CUADERNOS DE DOCENCIA UNIVERSITARIA 39

BUENAS PRÁCTICAS EN EL USO DE PRUEBAS DE ALTERNATIVA MÚLTIPLE

Jordi Renom Pinsach

Eduardo Doval Diéguez

Octaedro Editorial

Institut de Desenvolupament Professional

Cuadernos de docencia universitaria 39

Título: *Buenas prácticas en el uso de pruebas de alternativa múltiple*

Primera edición: octubre de 2019
Recepción del original: 11/03/19
Aceptación: 09/05/19

Ediciones OCTAEDRO
Bailèn, 5, pral. - 08010 Barcelona
Tel.: 93 246 40 02
www.octaedro.com - octaedro@octaedro.com

Universitat de Barcelona
Institut de Ciències de l'Educació
Campus Mundet - 08035 Barcelona
Tel.: 93 403 51 75

ISBN: 978-84-17667-82-5

Diseño y producción: Servicios Gráficos Octaedro

ÍNDICE

INTRODUCCIÓN

Desde hace décadas los test «objetivos» y exámenes «tipo test» han sido objeto de un debate controvertido. Parecía que los nuevos modelos educativos y las crecientes tecnologías harían que estas maneras de evaluar quedasen obsoletas frente a otras formas de evaluación más auténticas. Sin embargo, hoy en día siguen estando muy presentes en la evaluación universitaria, en procesos de acreditación y en la formación en general, siendo muy habituales en las plataformas formativas digitales.

Las pruebas tipo test o de elección múltiple tienen, como cualquier otro tipo de prueba, ventajas e inconvenientes, aunque, en este caso, los usuarios tienden a ponderar solo las primeras. Si atendemos al volumen de evaluaciones en que se aplican y a la repercusión de sus resultados para los examinados, sorprende la falta de normativa sobre la calidad de este tipo de exámenes. Existen regulaciones para evitar el fraude por parte de las personas que realizan la prueba y también directrices internacionales de buenas prácticas sobre la creación y administración de los exámenes, aunque estas últimas son poco conocidas y, en cualquier caso, al tratarse de recomendaciones, su uso queda a criterio de las personas que vayan a usar la prueba. Tampoco es muy habitual que los docentes sometan a control sus pruebas, efectuando o solicitando un análisis de su calidad. De hecho, de acuerdo con el tópico predominante de que los exámenes tipo test son fáciles de crear y que de por sí aportan un valor de objetividad a la evaluación, este control puede parecer innecesario. La experiencia acumulada, sin embargo, muestra que el principal valor de este tipo de pruebas se reduce a la autonomía (que no sencillez) en la creación y, por encima de todo, a la facilidad de corrección, pero la supuesta objetividad de sus resultados no queda, ni mucho menos, garantizada.

Como en cualquier otro procedimiento de evaluación, los test objetivos comportan unas reglas de juego que hay que conocer, porque, de lo contrario, el usuario asume una responsabilidad que parece que no le afecta directamente, pero siempre acaba perjudicando a sus examinados.

Manteniendo un enfoque crítico, esta obra va dirigida a las personas encargadas de realizar evaluaciones, porque son ellas quienes deben tomar conciencia de las prestaciones y limitaciones de este tipo de instrumentos. No se trata de un manual especializado ni exhaustivo sobre medición educativa; en la bibliografía se encuentran algunas de las principales referencias en este campo que puede ayudar a profundizar en aspectos concretos. Tampoco se pretende debatir las ventajas e inconvenientes ni el estado de la cuestión de esta modalidad de evaluación. A partir de la experiencia de los autores, se propone una metodología de base para docentes de diferente perfil interesados en esta modalidad de evaluación que les permita abordar de manera autónoma y con garantías el *test del test* de un examen.

1. OBJETIVO DEL TEST

En su sentido más convencional, un test es un instrumento compuesto por una serie de ítems (preguntas, tareas, problemas...) diseñados para evaluar conocimientos o aptitudes en el ámbito psicológico y educativo. En el caso de los test «cerrados» de alternativa múltiple (AM), la persona que se examina escoge la respuesta correcta entre varias opciones que cada pregunta ofrece. Los ítems se aciertan, fallan u omiten y, por lo general, la suma de aciertos (puntuación total) proporciona una supuesta medida del nivel del examinado.

La finalidad de este tipo de exámenes suele estar ligada a la evaluación de aprendizajes, básicamente conocimientos y competencias en toda clase de contenidos y ámbitos. También son habituales en procesos de selección y acreditación (oposiciones, idiomas, conducción...).

En términos pedagógicos, los exámenes se han relacionado con la evaluación sumativa y menos con la evaluación formativa. Esto, en parte, se debe a la forma en que se han empleado dentro de la estrategia de evaluación.

En cuanto a sus repercusiones, los exámenes varían en función de sus efectos. Hay pruebas de simple control y seguimiento, como, por ejemplo, la modalidad denominada *quiz* (exámenes) frecuentes en muchas plataformas virtuales de aprendizaje o LMS (*learning managment systems*), que sirven para que los estudiantes realicen ejercicios de prueba. En el otro extremo se encuentran los exámenes con efectos vinculantes para el futuro del examinado, donde aprobar o suspender la prueba tiene consecuencias claras (por ejemplo, repetir la asignatura, tener que volver a pagar la matrícula, superar o no una oposición...). Los exámenes también pueden diferir por la precisión de sus resultados. Los hay tipo *screening* o exploratorios, que indican, aunque con poca precisión, el nivel que tiene cada persona evaluada. En cambio, otros (por ejemplo, los de certificación) ofrecen puntuaciones más precisas con el fin de minimizar el número de errores en las decisiones del tipo apto/no apto, tomadas en relación con una nota de corte muy concreta.

Desde un punto de vista semántico, la principal diferencia entre test y examen está en el carácter estandarizado del primero. Para conseguir una prueba estandarizada son necesarios estudios previos, realizados con una o más muestras piloto de personas, con el fin de evaluar la calidad de los ítems y del conjunto de la prueba y concretar una versión definitiva de la prueba con garantías psicométricas. Por lo general, en el campo de la evaluación, este tipo de análisis y comprobaciones previas solo es posible en evaluaciones realizadas a gran escala, como en el estudio PISA. Por el contrario, en un examen no suele haber ensayos ni análisis previos de sus cualidades métricas con muestras piloto de estudiantes. Por lo general, un examen se crea con un fin concreto, para una ocasión determinada, y sus resultados se interpretan a partir de un criterio predefinido por las personas que realizan la evaluación. Un examen (el mismo examen) no se diseña para diferentes ocasiones y grupos. Al revés, su originalidad y confidencialidad son a menudo los valores que garantizan la equidad de la evaluación.

Pese a estas diferencias, test y exámenes comparten algunas cualidades y elementos metodológicos importantes, aunque en los exámenes pocas veces queden garantizados. Por ejemplo, tras aplicar un examen no es habitual analizar sus cualidades psicométricas. Por regla general, el examen simplemente se crea, se aplica y se evalúa con él a las personas que se presentan a la evaluación, asumiendo que la prueba tiene garantías (no comprobadas) para ello. No es que dichas garantías no se puedan evaluar. Desgraciadamente, en muchas ocasiones simplemente no se realizan las comprobaciones necesarias sobre las cualidades exigibles a la prueba, porque las personas responsables del examen desconocen su existencia o su importancia.

1.1. Otras cuestiones clave

Una vez definidos los objetivos evaluativos de la prueba, es necesario concretar la forma y estructura que esta va a tomar; lo cual guarda relación con cuatro preguntas clave: qué formato de ítem vamos a emplear, cuántas preguntas debe tener la prueba, qué taxonomías de conocimiento se van a utilizar y cómo se va a corregir y puntuar.

¿Qué formato de ítem vamos a emplear? Se trata del modo en que se responden los ítems. Hay exámenes de respuesta abierta en que el examinado elabora la respuesta, y cerrados del tipo Verdadero/Falso (VF), o AM, con diversas variantes. En esta obra el formato al que se hará referencia será el de AM convencional, con una sola respuesta correcta que puntúa. Muchos de los aspectos tratados serán también válidos para el formato VF y para otras variantes de AM.

¿Cuántas preguntas deben formar la prueba? A este respecto no hay una pauta estricta. Según el formato, un examen puede tener diferente número L (longitud) de ítems. En el caso de AM, también hay que decidir el número *n* de alternativas de respuesta (es recomendable que todos los ítems mantengan ese valor). Existe una relación matemática, descrita en la teoría clásica de los test (TCT, una disciplina psicométrica), entre *L* y *n*, y el coeficiente de fiabilidad de las puntuaciones del test. En general, la fiabilidad aumenta cuanto mayor es *L* y también *n*, aumentando, por tanto, también la precisión de la medida (dado que el error de medida de las puntaciones se reduce). Por este motivo muchas pruebas oficiales son largas, a veces de forma excesiva (es un recurso técnico para conseguir mayor precisión en los resultados y menor riesgo de reclamaciones, impugnaciones, etc.). En la práctica es posible estimar el punto en que se optimiza la relación entre el esfuerzo que supone crear y gestionar muchas preguntas en aras de una mayor fiabilidad/precisión.

En cuanto al valor idóneo de *n*, desde hace décadas ha sido objeto de estudio y discusión. La tendencia histórica ha ido a la baja, hoy en día se ha estabilizado entre cuatro y seis alternativas desde un punto de vista técnico (para maximizar la fiabilidad), o tres desde un punto de vista práctico (para conseguir buenas opciones de respuesta). Paradójicamente en muchas ocasiones el número de alternativas de la prueba se decide de manera puramente circunstancial, por ejemplo, en función del tipo de hoja de respuesta disponible, de las prestaciones de la lectora óptica o por costumbre («en esta asignatura el examen siempre ha sido así»).

A la hora de decidir *L* y *n*, también hay que considerar la estrategia con que se abordará el efecto de la conjetura de la respuesta (apartados 7.2,

7.3 y anexo) en caso de que se sospeche que las condiciones del examen la favorecen. Aquí la recomendación es simple: aumentar *n,* teniendo presente que a menor *n* se precisará mayor *L*, y viceversa. Un ejemplo habitual son los exámenes de VF que suelen tener mayor longitud que los de AM, porque con solo dos alternativas se necesitan más ítems para contrarrestar el posible efecto de la conjetura.

¿Qué taxonomías de conocimiento se van a utilizar? La pregunta no hace referencia al contenido o materia que evalúan los ítems, sino al enfoque de su planteamiento. Muchos exámenes AM identifican el dominio de una asignatura con la capacidad de recordar o evocar datos, conceptos, principios, etc. Sin embargo, esta limitación no es achacable a las preguntas AM, sino a la manera en que están planteadas. Las preguntas se pueden diseñar con enfoques diferentes a los puramente memorísticos (comprender, analizar, sintetizar...) como se verá en el apartado 2. Lo recomendable es que la taxonomía adoptada en el examen sea equivalente al enfoque empleado en el proceso de enseñanza-aprendizaje. La combinación entre la taxonomía y el contenido evaluado (los ítems también guardar relación con los diferentes aspectos tratados en el proceso de enseñanza-aprendizaje) determina la representatividad o validez de contenido del examen.

¿Cómo se va a corregir y puntuar? La mayoría de los ítems AM se puntúan como 1 (acierto) y 0 (error), y se corrigen manual o mecánicamente, a partir de una plantilla o clave. Solo se considera acierto cuando el examinado escoge la alternativa correcta. Sin embargo, existen variantes como el caso en que cada alternativa de respuesta se asocie con un tipo de conocimiento incorrecto, parcial o total. Por ejemplo, para un ítem con cinco opciones de respuesta, una totalmente incorrecta, tres parcialmente correctas en distinto grado y otra totalmente correcta, se le podría asignar un peso de 0, de 0,25, de 0,50, de 0,75 y de 1, respectivamente. Otra opción es que la máxima puntuación (acierto) corresponda a la elección de una combinación de alternativas (esto no es muy recomendable). Este último caso plantea dos cuestiones que cabe considerar: en primer lugar, cómo se tratará el efecto de la conjetura; en segundo lugar, cuál será la viabilidad de una corrección mecanizada, ya que muchas lectoras tienden a aceptar solo una alternativa como cierta.

1.2. Retos y cualidades generales de un examen

Las pruebas de evaluación asumen tres retos metodológicos importantes, ya que proporcionan medidas indirectas, probabilísticas y de interpretación relativa.

Las mediciones son indirectas porque el resultado de una prueba refleja parcialmente aquello que pretende medir. Esto conecta con la supuesta representatividad de los ítems. Tras un curso de primeros auxilios, ¿qué preguntas reflejan la capacidad de alumnos y alumnas para tratar una herida? ¿El examen teórico para obtener el permiso para el manejo de vehículos o embarcaciones informa del nivel de preparación del futuro conductor? Técnicamente, estas preguntas tratan de la validez de las inferencias que se realizan a partir de las puntuaciones del examen (por ejemplo, como la puntuación obtenida es alta, el alumno sabe tratar una herida, o como el alumno ha obtenido una puntuación baja, no tiene conocimientos suficientes para manejar una embarcación) y son frecuentes las críticas a exámenes memorísticos o sensibles al entrenamiento previo y a la conjetura, que no reflejan la supuesta preparación del examinado. ¿Un examen escrito de inglés refleja realmente el nivel la competencia lingüística del examinado? ¿Qué expresa realmente el resultado de los exámenes MIR (médico interno residente) o PIR (psicólogo interno residente)?

En el caso concreto de AM, estas dudas se extienden al propio formato. Muchos estudiantes prefieren exámenes tipo test por la creencia de que ofrecen más oportunidades de acierto que las preguntas abiertas. Por definición, si los ítems son representativos del nivel del examinado, el formato de las preguntas no debería intervenir en la puntuación, pero en realidad no es así. A menudo, el entrenamiento previo con modelos de examen, la conjetura y otros factores intervienen en la puntuación.

En segundo lugar, el problema de la medida probabilística se refiere a que toda puntuación de un test incorpora siempre un componente de imprecisión. Psicométricamente, las puntuaciones de un test nunca son valores exactos (por eso la imprecisión está ligada intrínsecamente a la fiabilidad, o más bien a la falta de ella), sino estimaciones de la verdadera medida. ¿En un examen en que se aprueba con un 5, una

puntuación de 4,95 implica suspenso? En muchos casos es así, aunque esto implique asumir que el instrumento tiene suficiente precisión para discriminar una diferencia de 0,05 puntos. ¿El examen tiene realmente esta sensibilidad? En la práctica es posible estimar los niveles de fiabilidad y precisión de una prueba. Lo más complicado es que el resultado suele desconcertar a los autores del test, ya que les obliga a tomar conciencia de la gran imprecisión que suelen tener las evaluaciones realizadas.

El último aspecto, la relatividad de las medidas, afecta especialmente a los test (por ejemplo, a las pruebas de evaluación psicológicas), no tanto a los exámenes. En psicología, la puntuación de un test suele interpretarse, siguiendo un enfoque normativo, en función de las puntuaciones obtenidas por un grupo de personas que se toma como referencia (el denominado grupo normativo); por tanto, la interpretación de una determinada puntuación dependerá de (será relativa a) las características de dicho grupo normativo. En cambio, en educación, las puntuaciones del examen se suelen interpretar bajo unos criterios establecidos por el/la autora, el/la docente, el programa, etc. (enfoque directo o criterial).

De estos tres problemas deriva otro más concreto: el sesgo de las medidas. En la teoría de los test se define el sesgo según alguna persona o un determinado colectivo queden sistemáticamente perjudicados al responder la prueba. Un caso cotidiano lo constituyen los exámenes con distintos modelos (versiones) para diferentes grupos de alumnos. Si, por ejemplo, estas versiones no son comparables en dificultad, y un grupo recibe un modelo con preguntas más difíciles que el resto, este grupo se verá claramente perjudicado en los resultados que obtenga. En otros casos, el resultado de un examen está condicionado por la capacidad de comprensión lectora de los alumnos. Por ejemplo, para responder y acertar problemas de cálculo mental es necesario entender las preguntas, y esto perjudica a los alumnos que, aun poseyendo una buena capacidad matemática, no tienen suficiente nivel de comprensión lectora.

Los test de AM suelen penalizar los errores (los errores restan) a fin de que el examinado sea prudente y evite responder a base de conjeturas

intentando adivinar la respuesta. Esto tiene efectos colaterales ya que involucra aspectos emocionales y personales (impulsividad, fatigas, ansiedad, autoimagen, aceptación del riesgo, miedo al fracaso, etc.) que intervienen al afrontar las preguntas. El sesgo en estos casos estriba en que algunos perfiles de personalidad tienden a puntuar peor por cuestiones independientes de su capacidad.

