de aciertos en los *O* ítems sería *O(1/K)=O/K*. Se propone como segunda fórmula la siguiente:

$$X_2^c = X + \frac{O}{K} \qquad [2.16]$$

La segunda fórmula añade los aciertos que cabe esperar obtener si se responde completamente al azar a los ítems de los que no sabe la respuesta correcta, en vez de quitar los aciertos que se suponen obtenidos por haber respondido al azar. Es evidente que las puntuaciones corregidas obtenidas por la segunda fórmula por lo general serán más altas que las obtenidas por la primera.

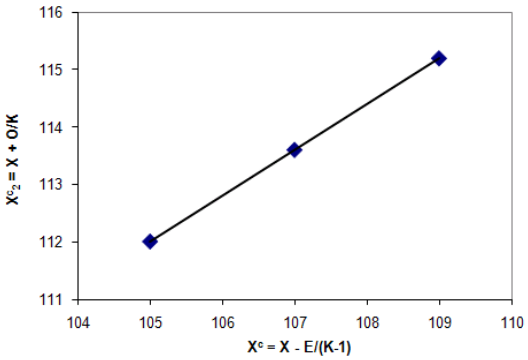
La Tabla 2.20 se ha construido a partir de los datos de la Tabla 2.19, que incluye el patrón de resultados de 3 evaluados en un test de 140 ítems de 5 opciones de respuesta.

Tabla 2.20. Aciertos, errores, omisiones y puntuaciones corregidas

Evaluado	*X*	*E*	*O*	X^c	X_2^c
1	112	28	0	105	112 + 0/5 = 112
2	110	12	18	107	110 +18/5 = 113,6
3	109	0	31	109	109 + 31/5 = 115,2

Se aprecia que las tres personas están ordenadas de la misma manera en ambas fórmulas correctoras. Es más, si representamos gráficamente las puntuaciones de las tres personas según las dos correcciones, vemos (Figura 2.6) que están en la misma recta:

Figura 2.6. Relación lineal entre las dos fórmulas correctoras



La relación observada se cumple siempre, pues existe una relación lineal entre los valores que se obtienen con ambas fórmulas:

$$\begin{aligned} X_2^c &= X + \frac{O}{K} = X + \frac{J - X - E}{K} = X - \frac{X}{K} - \frac{E}{K} + \frac{J}{K} = X\left(\frac{K-1}{K}\right) - \frac{E}{K} + \frac{J}{K} \\ &= \left(X\left(\frac{K-1}{K}\right)\left(\frac{K}{K-1}\right) - \frac{E}{K}\left(\frac{K}{K-1}\right) + \frac{J}{K}\left(\frac{K}{K-1}\right)\right)\frac{K-1}{K} \\ &= \left(X - \frac{E}{K-1} + \frac{J}{K}\left(\frac{K}{K-1}\right)\right)\frac{K-1}{K} \\ &= \frac{K-1}{K}X^c + \frac{J}{K} \end{aligned}$$

Ambas fórmulas son equivalentes, pues ordenan a las personas exactamente de la misma manera. Las puntuaciones obtenidas son, sin embargo, muy diferentes. La fórmula que premia las omisiones otorga puntuaciones más altas que la que penaliza los errores. Por tanto, a la hora de interpretar las puntuaciones habrá que tener esto en cuenta. No parece que pueda ponerse, por ejemplo, el mismo punto de corte de aprobado si se aplica una u otra. Algunos autores plantean que, desde un punto de vista ético, parece mejor estrategia premiar la omisión cuando no se sabe, que castigar lo que se quiere evitar (Frary, 2008).

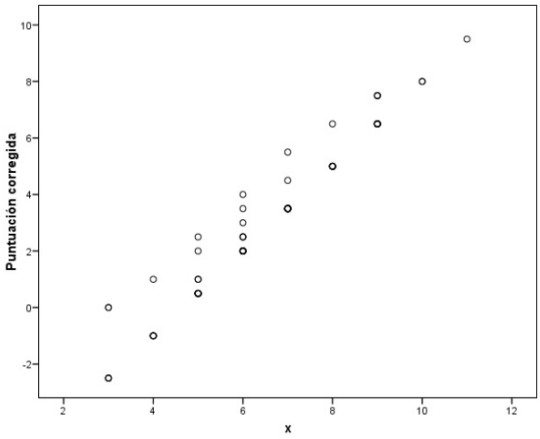
Hemos visto que esas posiciones relativas de los evaluados son las mismas cuando se aplica una u otra fórmula. La relación de X^c y X_2^c con la puntuación sin corregir, *X*, requiere algún comentario adicional.