En los exámenes de AM también se produce otro fenómeno asociado a la experiencia y entrenamiento. Tras resolver muchos modelos de prueba similares al examen real, el examinado adquiere una habilidad para identificar estilos y patrones en las preguntas que le ayudan a detectar la alternativa correcta. De hecho, en muchas certificaciones y acreditaciones oficiales se aconseja a los candidatos que entrenen a base de responder más y más modelos de pruebas. El sesgo aquí perjudica a los que no han entrenado y responden «solo» según su capacidad o los conocimientos relativos a lo que se evalúa.

Eliminar posibles sesgos ha sido uno de los argumentos tradicionales a favor del uso de test de AM, ya que en ellos el examinado no ha de crear ni elaborar la respuesta, solo escoger entre varias opciones, por lo que se excluye así cualquier subjetividad en la corrección. En los test de AM no interviene la buena o mala letra del examinado, tampoco su capacidad de expresión. Todo ello llevó en su momento a identificar los test de AM como «objetivos», puesto que la plantilla de corrección aparentemente elimina subjetividades y errores de puntuación (factores humanos). Por este motivo, en muchos ámbitos educativos se ha elevado a los exámenes de AM a un valor de rigurosidad que en realidad no está asegurado. En estos casos, simplemente se está confundiendo la objetividad del método de corrección con la objetividad (buen funcionamiento) del instrumento. Realizando auditorias (controles de calidad) de pruebas oficiales y exámenes académicos, hemos podido detectar muchas disfunciones graves en estas pruebas. Las causas suelen asociarse a ítems que no distinguen examinados con mayor o menor nivel (no discriminan) o a plantillas de corrección que presentan dudas respecto a su adecuación. En muchos exámenes, contra lo esperado, existen preguntas con más de una alternativa que funciona mejor que la que supuestamente es la correcta (los examinados de mayor nivel tienden a escoger sistemáticamente una alternativa que no es la que consta en la

plantilla). La situación más grave se da cuando la respuesta que desde un punto de vista psicométrico funciona mejor que la «correcta» es la omisión (los examinados de mayor nivel tienden sistemáticamente a dejar en blanco la respuesta).

Todas estas constataciones son habituales y llevan a situar el valor de la objetividad del test en muchos otros elementos diferentes al de su formato o la existencia de una plantilla de corrección.

1.3. Diseño del examen

Una vez decidido el formato de los ítems (AM) y el contenido que evaluar, el proceso completo de creación de un examen pasa por seis fases cualitativas y, opcionalmente, por cinco cuantitativas.

El primer bloque de fases afecta a todas las pruebas, y trata del proceso de creación y administración del examen. La información la genera o recoge básicamente el autor de la prueba. En cuanto al segundo bloque (análisis de los resultados), es recomendable, pero no siempre factible, puesto que precisa un software adecuado de análisis:

Bloque 1
1. Establecer tabla de especificación de objetivos (TEO)
2. Reglas de generación de ítems (RGI)
3. Crear un banco de ítems (BI) y establecer un vector descriptor de ítem (VDI)
4. Establecer un vector descriptor de personas (VDP)
5. Maquetar, editar la/s forma/s de la prueba y decidir el proceso de aplicación
6. Sistema de puntuación y publicación de resultados (sin garantías)

Bloque 2
7. Recoger la matriz de datos en bruto
8. Auditoria psicométrica de la prueba y propuesta de reajustes en el examen
9. Sistema de puntuación y publicación de resultados (con garantías)
10. Análisis comparativo de los datos de la auditoria con VDI y VDP
11. Efectos y reajustes en las fases 1, 2, 3 y 4

Esta secuencia de trabajo es orientativa y admite variantes. Por ejemplo, el entorno de aplicación cambia según sea un examen convencional en papel, o bien online. En el primer caso pueden definirse diversas versiones, modelos o formas con las mismas preguntas, aunque presen-

tadas en diferente orden. Esos modelos también pueden construirse con preguntas diferentes siempre que se asuma que su representatividad y dificultad son equivalentes.

En las pruebas online también se puede aleatorizar el orden interno de las preguntas y el de las alternativas de cada pregunta; en este caso, un orden diferente para cada examinado/a. Así, diferentes examinados pueden responder a un conjunto de preguntas, iguales o distintas entre ellos, presentadas en diferente orden y viendo sus alternativas en posiciones diferentes. Cuando se emplean como *quiz* también permiten asociar mensajes a cada alternativa de respuesta, de modo que, si son escogidas, ofrecen un *feedback* explicativo de las respuestas correctas y, en especial, de las incorrectas. En todos estos casos, el evaluador asume un principio de independencia local (IL) por el cual la respuesta a una pregunta solo viene dada por la capacidad del examinado y no por efecto de repuestas a otras preguntas (aprendizaje progresivo, eliminación, conjetura, efecto de halo...). La IL es una condición difícil de asumir, y más sin un análisis psicométrico que la garantice.

Otras ventajas de las pruebas online son el registro del tiempo dedicado a cada respuesta y de la secuencia de rectificaciones. En los siguientes apartados desarrollaremos con más detalle ambos bloques.

2. LA TABLA DE ESPECIFICACIÓN DE OBJETIVOS

El primer paso estratégico para crear un examen consiste en definir una matriz o tabla que combine las áreas de contenido a evaluar (filas) y alguna clasificación del tipo de conocimiento que se quiere evaluar (columnas). La combinación de ambos aspectos es una tabla de especificación de objetivos (TEO). Los ítems de la prueba presentarán las características identificados en las celdas de la TEO.

Los contenidos y objetivos del curso que evaluar (filas) tienen que ver con las partes, unidades didácticas o formativas, módulos, temas y subtemas, etc., en que se estructura la materia. Una misma materia puede ser abordada desde diferentes niveles de complejidad, dependiendo de los objetivos educacionales. Estos son los que se representan en las columnas de la tabla. No es lo mismo preguntar sobre qué es la «humedad relativa», que plantear cuál es su utilidad práctica o su relación con otros indicadores como la temperatura o presión atmosférica. En el primer caso, al examinado le basta con recordar la definición, mientras que en los otros debe haber comprendido el concepto y ser capaz de relacionarlo con otros.

Uno de los sistemas más reconocidos para estructurar los niveles de complejidad lo constituye la taxonomía de Bloom, que contempla estos seis niveles o enfoques, descritos desde el más básico al más complejo:

- **Conocimiento**: capacidad de recordar cosas (términos, principios, normas, métodos, teorías, etc.) previamente aprendidas. Se basa en la memoria, y no asume que necesariamente se comprenda lo que se recuerda En un examen son preguntas que requieren definir, distinguir, ilustrar, identificar, recordar, reconocer, etc. Son útiles para valorar vocabulario, terminología, definiciones, nombres, fechas, personas, lugares, propiedades, fenómenos, formas, usos, costumbres, reglas, símbolos, estilos, acciones, procesos, clasificaciones, categorías, métodos, técnicas, tratamientos, principios, fundamentos, leyes, elementos, teorías, modelos, fórmulas, etc.
- **Comprensión**: capacidad de captar el significado o sentido directo de la información presentada (de forma verbal, gráfica, simbólica, etc.).

Es un nivel superior al anterior, ya que implica una interiorización de lo aprendido. Las preguntas requieren una interpretación personal de los temas planteados. Ante un concepto conocido, el examinado debe describirlo empleando palabras distintas, o distinguir sus aspectos esenciales, o derivar consecuencias directas y evidentes. En un test son preguntas que requieren explicar, interpretar, diferenciar, distinguir, demostrar, inferir, concluir, predecir, etc.

- **Aplicación**: capacidad para utilizar los conocimientos en la resolución de problemas. Las preguntas plantean al examinado la solución de situaciones y problemas nuevos; para ello se emplean principios, reglas, métodos, teorías, etc., previamente aprendidos. Los modelos de problemas han de ser similares, pero no iguales a los tratados durante el aprendizaje. De esta forma, el examinado aplica, desarrolla, organiza... sus conocimientos en circunstancias diferentes de las empleadas como ejemplos.
- **Análisis**: capacidad para desglosar la información recibida identificando los elementos que la componen, así como sus interrelaciones y estructura. En estos ítems, el examinado ha de reconocer, detectar, distinguir, identificar, clasificar, discriminar, contrastar, comparar, distinguir, etc., las diferentes partes que constituyen el objeto de la pregunta.
- **Síntesis**: capacidad de reconocer los elementos y partes que forman un todo. Se trata de manejar fragmentos, partes o componentes, de organizarlos y combinarlos de manera que formen una estructura nueva que inicialmente no se mostraba como tal. Con la información disponible, el examinado ha de generar un producto original (idea sobre un tema, plan de acción, modelo explicativo...). En estos ítems, el examinado preferiblemente ha de poder escribir, producir, transmitir, modificar, documentar, planificar, producir, diseñar, modificar, especificar, etc.
- **Evaluación**: capacidad de emitir juicios cuantitativos o cualitativos sobre el valor o mérito de ideas, métodos, instrumentos, resultados, proyectos, programas, etc. Implica un pensamiento crítico (no una opinión), a partir de unos criterios preestablecidos. En estos ítems no basta conocer y comprender una determinada materia o contenido, sino que hay que formular juicios de valor en términos lógicos o ajustados a normas y reglas. El examinado ha de juzgar, argumentar, validar y decidir.

La tabla 1 muestra un ejemplo de TEO para un examen de una asignatura de educación industrial en un curso sobre sistemas de suspensión y dirección de automóviles y su reparación.

Tabla 1. Ejemplo de una tabla de especificación de objetivos

		Con.	Com	Apl.	Tot.	%
Chasis y suspensión	Elásticos y amortiguadores de choque	7	3		10	8.3
	Alineación	9	6	2	17	14.2
Principios de funcionamiento	Mecanismos de dirección	2	1		3	2.5
	Principios estabilizadores	3	5	1	9	7.5
Servicio y reparación	Diagnóstico de fallas		7	4	11	9.2
	Uso de las herramientas	7	1	3	11	9.2
	Técnicas de alineación	1	1		2	1.7
	Dirección y balanceo	1		1	2	1.7
Tipos de sistemas de frenos	Campanas y patines	2	2		4	3.4
	A disco	1	2		3	2.5
	Hidráulicos	4	1		5	4.2
Principios de funcionamiento	Presiones mecánico-hidráulicas	2	4	8	14	11.7
	Coeficientes de fricción	3	3		6	5
Diagnóstico y mantenimiento	Indicaciones de falla y ajustes	4	5	8	17	14.2
	Reparación de campanas, líneas y cilindros	2	1	1	4	3.4
	Uso equipos y herramientas	2			2	1.7
	Total	50	42	28	120	
	%	41.7	35	23.3		100

La TEO determina la estructura del futuro examen. En este ejemplo se ha decidido que la prueba conste de 120 ítems distribuidos en seis contenidos diferentes (filas principales) que a su vez se subdividen en otros más concretos (filas secundarias). Pese a la orientación aplicada de la asignatura en esta TEO, los autores optaron principalmente por las categorías de conocimiento» y comprensión» (76,7% de los ítems, como se puede ver en la parte inferior sombreada de la tabla) de la taxonomía de Bloom y solo un 23,3% de aplicación. En la práctica esto es habitual, ya que con ítems AM las categorías que mejor funcionan son los de conocimiento y comprensión, siendo progresivamente más difícil crear ítems de análisis, síntesis y evaluación, más idóneos con preguntas de respuesta abierta.

Si bien la taxonomía de Bloom es la más conocida, existen otras derivadas, también consolidadas, como la FIO (*the framework for intructio-*

nal objectives taxonomy) o LOGIQ (*logical operations for generate intended questions*). FIO es una propuesta práctica general que relaciona categorías intelectuales con otras de tipo psicomotor y de valores y actitudes en el ámbito afectivo. En el diseño de exámenes, las categorías intelectuales contempladas son diez: interpretar, clasificar, inferir, comparar, generalizar, sintetizar, analizar, hipotetizar, evaluar y predecir. La taxonomía LOGIQ está mucho más orientada a la creación de preguntas y contempla seis categorías: repetición, resumen, explicación, predicción, aplicación y evaluación.

No existe una regla fija sobre la cantidad necesaria de filas ni de columnas para elaborar una TEO. El número de filas debe venir determinado por las partes de la materia evaluada que se consideren relevantes y el de columnas, por aquellos enfoques o tipos de conocimiento que se desean alcanzar. En cualquier caso, como principio general, la TEO resultante debería ser un buen reflejo del proceso de enseñanza-aprendizaje seguido. Las celdas de la tabla (combinaciones de contenidos y niveles de complejidad) deben proporcionar elementos de ayuda en la creación de las preguntas correspondientes. La importancia o «peso» que tenga en la evaluación cada celda de la tabla y, por tanto, la cantidad de ítems necesarios para poder representar dicho peso vendrá dado, de nuevo, por: 1) el programa de la asignatura (extensión, estructura…); 2) la importancia asignada a cada una de las partes de la materia (filas), que puede ser proporcional a la cantidad de tiempo invertido en la docencia de cada una de dichas partes; 3) los tipos de niveles de complejidad ejercitados en el proceso de enseñanza-aprendizaje (columnas).

A modo de resumen, estas son algunas directrices y recomendaciones sobre el uso de TEO.

- La estructura de la TEO debe reflejar el contenido de la materia y el programa del curso a evaluar. Esto resulta fundamental para validar el examen.
- La TEO debe ser exhaustiva en cuanto al desglose de los contenidos de la materia a evaluar.
- También debe representar correctamente la complejidad de los niveles de conocimiento (no suele ser recomendable sobrecargar la prueba con ítems de conocimiento).

- Por su naturaleza, el formato AM no siempre se ajusta plenamente a la taxonomía de Bloom.
- En ocasiones, la estructura de la TEO acaba evidenciando la existencia de dos o más bloques de contenidos muy diferenciados. En estos casos es recomendable considerar si se trata de una prueba o bien son varias. Esto es importante si está previsto un análisis o auditoria.
- Conviene revisar el número y dificultad prevista para los ítems de cada celda de la tabla y detectar posibles sesgos indeseados (por ejemplo, es frecuente que los ítems más difíciles sean de una parte de la materia o categoría concreta de la taxonomía). La dificultad de la prueba tiene que guardar cierta relación con el nivel de los alumnos que van a ser evaluados (una misma prueba puede ser fácil para un grupo de alumnos y difícil para otro). Una vez determinado dicho nivel de dificultad global, conviene que no todos los ítems tengan un nivel de dificultad similar; algunas preguntas tienen que ser más fáciles y otras más difíciles, de manera que la mayoría de los examinados puedan encontrar un nivel similar al suyo.
- La dificultad de los ítems prevista por la persona que diseña la prueba debería distribuirse entre las diversas celdas.
- En relación con el valor o «peso» de los ítems de una celda, es preferible añadir o reducir su número, más que ponderar las respuestas correctas.

3. REGLAS DE GENERACIÓN DE ÍTEMS (RGI)

Una TEO es un buen recurso para iniciar el diseño de un examen, pero a la hora de crear el material que compondrá la prueba hacen falta pautas más concretas. Las RGI son precisamente el conjunto de directrices que guían el proceso de creación de las preguntas.

En su forma más estándar, los ítems AM están constituidos por un enunciado y varias opciones/alternativas, una de ellas correcta y el resto incorrectas (también denominadas distractores). La misión de la persona examinada es seleccionar la opción correcta de cada ítem. El enunciado se puede expresar directamente en forma de aseveración o de manera interrogativa. Las alternativas pueden expresarse de manera verbal (a través de frases), numérica (por ejemplo, con fórmulas matemáticas) o gráfica (por ejemplo, con imágenes).

Aparentemente, los ítems AM son fáciles de crear, aunque la experiencia muestra que construir un buen ítem AM no resulta una tarea sencilla. Las siguientes recomendaciones son directrices de buenas prácticas para crear ítems. La lista, que no es exhaustiva, combina pautas generales propuestas con constataciones de problemas habituales en los procesos de revisión y auditoria.

3.1. Recomendaciones generales

- El número de alternativas se debe mantener constante a lo largo del test. Esta es una condición importante cuando se aplican penalizaciones de los errores a partir de fórmulas establecidas (ver anexo).
- El número de alternativas puede oscilar de tres a diez, siendo habitual tres, cuatro o cinco. (El apartado 7.4 profundiza en este punto).
- Es preferible distribuir las alternativas verticalmente y no horizontalmente.
- Hay que comprobar que no hay errores ortográficos, gramaticales, abreviaturas no utilizadas previamente, etc. Los docentes tienden a prestar más atención al contenido y a la forma de las alternativas correctas, por lo que a menudo las erróneas presentan más faltas orto-

gráficas y de expresión (este detalle puede proporcionar pistas indeseadas).

- Hay que comprobar que el ítem trata un aspecto relevante del contenido de la asignatura (definido en las filas de la TEO) y un nivel de complejidad deseado (definido en las columnas de la TEO). Es tarea del docente prever las características del alumnado, que debería dar una respuesta correcta o una incorrecta (incluso del tipo de respuesta incorrecta) a cada ítem.
- Hay que comprobar que en cada ítem haya una única respuesta correcta (o en algunos casos, especificada en las instrucciones como mejor). En muchos auditorias de exámenes surgen problemas con la plantilla de corrección, ya que se detectan alternativas que, siendo consideradas erróneas, se comportan como alternativas correctas (por ejemplo, cuando los examinados de mejor nota escogen una alternativa errónea concreta u omiten su respuesta, mientras que los de peor nota escogen la alternativa supuestamente correcta).
- A no ser que se trate de una prueba de comprensión verbal, léxica, ortografía, etc., no hay que confundir la capacidad evaluada con la de comprensión del redactado. El razonamiento verbal y la comprensión lectora son importantes y necesarios, pero no deben constituir un elemento que condicione las respuestas a la prueba.
- Todos los ítems deben ser independientes (principio de IL). No debe haber conexiones ni interdependencia de contenidos entre los ítems (así, la respuesta a un determinado ítem no debe depender de haber contestado correctamente a un ítem previo).
- Es conveniente evitar términos absolutos como *todo*, *nada*, *siempre*, *nunca*, etc. Normalmente van asociados a alternativas erróneas.
- Es recomendable evitar frases hechas o tópicos, así como expresiones (o ejemplos) literales de libros de texto o apuntes.
- Es preferible evitar enunciados y alternativas en modo negativo y las dobles negaciones. Si hay ítems con elementos negativos en el enunciado, dicho elemento hay que indicarlo en mayúsculas (NO) o subrayándolo (no) para que sea bien visible.

3.2. Recomendaciones referidas al enunciado

- El enunciado del ítem ha de exponer una cuestión o formular un problema de manera comprensible, clara y específica, y sin necesidad de leer las alternativas. Hay que evitar explicaciones innecesarias.
- La información básica de la pregunta se debe encontrar en el enunciado, que debería ser más largo que las alternativas.
- Evitar regular la dificultad del enunciado (por tanto, de la pregunta) utilizando un vocabulario de difícil comprensión, distinto al habitual o que no resulta familiar para el examinado.
- Por regla general, una de las opciones de respuesta es correcta y el resto son incorrectas, pero en determinadas ocasiones ocurre lo contrario (una opción es incorrecta y el resto correctas). En este caso, cuando la respuesta solicitada es la incorrecta, esa especificación debería quedar destacada (por ejemplo, subrayando) en el enunciado.
- El enunciado de un ítem no debe poder ayudar a responder otro (principio IL).

3.3. Recomendaciones referidas a las alternativas

- Una buena manera de plantear las alternativas es preparar una breve justificación de cada una (por qué son o no correctas). Esto es habitual en pruebas online que ofrecen retroalimentación en cada posible respuesta. El hecho de redactar una justificación en cada opción obliga a reflexionar sobre su adecuación, mejorando así la construcción adecuada del ítem, y proporciona argumentos ante posibles reclamaciones.
- Las alternativas deben ser lo más cortas posibles.
- Las alternativas no han de comenzar todas con una misma expresión. Es muy frecuente y sobrecarga el tiempo de lectura. Cuando esto ocurre es recomendable trasladar la expresión repetida al enunciado.
- Si las alternativas son figuras o gráficos mantendrán un tamaño y aspecto proporcional.
- El orden de presentación de las alternativas debe ser neutro, preferentemente alfabético (o numérico), ya sea por orden creciente o decreciente.
- Un ítem estará bien formulado si las frecuencias de elecciones de las alternativas erróneas se distribuyen aproximadamente de manera uniforme al margen de la opción correcta

- Un ítem no debe poder ser acertado, es decir, contestado correctamente sin dominar el contenido. Las personas sin domino de los contenidos tienden a responder por descarte, y las alternativas que, por algún motivo, no resultan atractivas son candidatas a ser rechazadas Por ello, la redacción de los distractores debe tener igual atractivo, longitud, estilo gramatical y aspecto que la opción correcta.
- Es conveniente formular las alternativas incorrectas de modo que correspondan a los errores más típicos cometidos por los examinados.
- Las alternativas tienen que presentar contenidos diferenciados y no solapados.
- No son recomendables las alternativas que afirman o niegan otras. Por ejemplo, se desaconseja usar la opción «Ninguna de las anteriores». De cualquier forma, en caso de que este tipo de opciones sea necesario por algún motivo, es preferible «Ninguna de las otras opciones», especialmente en pruebas online en que la posición de las alternativas puede variar para cada examinado. Por otro lado, existe la creencia extendida de que «Todas las opciones anteriores/otras son ciertas» suele ser cierta mientras que «Todas las opciones anteriores/otras son falsas» suele ser falsa.
- Un caso problemático es «Ninguna de las opciones es cierta». Si es la correcta suele ser un argumento habitual en procesos de reclamación ya que incurre en una contradicción al implicar la propia alternativa invalidando con ella al ítem.
- Es mejor evitar las alternativas humorísticas. Estos ítems suelen mostrar un mal funcionamiento en una auditoria y propician la conjetura.

3.4. Claves y patrones

Hay que evitar las pistas, claves reveladoras y palabras coincidentes entre el enunciado y las alternativas. Las más frecuentes son las claves verbales (asociación verbal entre la pregunta y su respuesta), claves gramaticales (al concordar gramaticalmente o por género o número el enunciado con una alternativa) y claves por heterogeneidad (por discordancia entre los conceptos).

Un efecto habitual al homogeneizar las alternativas es solapar sus contenidos (efecto indeseado, como se ha indicado), lo cual lleva al examinado a inferir conexiones y posibles patrones de acierto.

Hay que evitar que la alternativa correcta sea la de más contenido o extensión (este es un hábito muy frecuente cuando se redactan opciones).

En exámenes convencionales, en papel, se observan habitualmente patrones sistemáticos en la disposición de los ítems. A menudo son fruto de rutinas y desatención por parte de autores que no son conscientes de su existencia. Los patrones más comunes son de tres tipos: posición de la alternativa correcta en el ítem, posición de la alternativa correcta en la página y transición entre alternativas correctas (se detallan a continuación). Estos patrones tienen efectos colaterales negativos en la evaluación y son muy fáciles de evitar (basta con aleatorizar la presentación de las opciones).

Respecto a la posición de la alternativa correcta en el ítem, en exámenes universitarios, en papel no siempre existe una distribución homogénea de la posición de la alternativa correcta. La tabla 2 corresponde a un caso real de 55 preguntas de cuatro alternativas A, B, C y D. En la columna «Total» de la derecha consta el número de veces que cada letra es correcta. Se aprecia una predominancia entre B y C respecto a A y D, así como una diferencia de 7 entre la más frecuente (17) y menos frecuente (10). En el caso de ser una distribución homogénea, cada alternativa debería ser cierta 13 o 14 veces para evitar así la creación de expectativas.

Tabla 2. Distribución de las alternativas de respuesta (A, B, C y D) y del número de opciones correctas en un examen real de 55 preguntas

Alt. ✓	1	2	3	4	5	Total
A	1	1	2	2	4	10
B	3	3	3	4	4	17
C	1	5	4	3	2	15
D	6	2	2	2	1	13
Total	11	11	11	11	11	55

Respecto a la posición de la alternativa correcta en la página, esta es otra constatación habitual. Si dividimos las páginas de un examen en tres secciones verticales (ver figura 1), es frecuente que en la parte superior las respuestas correctas sean las últimas (C, D, E). En la sección inferior sucede lo contrario, las correctas suelen ser las primeras alternativas (A, B...). En cuanto al sector central, no hay un patrón tan claro y predominan otros hábitos (alternativa con más extensión, «Todas las anteriores son ciertas»...).

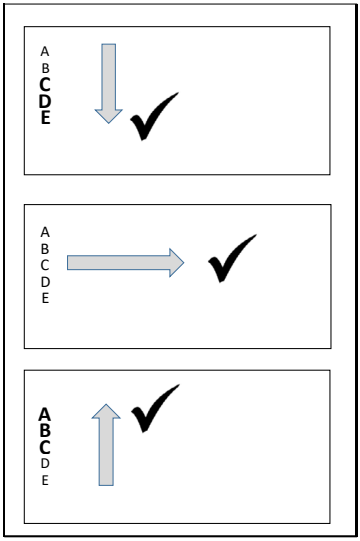


Figura 1. Distribución de alternativas correctas en un examen

Volviendo al ejemplo de la tabla 2, las columnas 1 a 5 corresponden a la posición de las respuestas correctas dentro de cada una de las 11 páginas que ocupa la prueba (cinco ítems por página). Las celdas indican el número de veces que cada alternativa es cierta destacando en gris el mayor valor. Se observa ya una cierta tendencia en las celdas sombreadas. La alternativa D es la más frecuente en el primer ítem de las páginas, mientras que A y B lo son en el último. Las posiciones segunda, tercera y cuarta se reparten principalmente entre las alternativas centrales. Este resultado vuelve a evidenciar una tendencia anómala que se podría evitar simplemente aleatorizando la posición de las alternativas en las páginas.

En cuanto a la transición entre alternativas correctas, en un examen no deberían existir patrones sistemáticos en la secuencia de aciertos (A-D-B-A-C...). De lo contrario, predisponen la elección de respuesta y en la decisión del examinado. La figura 2 muestra la frecuencia de transición entre dos ítems consecutivos de un examen de 160 preguntas de cuatro alternativas A, B, C y D.

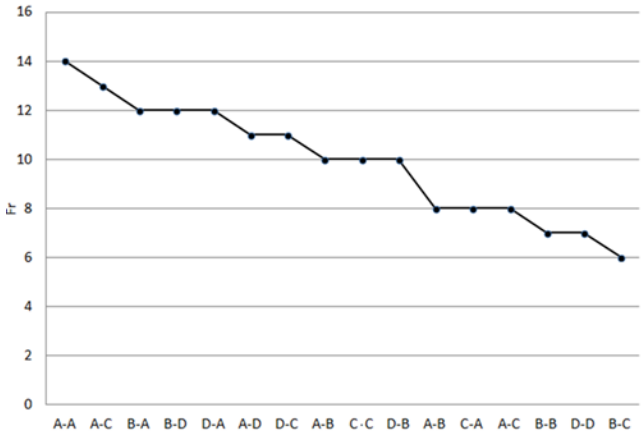


Figura 2. Frecuencia de transición entre alternativas (A, B, C y D) en pares consecutivos de un examen de 160 preguntas

En ítems de cuatro alternativas se pueden producir 16 transiciones (4^2). En la figura se han ordenado de mayor a menor frecuencia ocurrida en el examen. La transición más frecuente es A-A (14 veces) y la que menos B-C (seis veces). En caso de aleatorizar los tránsitos, cada uno de ellos debería ocurrir diez veces. Además, en este ejemplo destaca especialmente la diferencia del tránsito idéntico A-A por encima de los otros tres: C-C, B-B y D-D.

Una manera gráfica de comprobar este patrón es observar la secuencia completa del examen. La figura 3 corresponde a una prueba de 64 preguntas (eje X) de cuatro alternativas 1-2-3-4 (eje Y). El trazado grueso indica el itinerario entre alternativas. La línea delgada es un suavizado de la anterior para observar mejor la tendencia de la opción correcta. No se observa una pauta concreta (a veces el salto se da entre la alter-

nativa 1 y la 4, otra entre la alternativa 1 y la 2...), de manera que los «dientes» de la sierra son desiguales (a veces grandes, a veces pequeños). En caso de que la secuencia no estuviese suficientemente aleatorizada, podría verse una pauta concreta (por ejemplo, saltos más habituales entre las opciones extremas, o tendencias a repetir la misma opción de respuesta en ítems consecutivos).

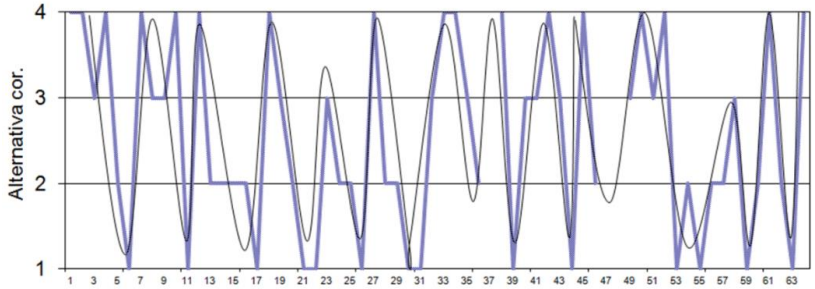


Figura 3. Patrón de tránsito entre alternativas correctas de ítems adyacentes

Sobre la combinación de patrones, estos y las claves reveladoras son la base no escrita de muchos recursos populares para acertar preguntas sin conocer la respuesta. Desde hace años circulan «reglas», «recetas», trucos o consejos entre examinados para afrontar los exámenes. En la red hay multitud de páginas, videos y foros dedicados a «cómo aprobar sin estudiar». Sus principales argumentos proceden básicamente del mal diseño de los test. Un examinado que no tiene nada que perder, a quien no le preocupa que le penalicen los errores, puede probar «suerte» y aplicar algunas reglas sencillas. Por ejemplo, ante cualquier duda, podría escoger siempre la alternativa con más palabras o contenido. Podría evitar marcar la A o la B en los primeros ítems de cada página y la D o la E en los últimos. También podría evitar las alternativas con errores ortográficos, etc. Si a estos aspectos añadiese las claves reveladoras particulares que facilitan algunos ítems, las posibilidades de su éxito aumentarían.

Respecto a los patrones de autor, la mayoría de docente tienen hábitos de repetición en sus clases. Repiten ciertas palabras y expresiones (latiguillos, muletillas, comodines, frases hechas, clichés...) que les caracterizan ante su alumnado, aunque sea de diferentes promociones.

Algo parecido sucede con sus apuntes y exámenes, ya que suelen arrastrar patrones personales que se mantienen curso a curso, de manera no consciente para el docente, pero sí para quien se va a examinar y dispone de modelos de apuntes e incluso de exámenes anteriores. Algunas páginas web ofrecen auténticas bibliotecas de recursos al respecto. Por ejemplo, algunos docentes se caracterizan por sus patrones de transición estables, otros convierten el «Ninguna de las anteriores» en un comodín para todas sus preguntas. Los patrones personales son difíciles de detectar, requieren una revisión cronológica de apuntes y ejemplares anteriores de exámenes que permitan detectar repeticiones sistemáticas.

En los exámenes se producen también otros patrones, que pueden ser peculiares y sorprendentes, como el de la ley de Benford. Según esta, en nuestro entorno predominan los números (no generados aleatoriamente) que empiezan por 1 respecto a los que comiencen por 2, 3 y así hasta el 9. Esta distribución sigue un perfil muy definido (30,1 % empiezan por 1; 17,6 %, por 2; 12,5 %, por 3, y así sucesivamente, hasta menos del 5 %, que empiezan por 9). Un estudio canadiense que revisó gran cantidad de ítems de exámenes confirmó que esta misma distribución se cumple con los números empleados por los docentes en las respuestas a preguntas de AM que implican cantidades y cálculos. Según esta ley, en caso de duda hay un 50 % de posibilidades de acertar si se escoge un resultado que empiece por 1, 2 o 3; mientras que estas se reducen al 15 % si comienza por 7, 8 o 9.

4. EJEMPLOS DE ERRORES HABITUALES EN LOS ÍTEMS

A continuación, veremos algunos ejemplos reales de problemas habituales en la redacción de ítems AM. Evitar estos errores no es solo una cuestión formal o de estilo. La experiencia práctica en auditorias psicométricas muestra que los ítems con estas anomalías suelen tener un mal funcionamiento evaluador y no resultan buenas «piezas» del examen.

En los ejemplos la respuesta especificada como correcta por el/la autor/a se destaca en **negrita**. Se ha respetado esta asignación, así como el redactado original. Varios ejemplos incumplen más de una directriz. Pertenecen a exámenes diferentes, por tanto, no pueden analizarse en conjunto intentando identificar los patrones antes expuestos.

En cada caso, primero se presentará el ítem y a continuación una breve descripción de los problemas asociados.

Ecosistema es...

A. La unidad fundamental para los estudios ecológicos.
B. Un conjunto de organismos vivientes en una comunidad, junto con su entorno.
C. Todo lo relacionado con las interacciones organismo-medio natural.
D. Todas las opciones anteriores son ciertas.
E. Ninguna de las opciones anteriores es cierta.

El enunciado es corto. Dos alternativas niegan o afirman otras. Utiliza el término *todo*. Este ítem no se podría administrar online aleatorizando las alternativas, ya que D y E condicionan la respuesta.

Marca la respuesta correcta:

A. Cuanto menos velocidad, más hacia la izquierda irá el viento (*veering*).
B. Cuanto más velocidad, más hacia la izquierda irá el viento (*baking*).
C. Cuanto menos velocidad, más hacia la izquierda irá el viento (*baking*).
D. Cuanto menos velocidad, más hacia la derecha irá el viento (*veering*).

El enunciado debe ser más largo que las alternativas. El término *cuanto* se repite y debería pasar al enunciado.

¿Cuándo no es incorrecto afirmar que una investigación es ex-post-facto?