Es fácil ver que existe una relación lineal perfecta entre X^c y *X* cuando no hay omisiones.

$$X^c = X - \frac{E}{K-1} = X - \frac{J - X}{K-1} = X + \frac{X}{K-1} - \frac{J}{K-1} = X\left(\frac{K}{K-1}\right) - \frac{J}{K-1}$$

Como existe una relación lineal entre X^c y X_2^c, se sigue que también la hay entre X_2^c y *X*. Por lo tanto, la posición relativa de las personas es la misma cuando son puntuadas con las dos fórmulas correctoras y la misma que se obtendría tomando la puntuación sin corregir, en el caso de que no hubiera omisiones. La relación entre la puntuación sin corregir y la corregida en el caso general, cuando hay omisiones, es de un fuerte componente lineal, pero no cabe esperar una relación lineal perfecta. La Figura 2.7 muestra los resultados de 80 estudiantes en un test de 14 ítems. Se obtuvo el siguiente diagrama de dispersión entre las puntuaciones corregidas X^c y las puntuaciones sin corregir *X*. La correlación entre ambas es de 0,971.

Figura 2.7. Relación entre el total de aciertos, X, y la puntuación corregida, X^c



Programas de ordenador para el análisis de ítems

Son muchos los programas disponibles para la realización del análisis clásico de ítems. Algunos son además de libre distribución. A continuación se detallan las principales características de algunos y cómo pueden conseguirse.

TAP (Brooks y Johanson, 2005) es un programa de libre distribución. Proporciona para cada ítem los índices de dificultad (p), discriminación (D), correlación biserial (r_b), correlación biserial puntual (r_{bp}) y correlación biserial puntual corregida. Permite el estudio del funcionamiento de las opciones incorrectas, pues proporciona para cada opción la frecuencia de elección y el índice de discriminación (D). No proporciona para los distractores las correlaciones ítem-test o ítem-resto del test. El tamaño del grupo superior e inferior lo fija por defecto en el 27% de la muestra, pero el usuario puede modificar el porcentaje. El programa puede obtenerse en la dirección: http://oak.cats.ohiou.edu/~brooksg/software.htm#TAP.

El programa CIA (http://shkim.myweb.uga.edu/), de libre distribución, obtiene para cada opción del ítem las correlaciones biserial y biserial puntual, con el test y con el resto del test. Divide la muestra en cinco subgrupos de igual tamaño (20%) y obtiene en cada uno cuantos evaluados han elegido cada una de las opciones. No permite cambiar el número de subgrupos.

Ledesma, Molina, Valero y Young (2002) han desarrollado un módulo, de libre distribución, que proporciona lo siguientes datos: 1) Los estadísticos descriptivos para los ítems y el test, 2) los efectos de la eliminación de cada ítem en los estadísticos descriptivos del test, y 3) las correlaciones entre ítems, ítem-total e ítem-resto del test. El programa da los resultados no sólo mediante tablas, sino también mediante gráficas.

López Pina (2005) proporciona otro programa de libre distribución para el análisis clásico de ítems, denominado CLM-1, válido para ítems de respuesta seleccionada. Obtiene los índices de dificultad y de discriminación estudiados y el índice de fiabilidad de cada ítem. Proporciona además datos psicométricos del test completo.

ITEMAN es un programa específico de análisis clásico de ítems de opción múltiple y de categorías ordenadas. Proporciona para cada ítem el índice de dificultad, el índice de discriminación y las correlaciones biserial y biserial puntual sin corregir y corregidas[13]. Más información en Lukas

[13] La versión 3.6 del programa da correlaciones biseriales puntuales corregidas incorrectas para los distractores.

(1998) y en http://assess.com/. En esta misma dirección se puede encontrar otro programa LERTAP 5. Es una herramienta muy completa para el análisis clásico de ítems y tests. En lo relativo específicamente al análisis de ítems, proporciona los indicadores de dificultad y discriminación, tanto los basados en la diferencia entre grupos, como en la correlación con el test o resto del test. Permite la inclusión de un criterio externo al test. Proporciona información gráfica del rendimiento del ítem para los distintos subgrupos.

El paquete SPSS no tiene específicamente un programa para el análisis de ítems, pero el procedimiento *Análisis de fiabilidad* puede resultar útil. Proporciona para cada ítem su media y varianza, la correlación entre el ítem y el resto del test, y el valor de la media, varianza y fiabilidad del test si se elimina cada ítem. Este procedimiento puede aplicarse a ítems de respuesta seleccionada y construida, así como a ítems de categorías ordenadas. Lei y Wu (2007) han desarrollado programas para SPSS y SAS que completan el análisis clásico de ítems dicotómicos y politómicos de ambos paquetes.

El grupo de investigación TIDE, de la universidad de Barcelona, ha desarrollado varios programas relacionados con el análisis de ítems y tests. METRIX Engine obtiene para cada ítem sus estadísticos descriptivos y los índices de dificultad y discriminación en el caso de ítems de opción múltiple. La aplicación SEDI (Renom, Rodríguez, Solanas, Doval, Nuñez y Valle, 2001) acepta la salida del módulo de análisis de ítems de METRIX, evalúa la calidad de cada ítem y recomienda qué hacer con cada uno de ellos. Más información en http://www.ub.es/comporta/tide/Index.htm.