A. Cuando el investigador trata con datos cualitativos
B. Cuando el investigador manipula la variable dependiente.
C. Cuando el investigador no manipula la variable independiente.
D. Cuando el investigador no tiene hipótesis.

Doble negación en el enunciado al combinar «no» con «incorrecta», además de otras negaciones en las alternativas C y D. Se repite «cuando» en las alternativas y debería pasar al enunciado: «Cuando el investigador…».

La trombosis venosa se manifiesta con los siguientes signos, excepto:

A. Dolor
B. Aumento de volumen
C. Calor local
D. Frío local

Si en preguntas anteriores se ha venido pidiendo la opción correcta, puede ser ahora una sorpresa en este ítem marcar la incorrecta (sería recomendable destacar «excepto» en el enunciado).

Ante un accidente, ¿cuál de las siguientes afirmaciones no es cierta?

A. Se recomienda reconocerlo en el mismo lugar en que se encuentra.
B. No se aconseja moverlo ni trasladarlo hasta que no se le hayan hecho la primera valoración de emergencia.
C. Trasladarlo puede agravar la situación o causarle nuevas lesiones.
D. Lo primero es trasladarlo a un lugar seguro.

Una negación en el enunciado y tres en una misma alternativa. En el enunciado debería quedar destacado «no es cierta» (por ejemplo, «NO es cierta» o «no es cierta»). Las alternativas tienen diferente longitud.

¿Cuál de las siguientes no corresponde a una lesión por congelación profunda?

A. Necrosis de la piel u óseas.
B. Curación en cuatro a seis semanas.
C. Hipersensibilidad al frío.
D. No deja secuelas.

Presencia de negaciones en el enunciado (sin que se destaque con negrita o subrayado) y la alternativa correcta lo que dificulta la comprensión: ¿La opción D significa que no corresponde que no deje secuelas?

Según la Organización Mundial de la Salud, en 2016 la esperanza de vida en Japón es superior a...

A. 80 años
B. 82 años
C. 84 años
D. 86 años

La esperanza de vida en Japón en 2016 era de 84 años, por tanto, las opciones A y B son correctas. Las alternativas de respuesta no son excluyentes.

Si sabemos que, en una población determinada de 1000 personas ,hay 300 mujeres y 700 hombres, y queremos realizar un muestreo estratificado-proporcional de 500 individuos, ¿cuántas mujeres tenemos que incluir?

A. 250
B. 175
C. 300
D. Ninguna es cierta.

Este ítem se puede acertar sin saber la respuesta correcta (150 mujeres). Una de las opciones (C. 300) es poco verosímil, porque se repite la misma cifra que en la población. Las alternativas no están ordenadas. En caso de mantener el «ninguna...», debería referirse a «las otras» o a «las anteriores».

Los golpes de calor:

A. No pueden darse en ambientes frescos y cálidos.
B. Se ven favorecidos por consumo de fármacos simpaticomiméticos.
C. No pueden prevenirse con una aclimatación previa.
D. No influyen factores nutricionales ni dietéticos.

El germen dentario comienza los movimientos eruptivos:

A. Cuando termina la calcificación coronaria y comienza la radicular.
B. Cuando llega a los dos tercios de raíz calcificada.
C. Justo al finalizar la calcificación radicular.

Exognasia es:

A. Falta de espacio para terceros molares.
B. Mordida cruzada bilateral.
C. Exceso de ancho de una arcada dentaria o de un maxilar en su totalidad.

El estilo consultivo en la toma de decisiones de los líderes es aquel que deja participar a sus subordinados en la decisión.

A. Verdadero.
B. Falso, solo a quienes tienen información.
C. Falso, solo les da o les pide información sobre su decisión.
D. Falso, solo recibe información o les da información sobre el problema.

En los cuatro ejemplos anteriores la alternativa correcta es la de más extensión. La alternativa A del primero parece incoherente.

Gen es...

A. La unidad del material reproductivo.
B. El fragmento de cromosoma que codifica la información genética del organismo.
C. El fragmento de un organismo que se reproduce sexuadamente.
D. Todas las opciones anteriores son ciertas.
E. Ninguna de las opciones anteriores es cierta.

Este ítem tiene el enunciado muy breve. Las alternativas tienen longitudes diferenciadas (y la correcta es más larga). En la opción correcta, la palabra *genética* alude a «gen», que aparece en el enunciado. Se incluyen opciones del tipo «Todas son...» o «Ninguna es...».

Las características de un grupo son varias, de entre ellas podemos definir:

A. Conjunto de personas que tienen interdependientes.
B. Grupo de personas que tiene uno o varios objetivos comunes y se organizan.
C. Conjunto de personas con intereses comunes.
D. Todas las anteriores.

La alternativa correcta tiene más contenido y comparte la palabra *grupo* con el enunciado (pista o clave). Se incluyen una opción del tipo «todas son…».

Biotecnología agraria es…

A. La modificación genética de moléculas de semillas o plantas con fines aplicados.
B. Un conjunto de técnicas de inseminación y polinización.
C. Una rama de la biología que estudia los problemas del campo.
D. Todas las opciones anteriores son ciertas.
E. Ninguna de las opciones anteriores es cierta.

Este ítem tiene un enunciado muy corto. Las alternativas D y E aluden a otras. La alternativa correcta tiene mayor contenido y una clave reveladora «semillas y plantas» conecta con el término «agrario» del enunciado. Se incluyen opciones del tipo «Todas son…» o «Ninguna es…».

La escala que se usa para explorar el estado de conciencia es:

A. Coma Glasgow Score.
B. Escala de Wales.
C. Escala de Greenwich.
D. Escala de Liverpool.

¿Qué tipo masticatorio favorece el crecimiento y desarrollo de los maxilares?

A. Masticación maseterina.
B. Movimientos de abre y cierre.
C. Movimiento de temporal.

En estos dos ejemplos, aun desconociendo la respuesta, existe una clave reveladora entre «conciencia»/«coma» y «masticatorio»/«masticación» en las respectivas alternativas correctas.

En el hemisferio norte, de espalda al viento, tienes a tu izquierda las bajas presiones y a tu derecha las altas presiones. Esto lo explica la ley de...

A. Buys Ballot.
B. Coriolis.
C. Murphy.
D. Ninguna es correcta.

En este ítem existe una alternativa humorística (C) que fue poco escogida y que aumentó las posibilidades de las otras tres. Incluye una alternativa (D) que niega todas, por lo que se incurre en una contradicción.

ADN o ácido desoxirribonucléico es...

A. El ácido que transmite la información heredada en los seres vivos.
B. Un ácido que caracteriza las células sexuadas.
C. Una molécula simple que compone a una bacteria.
D. Todas las opciones anteriores son ciertas.
E. Ninguna de las opciones anteriores es cierta.

Este ítem reduce las posibilidades a solo las alternativas A y B que comparten la palabra *ácido* (clave) con el enunciado. De nuevo, a la hora de escoger entre A y B la correcta es la que presenta mayor contenido. Se incluyen opciones del tipo «Todas son...» o «Ninguna es...».

¿Cómo se simboliza un frente frío en un mapa meteorológico? Una línea...

A. roja con triángulos.
B. amarilla con círculos.
C. verde con cuadrados.
D. Azul con triángulos.

Este ítem está afectado por la asociación habitual entre frío y color azul utilizada en muchos dispositivos cotidianos (grifos, refrigeración, conexiones...) e información meteorológica. La alternativa correcta es la única que empieza con mayúscula.

La rinorrea es:

A. La salida de sangre por la nariz.
B. Salida de líquido por la nariz.
C. Salida de sangre por la oreja.
D. Salida de líquido por la oreja.

En esta pregunta el enunciado es corto, se repite el inicio de cada alternativa («salida») aunque en la correcta es más completo («La») y vuelve a entrar en juego una asociación entre «rino» y «nariz».

No está permitido llevar:

A. Ornamentos en la cabeza, joyas y accesorios en el pelo.
B. Pantalones cortos.
C. Zapatillas deportivas.
D. Rodilleras recubiertas.
E. Protección para una nariz rota.

Pregunta sobre reglamento de baloncesto con dos alternativas B y C demasiado evidentes y sin atractivo que aumentan las posibilidades de las restantes. Enunciado corto y con una negación. La alternativa A es la más extensa.

La pelota se considera viva cuando:

A. La pelota sale a fuera por la línea de fondo.
B. La pelota sale a fuera por la línea de banda.
C. La pelota sale de la mano del árbitro en un salto entre dos.
D. La pelota está en las manos del árbitro.
E. La pelota está a disposición del árbitro.

En esta otra del mismo examen se repite el inicio de las alternativas («La pelota»). El enunciado debería ser «Se considera viva la pelota cuando….». De nuevo la correcta es la de más contenido. El 50 % de preguntas de esta prueba tenían la alternativa correcta con más texto.

I think the weather ..... be nice later.

A. will B. shall C. is going to D. is

En esta pregunta de un examen de inglés administrado a gran escala solo fueron escogidas las alternativas A y C (bajo atractivo de B y D). Aunque ofrece cuatro opciones funciona de hecho como un ítem de VF con solo dos. La disposición de las alternativas es horizontal.

Según los objetivos podemos distinguir diferentes tipos de grupos:

A. Grupos formales e informales.
B. Grupos de tarea y grupos de cohesión.
C. Grupos de relación y grupos de tarea.
D. Grupos de primarios y grupos de secundarios.

En esta pregunta se repite innecesariamente la primera palabra de cada alternativa.

En la deglución adulta o madura:

A. Se contraen los músculos posturales mandibulares.
B. Se contraen los orbiculares.
C. Se contraen los músculos cervicales.

En este ejemplo, el término «deglución» conecta con «mandibulares» y facilita la elección por conjetura. Además, la alternativa correcta es la de más contenido.

I ..... no satisfaction.

A. was able to get B. could get C. mustn't get **D. can't get**

La respuesta a este ítem viene condicionada por el título de una canción popular del grupo Rolling Stones. Al margen de su nivel de inglés, los aciertos aumentaban cuanta más edad tenían los examinados (sesgo).

5. BANCOS DE ÍTEMS (BI) Y VECTORES DESCRIPTIVOS

El resultado del proceso de creación de ítems siguiendo una estructura de TEO no es el examen final, sino un banco de ítems (BI). Al aplicar las RGI en cada una de las celdas de la tabla se está produciendo realmente una colección estructurada de ítems que sirven para configurar diferentes modelos o versiones del examen definitivo. Desde este punto de vista, un modelo concreto de examen constituye una muestra posible de todas aquellas que representen correctamente la configuración de la TEO. En función de la cantidad de ítems generados en cada celda de la TEO se podrán extraer más o menos modelos con ciertas garantías de equivalencia (los ítems de los diferentes modelos proceden de la misma TEO; por tanto, comparten contenidos y objetivos).

En el caso de pruebas online, algunas plataformas permiten clasificar los ítems en diversas categorías y niveles en función de criterios establecidos por el autor (por ejemplo, importancia, dificultad, frecuencia de aparición en la docencia...). En este caso, una vez introducidos los ítems en la plataforma y definidos los criterios de configuración del examen (por ejemplo, proporción de ítems de cada materia y nivel de complejidad, limitaciones en función de la dificultad...) bastará determinar la longitud deseada para la prueba, para que la aplicación genere diferentes modelos aparentemente equivalentes. Algunas de estas aplicaciones incorporan también datos posteriores a la administración del examen una vez analizadas las respuestas. Generalmente se trata de índices psicométricos que informan del funcionamiento de cada pregunta (dificultad, discriminación, conflicto entre alternativas, etc.) y que sirven luego para contrastar los datos propuestos por el autor con los datos del análisis.

La noción de BI ha tomado fuerza con los años gracias a las propuestas psicométricas como la teoría de respuesta de ítem (TRI) y los test adaptativos informatizados (TAI). En estos test, un algoritmo decide, en función de los aciertos y errores del examinado, las preguntas que le va presentando por pantalla. Bajo este enfoque el centro de atención es el banco del que se extraen consecutivamente las preguntas. El test que finalmente responde el examinado constituye solo una breve selec-

ción del contenido del banco ajustada al caso particular de la persona evaluada.

5.1. Vector descriptor de ítem (VDI)

Todos los ítems que forman un examen están caracterizados por múltiples aspectos de diseño que pueden aportan información útil a diversos niveles. La unión de todas estas características codificadas para cada pregunta constituye el vector descriptivo de ítem (VDI).

Ya sea un examen generado automáticamente desde una aplicación, ya sea elaborado de la manera convencional, un recurso interesante para revisar la estructura de la prueba consiste en agrupar todos los VDI. La tabla de la figura 4 muestra un ejemplo para un examen de 50 preguntas de cinco alternativas. Cada fila de la tabla corresponde a una de las 50 preguntas. Las columnas contienen la información que definen los VDI en este caso. De esta forma, una determinada fila (vector) contiene toda la información relevante para describir las características del ítem correspondiente (por ejemplo, en la figura 4 se ha sombreado el VDI de la pregunta 4). Según el contexto, cada examen permitirá dimensionar VDI de modo diferente. En este caso se han considerado relevantes los nueve criterios (columnas) siguientes:

VDI Pre

	Autor	Bloom	Tema	Dif.	Imp.	Doc	Alt.	Expo.	Long
ít 1	1	1	1	1	3	1	4	0	0
ít 2	1	1	2	2	2	3	5	1	1
ít 3	2	2	3	2	2	1	1	2	0
ít 4	1	3	4	3	3	4	3	0	1
ít 5	1	3	5	2	1	3	2	0	0
ít 6	2	2	1	1	2	1	2	3	0
ít 7	2	1	2	3	2	3	4	2	0
ít 8	1	1	5	3	2	1	4	2	0
ít 9	2	2	4	2	1	1	3	3	1
…									
…									
ít 50	2	3	4	2	3	4	5	2	1

VDI Post

Dif.	Discr.	Solap.	Atract
0,2	0,3		0
0,3	0,4		0
0,7	0,6		1
0,1	0,2	2	0
0,2	0,4		0
0,8	0,5		2
0,5	0,4		0
0,4	0,6		0
0,7	0,5		2
0,6	-0,2	3	1

Figura 4. Ejemplo del vector descriptor de un conjunto de 50 ítems

- Autor: esta prueba fue desarrollada por dos docentes; los valores 1 y 2 indican quién de los dos hizo cada ítem.
- Bloom: 1 (conocimiento), 2 (comprensión) y 3 (aplicación) informan de la categoría de la taxonomía de Bloom a la que corresponde cada pregunta. En este examen solo se utilizaron estas tres características de la taxonomía (columnas de la TEO).
- Tema: de 1 a 5 se indica a cuál de los cinco temas de la asignatura pertenece cada pregunta (filas de la TEO).
- Dif.: cada autor estimó la dificultad de sus preguntas en tres niveles: fácil (1), media (2) y difícil (3). Esta estimación requiere un ejercicio casi empático, en el que el autor debe ser capaz de prever si la mayoría de sus estudiantes contestarán correcta o incorrectamente el ítem.
- Imp.: cada autor estimó de 1 a 3 la importancia del tema tratado en la pregunta en relación con el dominio de la asignatura.
- Doc.: indica la fuente principal de preparación y documentación del tema tratado en la pregunta: (1) clase magistral, (2) apuntes, (3) prácticas, (4) lecturas complementarias).
- Alt.: es la alternativa correcta del ítem, en este caso de 1 a 5 (A..., E).
- Expo.: indica el número de veces en que ya se ha utilizado (expuesto públicamente) el ítem en exámenes anteriores (0 veces, 1 vez...).
- Long.: si la alternativa correcta es la de más contenido (longitud), aparece un 1.

Podría haber más datos descriptivos, pero esta información ya sirve de ejemplo para comprobar si se producen relaciones y efectos sistemáticos entre algunos aspectos que puedan alterar el funcionamiento del test. La revisión de esta tabla es previa al uso del test (VDI pre); es necesaria porque anticipa problemas y ayuda a mejorar la prueba. En general, salvo que el marco general de la asignatura y el del examen indiquen lo contrario, no deberían producirse asociaciones imprevistas. ¿Tiene sentido que un autor suela colocar las alternativas correctas en una misma posición para ítems con un mismo enfoque de la taxonomía de Bloom? ¿Se espera que los ítems considerados más difíciles tengan que ver con una u otra fuente de información? ¿La dificultad de una pregunta está relacionada con el número de veces que se ha utilizado anteriormente? Cada autor o equipo responsable de un examen debería constatar si se producen efectos sistemáticos de diseño no deseado en

su material. En este ejemplo hay pocas verificaciones previas de este tipo. Podríamos comparar las columnas «Bloom» y «Alt.», y en este caso parece que las alternativas correctas de los ítems Bloom 1 tienden a ser las últimas (4 y 5), algo que habría que corregir.

El VDI adquiere realmente valor cuando se incorporan nuevos datos objetivos procedentes del análisis psicométrico (VDI post). En la parte derecha de la figura 4 se muestran cuatro nuevos indicadores (columnas) para cada pregunta del examen de ejemplo. A menudo las lectoras de exámenes y sistemas de evaluación online incorporan módulos de análisis básicos que proporcionan datos como estos. Basta con localizarlos e incorporarlos al VDI. Las cuatro que se incorporan en la tabla 4 son las siguientes:

- Dif: Expresa la proporción de aciertos de la pregunta. Oscila entre 0 y 1 siendo 1 un ítem que nadie ha fallado (fácil) y 0 uno que nadie ha acertado (difícil).
- Discr: Es el índice de discriminación de la pregunta. Informa del grado en que esta es capaz de distinguir entre los examinados más y menos preparados. Un ítem que discrimina tiende a ser más acertado por los examinados más preparados y menos acertados por los menos preparados (permitiría diferenciar o separar a quienes tienen más y menos capacidad en el aspecto evaluado). Un ítem que no discrimina es acertado y fallado indistintamente por examinados con diferentes niveles de capacidad. La discriminación es un indicador muy útil en el control de calidad de un test. Un examen con preguntas que discriminan poco es un instrumento de evaluación defectuoso. La discriminación puede calcularse de muchos modos. En este ejemplo se ha obtenido a partir de una correlación, y se consideran con suficiente capacidad de correlación los ítems con valores superiores a + 0,3. (Volveremos a tratar este índice en el apartado 8.2).
- Solap: Una opción más avanzada de análisis de las respuestas consiste en recalcular la discriminación para cada alternativa simulando que esta fuera correcta. De este modo, para cada pregunta se obtienen tantas discriminaciones como alternativas. Si una o más alternativas incorrectas discriminan igual, o discriminan mejor que la que es «oficialmente» correcta (la considerada en la plantilla de corrección) entonces existe un problema de solapamiento. Puede

que, por algún motivo, la alternativa considerada inicialmente como correcta no lo sea, o que haya más de una alternativa que funcione como correcta (alternativas confusoras). En cualquier caso, este resultado constituye una mala noticia, ya que no está claro el funcionamiento del ítem en el conjunto del examen sea el esperado. En auditorias reales, tomar decisiones sobre solo unos pocos ítems que presenten solapamiento puede llegar a distorsionar significativamente la lista de resultados del examen (aprobados que pasan a suspensos, suspensos que pasan a aprobado...). En la tabla del ejemplo, la columna «Solap» indica el número de alternativas solapadas en los ítems conflictivos (el ítem 4 de la prueba es el único que presenta solapamiento en dos opciones de respuesta).

- Atract: Este cuarto dato indica si hay (1) o no hay (0) problemas de homogeneidad de atractivo. En el apartado sobre la RGI se indicaba que las alternativas incorrectas de un ítem debían tener un atractivo similar. Por ejemplo, en una pregunta de cinco alternativas acertada por un 60 % de examinados se espera que el 40 % de errores se distribuya homogéneamente entre las cuatro alternativas incorrectas (10 % cada una). Esta condición es especialmente importante cuando se aplican fórmulas de penalización de la conjetura (apartado 7.2).

Estas nuevas características «objetivas» (VDI post) pueden compararse con sus equivalentes «subjetivos» o con otras características previstas (VDI pre). En el caso de la figura 4, no parece, en general, que haya relación entre la dificultad que estimaban los autores (Dif. en VDI pre) y la dificultad final obtenida a través del análisis de las respuestas a los ítems (Dif. en VDI post). Este es un resultado importante, ya que cuando ocurre indica un desajuste entre las expectativas de diseño de los docentes y la situación real de los examinados. Aunque el principal objetivo de los exámenes es obtener evidencia acerca del nivel de los estudiantes, esta comparación entre las dificultades estimadas y reales también nos proporciona información acerca del nivel de «conocimiento» que el docente que ha preparado la prueba tiene de sus estudiantes. Así pues, es un efecto colateral, pero positivo, de las evaluaciones, y muchas veces proporciona resultados sorprendentes para el docente. Bien gestionado, este tipo de conocimiento puede redundar en una mejora de la docencia.

Continuando con la comparación, a simple vista parece que los ítems que mejor discriminan (Disc. en VDI post) corresponden a temas tratados principalmente durante las clases (Tema en VDI pre). También se observa que los ítems más difíciles con menor proporción de aciertos (Dif en VDI post) son los que nunca se habían empleado antes en otros exámenes (Expo en VDI pre) y han sido elaborados principalmente por el autor 1 (Autor en VDI pre). Los problemas de atractivo de opciones (Atract en VDI post) están más asociados al autor 2 (Autor en VDI pre).

Hay que tener en cuenta que durante la aplicación del examen pueden aparecer nuevas características relevantes que cabe considerar en un VDI. Por ejemplo, a menudo los examinados plantean dudas y realizan consultas sobre una determinada pregunta. En la tabla VDI se podría incluir una columna que refleje el grado en que los ítems han sido objeto de consultas (qué tipo de duda, cuántas veces ha sido consultado...). Más adelante, podrá utilizarse este nuevo dato para contrastarlo con indicadores de problemas (baja discriminación, solapamiento...).

Las comparaciones podrían extenderse, pero con lo visto hasta aquí ya es posible comprobar la utilidad de este recurso. Proporciona elementos de reflexión. Cada docente debe identificar relaciones no esperadas entre las características previstas para los ítems y las reales; asimismo, en caso de detectarse, debe valorar la conveniencia de efectuar modificaciones en los resultados del examen (por ejemplo, no considerar en el cálculo de la nota final una pregunta si su comportamiento claramente no es el deseado). Aunque nunca constituye una garantía total, una buena planificación de la prueba (definición de la TEO y definición del VDI) minimiza el número de problemas posteriores. Sea como fuere, los problemas detectados en un examen deberían considerarse en el futuro. Solo de esta forma la tarea (repetitiva) de elaborar exámenes propia de cualquier docente podrá mejorarse.

5.2. Vector descriptor de persona (VDP)

De la misma forma que conviene acumular evidencia sobre los ítems en el VDI, también es recomendable hacer lo mismo con los alumnos en un vector que describa las características de la persona que considere-

mos que puede tener una influencia en el resultado de la prueba, o al menos, que pueda proporcionarnos alguna luz sobre el mismo. Cada columna de la VPD definirá un tipo de información relevante referidas a las personas. El tipo de información que se considere relevante dependerá de muchas circunstancias (del tipo de curso, de la información que se pueda recoger sobre los estudiantes...). Alguna información suele proceder de la propia ficha del estudiante, como, por ejemplo, el grupo de matrícula, pero a veces es posible que el estudiante asista a un grupo diferente al que está matriculado. Esta es una información más interesante que la «oficial». Información como el grupo de asistencia a las clases de teoría y de prácticas (¿es un alumno de un grupo masivo de diurno o de uno minoritario de nocturno?), si suele venir o no a clases (se puede registrar, por ejemplo, la asistencia). La información que describe a alumnos y alumnas suele ser difícil de recoger. No es necesario que una variable del VDP esté completa para todo el alumnado. Por ejemplo, a veces, en una tutoría, un alumno comenta que, además de venir a clases, recibe clases particulares. O que la materia que impartimos le «cuesta» particularmente. Puede que solo tengamos este tipo de información para este alumno, pero sería interesante disponer de ella incluyéndola en el VDP, porque es posible que nos ayude a entender mejor su resultado en la prueba.

Otra información relevante, asociada al alumno, pero muy contextual, es la que hace referencia a las condiciones de un examen en concreto. Sería muy interesante saber el lugar exacto que ocupó en el examen (número de asiento, fila y columna del aula...). En algunas ocasiones también si entregó rápido el examen o tardó en acabarlo. Si durante el examen tenía dudas y en qué ítems. Somos conscientes de que este tipo de información es muy difícil de recoger; en especial cuando se tienen grupos con muchos estudiantes, pero no queremos dejar de indicar lo útil que podría ser disponer de este tipo de información para contextualizar los resultados referidos al alumno que se describen en el apartado 8.2.

6. CONFIGURACIÓN FORMAL DEL EXAMEN Y PROCESO DE APLICACIÓN

Imaginemos un examen compuesto por diez preguntas. Supongamos que lo imprimimos en una sola hoja, de manera que, en la primera página, después de un encabezado en donde consta el nombre de la asignatura, la fecha de la evaluación, un espacio para la identificación del alumno y unas instrucciones globales sobre la prueba, queda espacio para ocho preguntas, de manera que las otras dos tienen que imprimirse en la página reversa. Si al final de la primera página no avisásemos de que las preguntas continúan en la otra cara de la hoja, ¿cuántos alumnos/as habrían dejado sin contestar las dos últimas preguntas simplemente por no haberlas visto?

Este ejemplo pretende alertar sobre la importancia de los detalles formales al confeccionar los ejemplares de examen. El objetivo es minimizar los aspectos formales que pueden generar errores indeseados. Hay que recordar siempre que los alumnos han de poder contestar correcta o incorrectamente a los ítems única y exclusivamente por su nivel de competencia en la materia evaluada, y no por otros factores irrelevantes.

La hoja o el cuaderno de examen debe comenzar con un espacio reservado a la identificación de la prueba y del alumno. A continuación, han de incluirse unas instrucciones donde quede claro, entre otros aspectos, el número de preguntas, de qué manera han de indicarse las respuestas o cómo se calculará la puntuación (por ejemplo, si se penalizan los errores o no, y de qué manera en caso afirmativo). Los ítems se presentarán a continuación, y si la impresión de los mismos ocupara más de una página, debería indicarse al final de las mismas una leyenda del tipo «Continúa en la página siguiente» o simplemente «Continúa». De la misma forma, es recomendable indicar al final «Fin de la prueba».

La forma en que se presentan las alternativas de respuesta puede complicar de manera innecesaria la comprensión del ítem. Veamos un ejemplo de un mismo ítem con las opciones de respuesta presentadas de maneras diferentes:

Presentación 1
¿Cuál de las siguientes capitales es de un país asiático?

1. Dacca 2. Manama 3. Phnom Penh
4. Suva

Presentación 2
¿Cuál de las siguientes capitales es de un país asiático?

1. Dacca 2. Manama
3. Phnom Penh 4. Suva

Presentación 3
¿Cuál de las siguientes capitales es de un país asiático?

1. Dacca 2. Manama 3. Phnom Penh 4. Suva

Presentación 4
¿Cuál de las siguientes capitales es de un país asiático?

1. Dacca
2. Manama
3. Phnom Penh
4. Suva

En general, se considera que la última presentación, que muestra las opciones de respuesta de manera vertical, es la que destaca con más claridad dichas opciones; por tanto, es la forma de presentación recomendada.

Si la corrección de las respuestas se va a realizar de forma mecánica, será necesaria una hoja de respuestas específica para que pueda ser interpretada por una lectora óptica. En caso contrario, las respuestas pueden marcarse en cada una de las opciones o se puede confeccionar una hoja de respuesta *ad hoc* que se añadirá al final de la prueba. En cualquier caso, las instrucciones de la prueba tienen que dejar claro cómo y dónde han de marcarse las opciones (con una cruz, sombreando una casilla...) y de qué manera se pueden anular opciones marcadas inicialmente como correctas.

Otro aspecto formal que hay que cuidar es el tipo y tamaño de letra. Los tipos de letra pueden clasificarse según diversos criterios, pero uno relevante para facilitar la lectura es el del remate en la base. Las hay con remate (tipos Serif), como Times Roman, y sin remate (Sans Serif) como Arial. El remate en la base de las letras proporciona una cierta continuidad o ligazón al configurar palabras; además, distinguen perfectamente mayúsculas o minúsculas (por ejemplo, l, L), algo que no siempre ocurre con los tipos Sans Serif. Así pues, los tipos Serif suelen facilitar la lectura, y por ello acostumbran a ser los elegidos por las editoriales para la publicación de libros impresos. Sin embargo, cuando lo que se desea es llamar la atención del lector (como en los avisos al final de página) se recomiendan los tipos Sans Serif por su contundencia, en especial cuando se trata de mayúsculas o negritas. En cualquier caso, se recomienda evitar la utilización de tintas de colores (cerca de uno de cada 10 varones es daltónico).

Por lo que se refiere al tamaño de la letra, el mínimo aconsejado es de 12 puntos (aunque eso depende también del tipo de letra concreta que se use). Más que el tamaño de letra, lo que afecta a la legibilidad del texto, es el espaciado interlineal. En este sentido, lo que se recomienda es que el espaciado sea al menos de igual tamaño que la letra, aunque la lectura del texto se optimiza cuando el interlineado es entre el 20 % y el 30 % (si se utiliza una letra de tamaño 12, un interlineado óptimo debería ser de entre 14 y 15 puntos). Todo esto suele aumentar el número de páginas necesarias para que quepan todas las preguntas del examen. En este caso, si el formato de los ítems lo permitiera, estos podrían distribuirse en dos columnas por página, con una separación evidente entre ambas.

Las instrucciones que se den a los alumnos en el momento de comenzar el examen no se pueden improvisar. Con anterioridad tienen que preverse todos aquellos aspectos que tengan una consecuencia relevante (tiempo de administración, respuesta a las dudas...). La persona encargada de administrar la prueba debe exponer las instrucciones de la manera más clara y neutra posible. Si la misma prueba se administra simultáneamente en salas diferentes, las personas que se encarguen de informar tienen que asegurarse de que ofrecen una información comparable en contenido, forma y tiempo.

Son las personas encargadas de la administración quienes tienen que asegurar un mínimo de condiciones de confort (temperatura, iluminación...).

A menudo ocurre que, una vez comenzada la prueba, hay que interrumpirla para proporcionar datos adicionales sobre ella (por ejemplo, una aclaración sobre algún ítem o sobre algún detalle de algún ejercicio). Abusar de este recurso no es recomendable porque suele desconcentrar al alumnado; además, no se suele garantizar que la información que se añade llegue a todos (muchos están tan preocupados por contestar a las preguntas que no se dan cuenta de lo que se está diciendo). Sin embargo, en caso de necesidad, conviene concentrar todas las informaciones de manera que el número de interrupciones sea mínimo. Si las interrupciones tienen que ver con dudas de los examinados, conviene recordar la necesidad de incluir dichas consultas a la tabla VDI para contextualizar mejor el posterior análisis de los ítems.

La mayoría de los docentes toman algún tipo de precaución relativa al fraude en las respuestas. La posibilidad de algún alumno vea y copie las respuestas de otro es la más frecuente. Para que se pueda dar una copia, debe haber una fuente de información (FI voluntaria o involuntaria) y una fuente receptora (FR siempre voluntaria); por lo que una prevención básica consiste en separar estas dos fuentes. Es sabido que la proximidad lateral entre el informante y el receptor favorece la copia, pero también otras disposiciones, como la conocida disposición en V que se muestra en la figura 5. En la figura se ve que la información procedente de la fuente de información situada en el centro de la primera fila puede transmitirse oblicuamente hasta llegar a la última fila. Una manera de contrarrestar este posible efecto es confeccionar pruebas equivalentes (modelos) o, por lo menos, permutaciones de las preguntas y de las opciones de respuestas, y repartirlos de manera que no coincidan los modelos ni lateral ni diagonalmente.

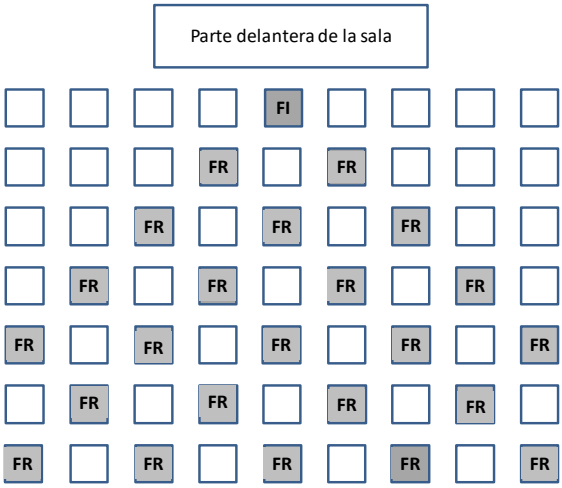


Figura 5. Disposición en forma de V que favorece la copia

7. SISTEMA DE PUNTUACIÓN

Una decisión muy importante a la hora de crear un test es la que hace referencia a cómo se puntuará a los examinados. La tendencia predominante es que todas las preguntas aporten la misma puntuación (respuestas incorrectas = 0, respuesta correcta = 1) y la suma del conjunto sea la puntuación final. Sin embargo, existen otras opciones que básicamente consisten en ponderar y penalizar o corregir la conjetura.

7.1 Ponderación de las respuestas

La ponderación de las respuestas consiste en dar más valor a unos aciertos que a otros. Esto puede hacerse bajo el criterio del docente/autor, que, como experto en la materia, decide el peso de cada pregunta. En estos casos conviene comparar los datos VDI subjetivos previos (importancia, dificultad...) con los posteriores más objetivos (apartado 5.1) y confirmar que realmente concuerdan. Sin esta constatación será dudoso que la ponderación refleje el valor/funcionamiento real de la pregunta.

Otro tipo de ponderación consiste en dar un valor diferente a la alternativa escogida o a una combinación de alternativas escogidas (aunque no es muy recomendable, en ocasiones se admite más de una opción como correcta). Aquí también interesa comprobar que el análisis de los ítems (indicadores) confirma la ponderación.

Existen ponderaciones que combinan el acierto con el nivel de certeza que manifiesta el examinado en cada respuesta. En estos exámenes, los ítems aportan dos datos, la respuesta y el porcentaje o grado de seguridad del alumno al darla (en una escala graduada arbitraria). Un acierto con una seguridad en la respuesta del 70 % pesará menos que otro acierto con una seguridad del 90 %. Esta forma de responder, y de ponderar, depende en gran medida de factores personales del examinados (autoconfianza, asertividad...) que pueden interferir en las respuestas y, por tanto, en las notas finales.

En otros casos, la ponderación afecta directamente a la puntuación global y no a los ítems. Se trata de ponderaciones basadas en la coherencia del patrón de respuesta del examinado (ver más adelante los patrones atípicos de respuesta).

7.2 Penalizar los errores

Consiste, a criterio del autor, en restar los errores, o una fracción de ellos, de los aciertos. En este enfoque no hay ninguna base matemática o fórmula que justifique la penalización. Básicamente, es una manera drástica de desincentivar la tendencia a responder cuando no está clara la respuesta.

Los examinados que desconocen la respuesta a una pregunta tienden a activar recursos alternativos para contestarla (intentar adivinarla, descartar opciones poco plausibles, etc.). La presencia de conjetura en un examen puede invalidarlo, ya que distorsiona el cálculo de las puntuaciones (no habría garantías de que la nota de un alumno sea un fiel reflejo de su nivel de conocimientos de la materia evaluada).

Tanto la penalización como la corrección de la conjetura, que veremos más adelante, comparten un efecto disuasorio que involucra otros factores personales. Un examen suele comportar presión, dadas sus consecuencias, generalmente vinculantes para el futuro de los examinados (superar la asignatura, pasar o no el curso, tener o no que recuperar la materia...). Todo ello provoca emociones, pensamientos negativos, etc., que condicionan de alguna forma la manera de responder (certeza, aceptación de riesgo, auto-concepto...) y determinan parte del resultado del examinado (una parte que no es la que se desea que quede reflejada en la nota, puesto que la nota debería poder interpretarse única y exclusivamente por el nivel de conocimientos evaluados).

Las estrategias anteriores no son las únicas para contrarrestar la conjetura. Una alternativa consiste en aumentar el número de aciertos (puntuación de corte) para la toma de decisiones finales (apto/no apto, baremo de calificaciones...) de modo que el peso de la corrección ya no reside en el examinado, sino en la exigencia de la prueba. En estos casos se suele aumentar el baremo de puntuación (subir el «listón») para compensar el posible efecto de la conjetura.

Otras opciones se basan en mantener la puntuación de corte, pero aumentando la dificultad de las preguntas o simplemente aumentar el número de alternativas. Por último, también existen estrategias analíticamente más complejas basadas en modelos psicométricos que tratan el efecto de la conjetura de manera diferenciada para cada ítem. Esta es la vía más reciente y sofisticada, ya que, al calcular la puntuación de un examinado, se tiene en cuenta el efecto diferenciado de la conjetura en cada uno de los ítems que ha respondido. De este modo, diferentes examinados con distintos patrones de respuesta (aciertos y errores en diferentes ítems) tendrán unas puntuaciones distintivas en función de su patrón.

Salvo casos a gran escala, este tratamiento de la conjetura aún es inusual en exámenes académicos, puesto que precisa tamaños de muestra importantes. Para ello se utilizan métodos analíticos y software especializado de la ya citada teoría de respuesta de ítem, especialmente el modelo de análisis de tres parámetros.

7.3. Corregir la puntuación

Desde hace años, el recurso más popular es la «corrección» (reducción) de la puntuación del examinado a partir de una supuesta justificación matemática (ver fórmula en el Anexo). Igual que ocurre con la penalización, esta opción conlleva unos riesgos y unos efectos colaterales que muchos evaluadores asumen sin saberlo.

La mayoría de fórmulas de corrección permiten estimar la cantidad de aciertos por conjetura y restarlos del total de aciertos obtenidos por el examinado. Con ello se «corrige» la puntuación obtenida permitiendo recalcular para cada examinado sus aciertos «reales» libres de efectos extraños. Muchas aplicaciones de corrección de test y exámenes incorporadas en lectoras ópticas y plataformas online suelen ofrecer el cálculo del anexo 1 entre sus *outputs*. Aparentemente, la fórmula soluciona de manera sencilla un problema que parecería mucho más difícil de abordar. Sin embargo, es interesante recapitular de nuevo las condiciones de aplicación en que se basa:

- Condición 1: Todos los ítems del examen deben tener el mismo número k de alternativas.

- Condición 2: Solo hay aciertos y errores, no omisiones (no se permiten «respuestas en blanco»).
- Condición 3: Todos los errores de los examinados se deben a que intentan acertar conjeturando y no lo consiguen.
- Condición 4: Cuando se enfrentan a preguntas que no dominan todos los examinados tienden a responder conjeturando y los fallos son debidos a no haberlo conseguido.
- Condición 5: Se asume que la posibilidad de escoger la alternativa correcta es equiprobable entre las alternativas. Dicho de otro modo, todas tienen el mismo atractivo (ya descrito en el apartado 3).

A efectos aplicados, esta lista plantea una primera duda general: ¿Se cumplen las condiciones en nuestros exámenes? De la que se derivan otras seis:

- ¿Todas las preguntas tienen igual número de alternativas? Hay exámenes en que varían. Si es así, no sería lícito aplicar esta corrección.
- ¿Aceptamos que en una fracción de las preguntas en que prueben suerte van a tenerla y acertaran? Esto implica que los examinados comparten un mismo estilo de respuesta.
- Aun compartiendo todos los ítems el mismo número de alternativas de respuesta (condición 1), ¿estas tienen un atractivo similar en cada pregunta (condición 5)?
- Esto es, puede asumirse que el número de aciertos por conjetura es proporcional al número de intentos contestados por conjetura (ver anexo: $C = C1/k$). En la mayoría de casos esto no se verifica. En la práctica, al pedir a los autores de un examen datos descriptivos de las preguntas (VDI pre), muchos admiten desequilibrios en el interés que pueden suscitar las diferentes alternativas entre sus alumnos/as. Al compararlo después con los resultados del análisis de los ítems (VD I post) se suele confirmar la existencia de un problema de diseño.
- ¿Es aceptable el supuesto de que, cuando los examinados no tienen suficiente capacidad para responder las preguntas, todos van a probar suerte (condiciones 3 y 4)?
- ¿En el examen hemos insistido en que no haya preguntas sin contestar (condición 2)? En la mayoría de exámenes hay omisiones, y más si los errores penalizan.

En este último punto hay una paradoja interesante: por un lado, esta corrección potencia la conjetura (responder todo, no omitir nada), mientras que luego se penaliza los fallos. Sea como sea, los examinados han de escoger una alternativa, no importa cómo la elijan. Dicho de otro modo, se promueve la conjetura para luego penalizarla asumiendo que nadie falla por desconocimiento, sino por «mala suerte» al escoger.

Frente a todo esto, en la práctica real se observan escenarios muy diferentes. Los examinados temen ser penalizados y evitan responder las preguntas muy difíciles. En la mayoría de situaciones, los estudiantes responden las preguntas accesibles para su nivel dejando sin contestar las restantes. Si fallan en alguna respuesta, no es siempre por «mala suerte», sino porque han escogido deliberadamente una respuesta incorrecta pensando que era correcta (han contestado basándose en lo que sabían, aunque lo que sabían no era correcto).

Otro aspecto que tener en cuenta es que este tipo de correcciones no distinguen entre examinados con más o menos nivel (capacidad), mientras que la tendencia a responder por conjetura sí suele variar según este factor. En muchos análisis de respuestas a test reales, los examinados del tercio inferior de puntuaciones (los que tienen menos capacidad sobre los conocimientos evaluados) son los que maximizan las respuestas por conjetura. Tienen menos que perder y más que ganar. Por el contrario, el comportamiento del tercio superior es mucho más prudente. Se trata de los examinados con más conocimiento que se plantean las posibles penalizaciones de manera más conservadora. Con ello se constata un efecto habitual; las correcciones tienden a perjudicar especialmente a los examinados de menor capacidad. A la vista de estos factores aún se hace más difícil admitir que todos los examinados funcionen estadísticamente como un solo individuo modal. Además, muchos exámenes y encuestas mezclan ítems con diferente número de alternativas de respuesta, lo que afecta a la probabilidad de acertar, ya que varía el valor k, por lo que se incumple claramente la condición 1 (por este motivo no es recomendable variar el número de alternativas en los exámenes). Algo parecido sucede cuando k es constante estructuralmente, pero no funcionalmente. Supongamos que un ítem tiene cinco alternativas como todos los del test. Imaginemos que, por algún motivo, el diseño del ítem hace que nadie, o casi nadie, escoja un par de

sus alternativas. La forma como están redactadas o el contenido que tratan carece de atractivo, y los examinados tienen muy claro que estas dos alternativas no pueden ser correctas. Pese a que este ítem estructuralmente tiene cinco alternativas, en realidad solo tres funcionan como tales. Es evidente que esta situación facilita la elección de la respuesta correcta (cuando no se conoce la opción correcta, resulta más fácil acertar entre tres opciones que entre cinco). Cuando esto se repite en varios ítems, el efecto de la conjetura aumenta, y con ello la distorsión de las puntuaciones de los examinados. En muchos exámenes, el valor funcional de *k* es menor que el estructural. Basta comprobar las frecuencias de elección de las alternativas y constatar que el atractivo (tendencia a escogerlas) varía mucho. Hay pruebas con ítems de cuatro alternativas estructurales, pero en la práctica funcionan como ítems de VF, ya que solo dos alternativas de cada pregunta atraen realmente la atención, por lo que quedan descartadas de antemano las otras dos por la mayoría de examinados.

Todo lo anterior puede llevar a reflexionar sobre la aplicación de este tipo de correcciones, puesto que solo son lícitas bajo condiciones muy concretas y controladas. Paradójicamente, su popularidad ha ido por delante del conocimiento de sus fundamentos. Además, existen otras variantes, como las que aceptan la existencia de omisiones, pero también añaden nuevas condiciones difíciles de cumplir. Por su extensión y especificidad no las abordaremos en esta obra.

En general, todas estas expresiones fueron propuestas en la primera mitad del siglo XX pensando en cómo corregir de manera objetiva (justificada analíticamente) el efecto de la conjetura en unas situaciones experimentales muy concretas. Originalmente se empleaban con test denominados de velocidad de cinco o más alternativas de respuesta y longitud superior a 20 ítems. En este tipo de pruebas, la puntuación del examinado viene dada principalmente por su velocidad de respuesta. En estas pruebas, los ítems no suelen ser muy difíciles, puesto que lo importante es comprobar cuántos aciertos se consiguen en un tiempo breve. Este escenario no es generalizable para la mayoría de exámenes actuales, ya sea en papel u online. Por otro lado, si ya en su momento se plantearon como una estimación (corrección) arriesgada de la puntuación del examinado,

más difícil es aún hoy en día, a la vista de las condiciones, aceptar su validez en escenarios complejos.

7.4. Corrección y número de alternativas

Estas fórmulas de corrección también han servido de base para justificar algunos tópicos como el número idóneo de alternativas que debería tener un test.

Evidentemente se reducirá el efecto de la conjetura cuantas más alternativas haya y mejor estén creadas. Sin embargo, esto exige un sobreesfuerzo de diseño que no es sostenible en la mayoría de evaluaciones reales.

Desde hace años han aparecido propuestas defendiendo «números ideales» de alternativas a la hora de crear ítems. Una procede de la fórmula de corrección antes vista y lleva a considerar el «cinco» como el mínimo número de alternativas idóneo y realista. La figura 6 expone esta propuesta. El gráfico muestra cómo sería la puntuación corregida N (eje ordenadas) para el caso de nueve examinados que hubieran obtenido respectivamente 10, 20, 30, 40, 50, 60, 70, 80 y 90 aciertos en total (A) en un examen de 100 preguntas que podría variar de tres a siete alternativas (abscisas).

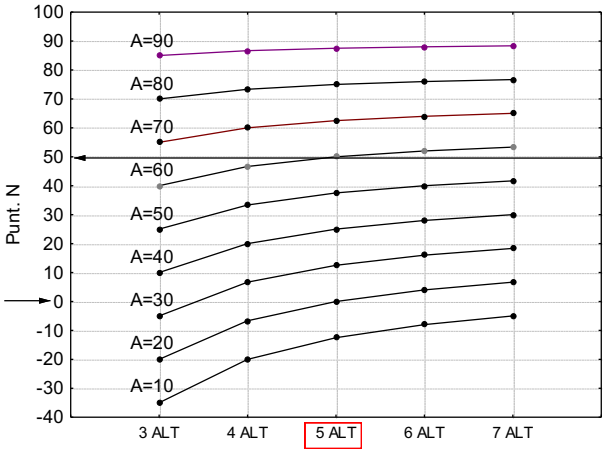


Figura 6. Relación entre la corrección por conjetura aplicada a la nota y número de alternativas de respuesta de los ítems

Cada perfil representa la puntuación corregida que tendría un examinado en función del número de alternativas del supuesto test. La flecha larga horizontal señala la puntuación 50 de corte del examen (apto/no apto).

En la parte inferior se encuentra el perfil del examinado que ha obtenido solo 10 puntos. Se observa que tanto si el test fuera de tres alternativas como de siete, la corrección siempre rebaja N, que por otra parte varía mucho entre tres y siete alternativas y siempre es negativa (la flecha pequeña indica el valor 0 de puntuación corregida).

En el extremo superior se encuentra el perfil del examinado con 90 aciertos (A = 90). En este caso, la variación de la corrección según el número de alternativas es mínima. Se mantiene casi horizontal y parece que el número de alternativas no incide demasiado en la corrección. Para los casos intermedios, la lectura de los perfiles es similar. En general, se observa una tendencia asintótica a medida que el número de alternativas aumenta, pero varía en función del total de aciertos. La variación en función del número de alternativas disminuye a medida que la puntuación total es mayor.

Este tipo de gráficos han llevado a considerar el valor 5, y superiores, como idóneos para contrarrestar la conjetura. De hecho, muchos modelos de hojas de respuesta admiten un máximo de entre cuatro y seis alternativas. Sin embargo, estas consideraciones acerca del número óptimo de alternativas tienen que valorarse bajo el escenario, poco realista, de que todas las alternativas incorrectas tienen características equivalentes (son comparables en calidad). Lo que suele ocurrir en la práctica, sin embargo, es que la calidad de los distractores es diferente, en parte por cómo ha sido el proceso de creación de la pregunta. Generalmente, la persona que redacta una pregunta tiene muy claro cuál es la respuesta correcta y muy probablemente también una de las respuestas incorrectas. No suele costarle mucho encontrar una segunda opción incorrecta, pero la tercera es más costosa (a veces demasiado costosa), y así sucesivamente. Por eso los ítems con tres opciones de respuesta suelen «funcionar» mejor (desde un punto de vista métrico) que los que tienen un número mayor. Obsérvese que este condicionante práctico y frecuente no es muy compatible con la recomendación psicométrica

de aumentar el número de opciones (para aumentar la fiabilidad, para minimizar la conjetura...).

Dos últimos aspectos relacionados con la puntuación del test son su distribución y el punto de corte. En la mayoría de pruebas psicológicas (normativas) interesa que las puntuaciones se distribuyan siguiendo la curva normal. El análisis psicométrico de las cualidades de un test suele asumir esta forma como una condición importante. Por el contrario, los exámenes no responden a este requerimiento. De hecho, interesa que no lo cumplan

La figura 7 representa la distribución esperada para un test de norma de grupo (TNG) y la de un test referido al criterio (TRC) (ya descritos en el apartado 1.2). Si la evaluación de una asignatura se realiza mediante un examen, se espera que la distribución de puntuaciones tienda hacia la parte alta de puntuación. Si el proceso de EA se ha desarrollado convenientemente, el rendimiento de los estudiantes debe mostrar un desvío a la derecha.

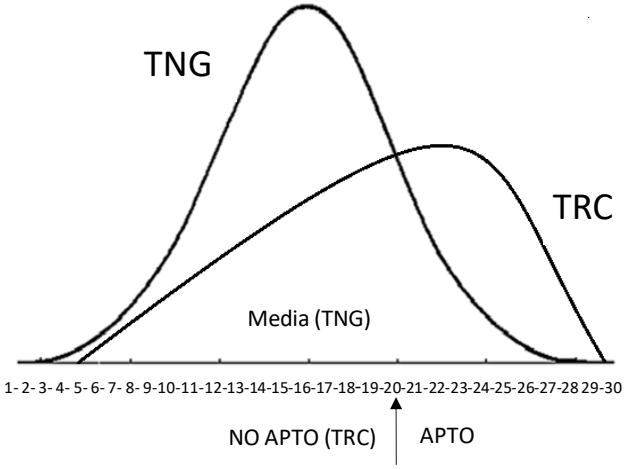


Figura 7. Distribuciones de las puntuaciones de dos test, uno interpretado según una norma de grupo (TNG) y el otro según un criterio (TRC)

Otra cuestión es el punto de corte para la toma de decisiones importantes como apto/no apto, aprobado/suspenso, etc.

Existe la tendencia a confundir la mitad de aciertos con el 5 como puntuación de corte más asociada al aprobado/apto. Al crear una prueba hay que tener previsto cuál va a ser esta puntuación (afecta a la dificultad de los ítems) y si debe corresponder con este 50 % de aciertos o a un valor superior. De hecho, acertar la mitad puede parecer contradictorio con la práctica profesional, y en muchos exámenes y acreditaciones el nivel de exigencia es muy superior (¿qué diríamos de un médico que solo acierta la mitad de los diagnósticos o de un ingeniero que recibe críticas por fallos en la mitad de los dispositivos que diseña?). En muchos entornos, la puntuación de corte oscila alrededor del 70 % de dominio del examen y este dato ha de tenerse en cuenta a la hora de diseñar el examen.

8. AUDITORÍA CUANTITATIVA

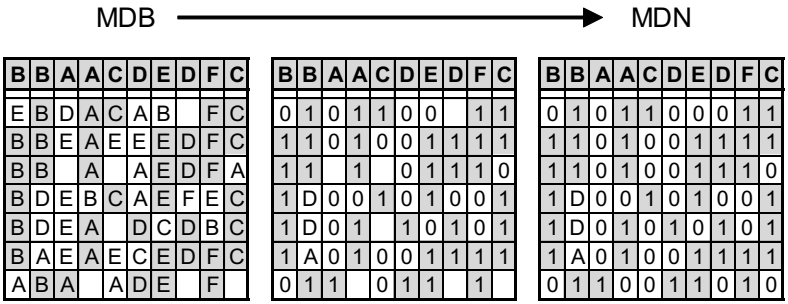
Consiste en el análisis psicométrico del funcionamiento tanto de los ítems como del test en su conjunto. Para ello se obtienen indicadores numéricos y gráficos de cada una de las alternativas de los ítems, incluyendo la omisión como una enésima opción complementaria. También se valora la adecuación de la plantilla y la coherencia de las respuestas. Durante el proceso se comprueba si cada alternativa cumple su función y en qué grado aportan valor al test. Cuando se detecta alguna anomalía en un ítem o en la plantilla, esta se contrasta con el VDI pre y con el autor a fin de considerar si es adecuado o no incluir la pregunta afectada en el cómputo de la puntuación total de los examinados. El proceso de auditoría varía en el nivel de detalle. Algunos modelos de lectoras y LMS ofrecen análisis preliminares de los test que facilitan la primera valoración del material. Para análisis más profundos es necesario un software especializado o acceder a un servicio de análisis de test.

8.1. Datos necesarios

Muchas auditorias quedan limitadas desde un principio por la falta de datos adecuados. Existe la creencia de que los datos que hay que procesar son los datos netos, es decir, las respuestas ya corregidas de los examinados (10101011...). En realidad, esto reduce las opciones de análisis, ya que lo realmente importante son los datos brutos (ABADEAD...). Los datos netos proceden de los brutos tras aplicar la plantilla de corrección, y esto es contradictorio con el principio de auditoría, puesto que, hasta que el análisis demuestre lo contrario, la plantilla se mantiene en cuarentena y es tratada como dudosa.

La figura 8 muestra tres versiones de datos para un mismo test de 10 ítems y seis alternativas. La tabla de la izquierda corresponde a la matriz de datos brutos (MDB) con las respuestas marcadas por los examinados y las omisiones (la fila superior es la plantilla de corrección). La tabla central corresponde a la matriz de datos netos (MDN) una vez corregidas las respuestas con la plantilla y respetando las omisiones. La tercera tabla es la misma que la central, pero convirtiendo estas omi-

siones en 0. Las tres tablas representan las tres situaciones posibles, la primera es indispensable para una auditoría completa y la tercera la menos útil.



B	B	A	A	C	D	E	D	F	C
E	B	D	A	C	A	B		F	C
B	B	E	A	E	E	E	D	F	C
B	B		A		A	E	D	F	A
B	D	E	B	C	A	E	F	E	C
B	D	E	A		D	C	D	B	C
B	A	E	A	E	C	E	D	F	C
A	B	A		A	D	E		F	

B	B	A	A	C	D	E	D	F	C
0	1	0	1	1	0	0		1	1
1	1	0	1	0	0	1	1	1	1
1	1		1		0	1	1	1	0
1	D	0	0	1	0	1	0	0	1
1	D	0	1		1	0	1	0	1
1	A	0	1	0	0	1	1	1	1
0	1	1		0	1	1		1	

B	B	A	A	C	D	E	D	F	C
0	1	0	1	1	0	0	0	1	1
1	1	0	1	0	0	1	1	1	1
1	1	0	1	0	0	1	1	1	0
1	D	0	0	1	0	1	0	0	1
1	D	0	1	0	1	0	1	0	1
1	A	0	1	0	0	1	1	1	1
0	1	1	0	0	1	1	0	1	0

Figura 8. Tres versiones de una matriz de datos para un mismo test de 10 ítems y seis alternativas: una de datos brutos (MDB con respuestas originales, A, B, C...) y dos de datos netos (MDN); una codificando los aciertos (1) y errores (0), y la otra también las omisiones (0)

8.2 Indicadores

Disponiendo de la tabla MDB, es posible obtener una gran variedad de indicadores tanto globales como de los ítems, aunque por su extensión no será posible tratarlos todos aquí en detalle. A continuación, aparece una descripción de los más comunes.

Globales

- Distribución de las puntuaciones: Corresponde al gráfico de la figura 6.
- Coeficiente alfa: Es un tipo de coeficiente de fiabilidad (Alpha de Cronbach) denominado «de consistencia interna», que oscila entre 0 y 1, siendo recomendable aceptar valores superiores a 0,8.
- Error de medida (SEM): Indica la imprecisión de los resultados del test. Como regla orientativa, basta con sumar y restar dos veces el valor de SEM (± 2 SEM) a una puntuación del test (examinado) para estimar el intervalo de error en que esta oscila con un nivel de confianza del 95 %. Cuanto mayor sea este intervalo, menos preciso es el test.

- Comparación entre omisiones y puntuación: Consiste en un gráfico de dispersión que muestra la relación y la correlación entre las puntuaciones totales de los examinados y la cantidad de respuestas que omite cada uno.

De ítem

- Dificultad de ítem: Es la proporción de examinados que seleccionan la alternativa correcta. Oscila entre 0 y 1. Un ítem «fácil» ofrece valores cercanos a 1; uno «difícil» se aproxima a 0.
- Varianza: Es un indicador de dispersión cuyo valor mínimo es 0 y el valor máximo depende del número de alternativas (por ejemplo, para ítems dicotómicos oscila entre 0 y 0,25). La varianza es necesaria para que un ítem discrimine, pero no garantiza que lo haga.
- Discriminación de ítem: Habitualmente se calcula a través de la correlación entre las puntuaciones del ítem con las puntuaciones totales. Puede oscilar entre -1 y +1, pero solo es recomendable aceptar ítems con valores positivos superiores a 0,3. La discriminación informa del funcionamiento del ítem; si muestra un valor negativo, tal vez el ítem esté funcionando a la inversa de los esperado (lo aciertan los examinados menos preparados y lo fallan los más preparados).
- Discriminación corregida: Se calcula de la misma forma que la anterior, pero sin incluir los datos (respuestas) del ítem analizado en la puntuación total. El resultado tiende a dar valores menores que la discriminación sin corregir y es aconsejable en pruebas cortas. A medida que un test tiene más ítems, ambos tipos de discriminaciones tienden a coincidir.
- Elecciones de las alternativas: Es la proporción de examinados que escoge cada una de las alternativas. La dificultad del ítem coincide con la proporción de la alternativa correcta.
- Omisión: Es la proporción de sujetos que no han respondido al ítem.
- Discriminación de las alternativas: Es la discriminación (con o sin corrección) que tendría el ítem considerando que una alternativa errónea fuera cierta. Si el ítem está bien construido, la discriminación de la alternativa correcta debe producir un valor mayor que el resto.
- Discriminación de la omisión: Es la discriminación (con o sin corrección) que tendría el ítem considerando que la omisión fuera la respuesta correcta.

- Igualdad de atractivo: Es una prueba de ajuste entre la proporción de elecciones de cada alternativa errónea con respecto a la proporción esperada en caso de que todas tuvieran un atractivo similar. Si varios ítems muestran desajustes, se invalida la condición 1 del apartado 7.3. Aunque aparentemente estos ítems tengan *k* alternativas, en realidad funcionan con menos y se incrementa el efecto de la conjetura.
- Análisis de la plantilla: Cuando una alternativa incorrecta discrimina más que la correcta puede indicar un conflicto entre alternativas o un error de asignación en la plantilla. En cualquier caso, hay que revisar el contenido del ítem.

Parte de estos indicadores pueden verificarse visualmente a partir de un solo gráfico de perfiles de respuesta de un ítem como el de la figura 9. El gráfico corresponde a un ítem cualquiera de un test de cuatro alternativas (A, B, C y D). El eje de abscisas representa el rango en que oscila la puntuación total del test. En este ejemplo, las puntuaciones se agrupan en diez niveles o intervalos que van de menos a más puntuación o capacidad (N1 a N10). El eje de ordenadas expresa el porcentaje de examinados de cada nivel de puntuación de las abscisas que escogen una determinada alternativa de respuesta. Cada perfil corresponde a una alternativa e indica el porcentaje de examinados de cada nivel que la ha escogido. También se incluye el perfil de la omisión (X).

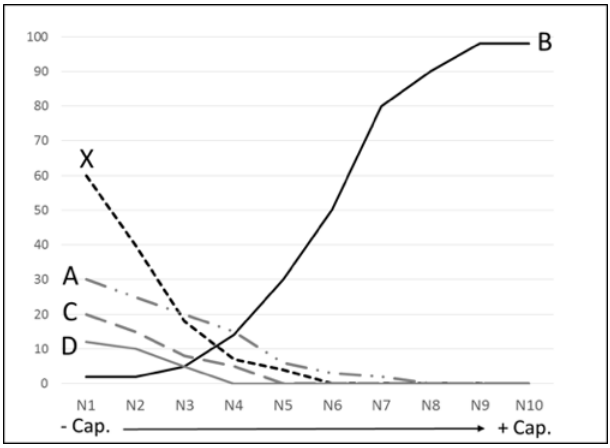


Figura 9. Curvas de respuesta de las alternativas de un ítem. La opción correcta (B) es la creciente

El patrón deseable es que haya un solo perfil creciente, el de la alternativa correcta, y otros decrecientes para el resto de perfiles (respuestas incorrectas y omisión). Dicho de otro modo, la tendencia a escoger la alternativa correcta debe aumentar paulatinamente hacia la derecha a medida que los examinados tienen más capacidad (puntuación total). Por el mismo motivo, la elección de las alternativas incorrectas y de la omisión debe decrecer.

En la figura 9, la alternativa correcta es la B y muestra el patrón esperado. Lo mismo sucede con A, C, D y X (los examinados de menor capacidad son quienes más responden incorrectamente u omiten la respuesta).

Diversas aplicaciones psicométricas ofrecen gráficos similares al de este ejemplo y son muy útiles a la hora de valorar la adecuación de los ítems.

Las directrices básicas de interpretación para el perfil de **la alternativa correcta** son:

- Cuanto más rápidamente crezca el perfil, más discrimina el ítem.
- Si lo hace suavemente a lo largo de todo el eje de abscisas, puede que discrimine, pero en menor grado.
- Si el perfil está desplazado hacia la derecha, se trata de un ítem difícil; si lo está a la izquierda, es un ítem fácil.
- Si el perfil es plano a lo largo del eje de abscisas o decreciente, indicará que el ítem no discrimina o lo hace a la inversa. Es conveniente revisarlo.
- Si la parte izquierda del perfil ya es inicialmente elevada y paralela al eje de abscisas (asintótica) para luego crecer a medida que va hacia la derecha, se trata de un ítem afectado por respuestas conjeturadas (adivinación, azar...).

Para las **alternativas incorrectas**:

- La regla general establece que deben decrecer de izquierda a derecha.
- Si muestra un perfil creciente (como el que correspondería a una correcta) y el de la correcta es decreciente o crece menos que el de la

errónea, es importante revisar la plantilla de corrección para confirmar qué alternativa es realmente la correcta (puede que haya algún error en la plantilla).
- Si todos los perfiles se mantienen planos y estables, puede indicar que la mayoría de respuestas se hacen por conjetura (adivinación, descarte, azar...). En cualquier caso, el ítem no funciona bien y no deberían utilizarse para evaluar a los examinados.

En cuanto a la **omisión**:

- Como regla general, debe mostrar una tendencia decreciente. No obstante, para efectuar una interpretación adecuada hay que considerar si los errores penalizan o no y, si lo hacen, en qué grado.
- Si se mantiene estable (plano) a lo largo de todo el eje de abscisas, indica que los sujetos con más puntuación total también tienden a dejar en blanco esa respuesta. No es un resultado esperado.
- Si el perfil es creciente, hay que revisar el ítem, ya que, contrariamente a lo esperado, los sujetos con más puntuación tienden a omitir la respuesta. Si esto coincide con que el perfil de la especificada como correcta es decreciente, o crece menos que el de omisión, hay que verificar la clave de corrección y confirmar cuál es realmente la alternativa correcta (los sujetos con mayor puntuación tienden a omitir esta pregunta y esto puede evidenciar un error de diseño que solo ellos detectan). Este es uno de los peores casos posibles.

De examinado

Imaginemos dos estudiantes A y B que empatan con 12 aciertos en una prueba de 20 preguntas. A efectos oficiales, se les considera con el mismo nivel de capacidad ya que su puntuación es una simple suma sin entrar a mirar en qué ítems han conseguido los aciertos. No obstante, si ordenamos las 20 preguntas del examen de la más fácil a la más difícil (en función de la proporción de aciertos de todo el grupo examinado) observamos que A ha conseguido sus 12 puntos acertando y fallando ítems fáciles y difíciles indistintamente. Dicho de otro modo, ha acertado preguntas muy difíciles, pero también ha fallado otras fáciles (poco esperado). En cuanto a B, ha conseguido los 12 puntos de manera más coherente, ha acertado los ítems de poca y media dificultad hasta

que ha empezado a fallar u omitir las respuestas y así hasta la pregunta más difícil.

Examinado A: fácil 01011010111010110101 difícil
Examinado B: fácil 11111111111100000000 difícil

A efectos de evaluación, las preguntas aquí son: ¿Ambos examinados tienen realmente la misma capacidad? ¿Están empatados? Observando la serie de respuestas, parece que B tiene un patrón de respuesta más coherente que A; por tanto, su puntuación global debería reflejar de algún modo esta cualidad (quedar mejor ponderada). Si el ejemplo fuera un examen de cálculo, A habría acertado raíces cuadradas y fallado en simples sumas, lo cual no contribuye mucho a saber cuál es su verdadera capacidad o nivel (se pone en duda la validez de las inferencias sobre el nivel de conocimientos que se realizan a partir de las notas obtenidas). Este tipo de reflexiones pone en duda la validez de la puntuación que otorgamos al examinado A, o, mejor dicho, la validez de lo que inferimos a partir de dicha nota (¿habrá contestado al azar? Si es así, ¿la nota que ha conseguido describe correctamente su nivel de capacidad?).

Existen índices psicométricos que permiten detectar patrones incoherentes o atípicos de respuesta (PAR). Pero no pensemos que todos los PAR son debidos a «trampas» en el examen para conseguir más nota. En algunas ocasiones, alumnos de alto nivel contestan mal preguntas muy fáciles (para ellos son tan fáciles que no creen que se puedan preguntar en el examen; por eso «le buscan los tres pies al gato»). O le dedican tanta atención a los aspectos difíciles de la materia que dejan de lado los más sencillos. En estos casos, al contestar de forma incoherente con su nivel, estos estudiantes tendrán una nota más baja de los que los correspondería por su capacidad. La auditoría de los patrones de respuesta nos puede identificar estos y otros casos. Sin embargo, no nos puede explicar sus razones. Es aquí cuando la combinación de los resultados cuantitativos con la información recogida en los VDP puede ser crucial (por ejemplo, es posible que las pautas de respuestas incoherentes de un examinando puedan justificarse por asistencia a clases particulares).

Otro tipo de índices se centra exclusivamente en intentar detectar «proximidad» en las pautas de respuesta, en especial de patrones de

errores similares (PES). Este tipo de índices pretender detectar la probabilidad de copia entre examinandos, pero, como ocurre con el análisis de los PAR, existen otras explicaciones alternativas a los PES. Por ejemplo, es habitual que grupos de estudiantes se reúnan para estudiar la materia del examen, por lo que no sería extraño encontrar patrones similares de aciertos y de errores en diferentes miembros del mismo grupo de estudio. También aquí la información de la VDP es necesaria (por ejemplo, si se ha recogido la localización fila-columna en el aula de los dos examinandos «sospechosos» de copiar el día de la prueba).

En la práctica real aún es poco frecuente el análisis de patrones de respuesta. Al corregir exámenes, la mayoría de docentes no tiende a considerar la pauta u origen de los aciertos, errores y omisiones ni su coherencia. Nuestra experiencia indica que, cuando se realiza, el profesorado está más interesado por los aspectos «negativos» (copia, respuestas al azar) que por los positivos (detectar a alumnos cuya capacidad queda infravalorada en la prueba). En este sentido, destacamos la capacidad formativa que tienen estos instrumentos analíticos (por ejemplo, detectar pautas incorrectas de estudio que pueden ser mejoradas). Todo es cuestión de interés.

ANEXO. CORRECCIÓN DE LA PUNTUACIÓN POR CONJETURA

Una de las expresiones más extendida entre usuarios de test y productos on-line es:

$$N=A-(F/(k-1))$$

Para conocer la puntuación corregida *N* de un examinado solo hay que restar del número obtenido de aciertos *A* el número de fallos *F*, dividido entre el número *n* de alternativas de respuesta de los ítems menos 1. Aparentemente es un cálculo simple, el problema es que a menudo se desconocen las condiciones que asume. Un ejemplo hipotético servirá de base para conocer el origen de esta expresión y sus condiciones de aplicación (en los paréntesis aparecerán citadas las condiciones que se describen más adelante).

Imaginemos un examen de *K* ítems con el mismo número *k* de alternativas de respuesta cada uno (condición 1). Un examinado cualquiera acierta un número *A* de preguntas y falla un número *F*. No existe la posibilidad de omitir (condición 2). De los *A* aciertos conseguidos se asume que una cantidad *N* se debe a su nivel de capacidad y otra cantidad *C* a que ha escogido la respuesta correcta por conjetura (condición 3). Así, estos *C* ítems que ha acertado por conjetura solo son una parte del total de preguntas *C1* en que, supuestamente, ha intentado adivinar la respuesta (condición 4). De este modo podemos desglosar la puntuación del examinado en los siguientes componentes:

K: número de ítems del examen

k: número de alternativas de respuesta de cada pregunta

A: número total de aciertos

F: número total de ítems fallados

N: número de ítems acertados por el propio nivel de capacidad y sin conjeturar

C1: número de ítems en que el examinado ha intentado conjeturar

C: número de ítems acertados por conjetura

El total de *K* ítems del examen es la suma de los aciertos y fallos.

$$K = A + F$$

A su vez, el total de aciertos es la suma de aciertos sin conjeturar N y aciertos gracias a la conjetura C.

$$A = N + C$$

De esto se deduce que los aciertos N fruto de la capacidad del examinado son:

$$N = A - C$$

En cuanto a los aciertos conseguidos conjeturando (C), podemos asumir que son solo una parte de aquellos en que lo ha intentado (C1). Como cada ítem tiene la misma cantidad *k* de alternativas, el valor *C* será la siguiente fracción de *C1* (condición 5).

$$C = C1/k$$

Por otro lado, el total de fallos F cometidos será la diferencia entre el número de ítems en que el examinado ha intentado conjeturar (C1) y la cantidad en que ha conseguido acertar conjeturando (C).

$$F = C1 - C$$

Como hemos visto antes, *C* también puede expresarse como una fracción de modo que

$$F = C1 - (C1/k)$$

Lo que equivale a:

$$C1 = (k*F)/(k-1)$$

Y, aceptando antes que $C = C1/k$ y substituyendo ahora *C1* queda como:

$$C=(k*F) / [(k*(k-1))]$$

Donde, simplificando k, queda que:

$$C = F/(k-1)$$

Volviendo a la expresión inicial referida a los aciertos debidos al propio nivel de capacidad N, podemos ahora tomar la expresión:

$$N = A - C$$

Y substituyendo *C*, reencontrar la fórmula presentada al principio.

$$N=A-(F /(k-1))$$

GLOSARIO

Administrar un test: aplicar la prueba a uno o más individuos.

Alternativa correcta: letra o número que identifica la respuesta que puntúa.

Alternativa múltiple (AM): ítem formado por un enunciado y unas opciones de respuesta donde hay que seleccionar la respuesta correcta o la mejor posible.

Análisis (en Bloom): capacidad de subdividir la información recibida.

Análisis de alternativas incorrectas: discriminación del ítem en el caso de que cada uno de los distractores fuera correcto.

Análisis de los ítems: proceso de examen de los indicadores los ítems del test.

Aplicación (en Bloom): capacidad de abordar situaciones o resolver problemas nuevos utilizando principios y reglas previamente aprendidos.

Auditoría de test: control de la calidad del test. Es el «test del test».

Auditoria cualitativa: revisión de los aspectos formales y de diseño de la prueba.

Auditoria cuantitativa: obtención de indicadores numéricos y gráficos que informan del funcionamiento del test a partir de las respuestas recibidas.

Banco de ítems (BI): colección o biblioteca de ítems de formato y contenido estandarizado

Clave o plantilla de corrección: código de aciertos, fallos y puntuación que se otorga a las respuestas de los ítems.

Coeficiente de fiabilidad: indicador basado en la equivalencia o consistencia de medidas de un mismo grupo de examinados (expresado como correlación).

Coherencia de respuesta: patrón de respuesta de un examinado ajustado a lo esperado.

Comprensión (en Bloom): capacidad de captar el significado o sentido directo de la información presentada.

Conjetura: recursos que ayudan a acertar un ítem cuando se desconoce la respuesta.

Conocimiento (en Bloom): capacidad de recordar términos, principios, normas, etc.

Consigna: instrucciones verbales o escritas para responder una prueba.

Corrección de la conjetura: descuento de la puntuación del test supuestamente debido a la conjetura.

Curva de ítem: gráfica que relaciona la capacidad de un examinado con la probabilidad de acertar un ítem.

Curva de omisión: gráfica que relaciona la capacidad de un examinado con la probabilidad que omita un ítem.

Curva normal: modelo matemático que relaciona la desviación típica de las puntuaciones con la proporción de casos, o área de la curva.

Dificultad de alternativa de ítem: proporción en que una alternativa es escogida.

Dificultad de ítem: proporción de examinados que escoge la alternativa correcta.

Discriminación: capacidad del ítem para diferenciar entre examinados de diferente capacidad.

Discriminación de alternativa: discriminación considerando que una alternativa errónea de un ítem fuera la cierta.

Discriminación negativa: quienes aciertan el ítem son los examinados con menor puntuación total y quienes lo fallan los de mejor puntuación total.

Enunciado: premisa, núcleo o parte introductoria del ítem que plantea al examinado la tarea a desempeñar.

Error estándar de medida (SEM): indicador de la precisión de las puntuaciones de un test.

Igualdad de atractivo de las alternativas: distribución homogénea de las elecciones erróneas entre las alternativas incorrectas.

Independencia local (IL): cuando solo la capacidad del examinado determina su respuesta a los ítems.

Ítem: elemento de un test. Una pregunta es un tipo de ítem pero un ítem no siempre es una pregunta.

Ítem abierto: el examinado elabora la respuesta

Ítem cerrado: el examinado elige la respuesta entre varias opciones.

***Learning managment system* (LMS)**: sistema online de gestión del aprendizaje.

Matriz de datos brutos (MDB): letras o números de las alternativas marcadas por cada examinado (AABDADCCA...).

Matriz de datos netos (MDN): proceden de los datos brutos tras comparar o corregir con la plantilla o clave de corrección (01101001...).

Longitud del test (L): número de ítems del test.

Patrones atípicos de respuesta (PAR): pautas de respuesta que no se corresponden con el modelo psicométrico de la prueba.

Patrones de error similares (PES) en parejas o grupos de examinados con un patrón de errores muy similar especialmente en los ítems difíciles.

Quiz: examen o prueba de ensayo y autoevaluación habitual en entornos LMS.

Reglas de generación de ítems (RGI): directrices para crear ítems adecuadamente.

Sesgo: para grupos homogéneos de examinados, un mismo ítem o test da puntuaciones diferentes en función de una característica ajena al objetivo del test (cultura, raza, nivel social, etc.).

Síntesis (en Bloom): capacidad de reunir elementos/partes para formar un todo.

Solapamiento: una alternativa errónea hace el papel de la alternativa especificada como correcta.

Tabla de especificación de objetivos (TEO): estructura que relaciona los contenidos a evaluar con la forma como se evaluaran.

Teoría clásica de los test (TCT): teoría psicométrica basada en el concepto de que la puntuación que obtiene una persona en una prueba es el resultado de sumar a su puntuación verdadera una puntuación de debido al error de medida. Permite cuantificar dicho error y por tanto estimar la fiabilidad.

Teoría de respuesta al ítem (TRI): conjunto de principios psicométricos y modelos matemáticos que relacionan la capacidad de los examinados con la probabilidad de obtener determinada puntuación en los ítems.

Test adaptativo informatizado (TAI): test personalizado administrado por ordenador. Se basa en algoritmos de presentación de ítems que de-

ciden, para cada persona y en cada paso de la prueba, el ítem óptimo para evaluar el nivel de conocimientos.

Test de norma de grupo (TNG): interpretan la puntuación de un individuo en relación a la ejecución global del grupo al que pertenece.

Test referidos al criterio (TRC): producen medidas directamente interpretables independientes a los resultados del grupo.

Validez: cualidad del test que informa de si mide lo que pretende.

Validez de contenido: basada en la representatividad de los ítems.

Vector descriptor de ítem (VDI): unión de descriptores codificados de las características de los ítems

Vector descriptor de persona (VDP): unión de descriptores codificados de características relevantes de las personas examinadas antes (pre) y después del examen (post).

Verdadero-falso (VF): formato de ítem en que solo hay que valorar si lo que plantea es correcto o no.

BIBLIOGRAFÍA

AENOR (2015). *Norma ISO 10667 para la evaluación de personas en entornos laborales*. Madrid: AENOR.

American Educational Research Association (AERA); American Psychological Association (APA); National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, L. W.; Krathwohl, D. R.; Airasian, P. W.; Cruikshank, K. A. (2001). *A taxonomy for learning, teaching and assessing. A revision of Bloom's taxonomy of educational objectives*. Londres: Addison-Wesley Longman.

Bloom, B. S. et al. (1956). *The taxonomy of educational objectives, handbook I: the cognitive domain*. Nueva York: David McKay.

Cizek, G. J.; Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. Nueva York: Routledge.

Covacevic, C. (2014). *Cómo seleccionar un instrumento para evaluar aprendizajes estudiantiles*. Washington: Banco Interamericano de Desarrollo.

Doval, E.; Renom, J. (2007). *Formatos de ítems en los exámenes universitarios*. Comunicación presentada en el XI Congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga.

—— (2009a). *Nuevos usos de los formatos de respuesta de selección en la evaluación diagnóstica y formativa*. Comunicación presentada en el VI Congreso Internacional de Docencia Universitaria e Innovación. Barcelona.

—— (2009b). *Los formatos de respuesta de elección múltiple y alternativas frente al reto evaluativo del Espacio Europeo de Educación Superior*. Comunicación presentada en el XI Congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga.

Doval, E. *et al.* (2015). *Las puntuaciones obtenidas en los test de conocimientos: ¿son siempre indicadores válidos del aprendizaje?* XXII Congreso Internacional Educación y Aprendizaje. Madrid.

Downing, S. M.; Haladyna, T. M. (2011). *Handbook of test development*. Mahwah, Nueva Jersey: Lawrance Erlbaum.

Elosua, P. (2003). «Sobre la validez de los test». *Psicothema*, 15 (2): 315-321.

Gómez, J; Hidalgo, M. D.; Guilera, G. (2010). «El sesgo de los instrumentos de medición. Test justos». *Papeles del Psicólogo*, 1 (1): 75-84.

Haladyna, T. M.; Downing, S. M. (1989). «The validity of a taxonomy of multiple-choice test item». *Applied Measurement in Education*, 1 (1): 51-78.

Haladyna, T. M.; Rodríguez, M. C. (2013). *Developing and validating test items*. Nueva York: Routledge.

Haladyna, T. M.; Downing, S. M.; Rodríguez, M.C. (2002). «A review of multiple-choice item-writing guidelines». *Applied Measurement in Education*, 15 (3): 309-334.

Hanna, L. S.; Michaelis. J. U. (1977). *A comprehensive framework for instructional objectives: a guide to systematic planning and evaluation*. Reading, MA: Addison-Wesley.

International Test Commission (2013). *International guidelines on quality control in scoring, test analysis, and reporting of test scores*. Disponible en: <www.intestcom.org>. Disponible en castellano en: <www.cop.es/pdf/ITC2015-Directrices-Control-Calidad.pdf>.

Lane, S.; Raymond, M. R.; Haladyna, T. M. (2016). *Hondbook of test development*. Nueva York: Routledge.

Marrelli, A. F. (1995). «Writing multiple-choice test items». *Performance and Instruction*, 34 (8): 24-29.

Martínez, R.; Muñiz, J. (2011). «Calidad de los ítems de los exámenes PIR». *Papeles del Psicólogo*, 32 (3): 254-264.

MIT (2013). «La ley de Benford y el arte de tener éxito en exámenes tipo test». *MIT Technology Review*. Disponible en: <https://www.technologyreview.es/s/7143/la-ley-de-benford-y-el-arte-de-tener-exito-en-examenes-tipo-test>.

Moreno, R; Martínez, J.; Muñiz, J. (2004). «Directrices para la construcción de ítems de elección múltiple». *Psicothema*, 16 (3): 490-497.

—— (2006). «New guidelines for developing multiple-choice items». *Methodology. European Journal of Research Methods for the Behavioral and Social Sciences*, 2 (2): 65-72.

Muñiz, J; Hernández, A.; Ponsoda, V. (2015). «Nuevas directrices sobre el uso de los test: investigación, control de calidad y seguridad». *Papeles del Psicólogo*, 36 (3): 161-173.

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

Prieto, G.; Muñiz, J. (2000). «Un modelo para evaluar la calidad de los test utilizados en España». *Papeles del Psicólogo*, 77: 65-75.

Renom, J. (1992). *Diseño de test*. Barcelona: IDEA I+D.

—— (1994). *Test adaptativos computerizados: fundamentos y aplicaciones*. Barcelona: Edicions UB.

—— (2002). *Metrix Engine UB: analizador de test y cuestionarios*. Barcelona: Edicions UB.

—— (2011). *Servicios de test universitarios*. XII Congreso de la Asociación Española de Metodología de las Ciencias del Comportamiento. San Sebastián.

—— (2013). «La auditoría de test». *Revista PSIARA COPC*. Disponible en: <http://www.psiara.cat/view_article.asp?id=4329>.

Renom, J.; Doval, E. (1999). *Test adaptativos informatizados: estructura y desarrollo*. En: Olea, J.; Ponsoda, V.; Prieto, G. (eds.). *Test adaptativos informatizados*. Madrid: Pirámide.

Renom, J; Solanas, A; Doval, E.; Núñez. M. (2001). *SEDI: sistema experto para el diagnóstico de ítems*. Comunicación presentada en el VII Congreso de Metodología de las Ciencias Sociales y de la Salud, Madrid.

—— (2002). *Piert: tutorial multimedia para el diseño de pruebas de rendimiento (versión profesional con herramientas)*. Barcelona: Edicions UB.

Renom, J. *et al*. (2014). «Proyecto UB-AUDIT: *plugin* para el análisis e informe de calidad de cuestionarios Moodle». *Revista del CIDUI*, 2. Disponible en: <https://www.cidui.org/revistacidui/index.php/cidui/article/view/538>.

Riba, M; Doval, E.; Fauquet, J. (2016). «Pruebas tipo test como instrumentos de evaluación diagnóstica y formativa». *Revista del Congreso Internacional de Docència Universitària i Innovació (CIDUI)*, 3. Disponible en: <https://www.cidui.org/revistacidui/index.php/cidui/article/view/968>.

Sans, A. (2008). *La evaluación de los aprendizajes: construcción de instrumentos*. Barcelona: Octaedro.

Williams, R. G.; Haladyna, T. (1982). «Logical operations for operating intended questions (LOGIC): a typology for higher level test items». En: Roid, G. H.; Haladyna, Y. T. (eds.). *A technology for test item writing*. Nueva York: Academic Press.

NORMAS PARA LOS COLABORADORES

http://www.ub.edu/ice/sites/default/files/docs/normas_pres.pdf

EXTENSIÓN

Las propuestas de cada cuaderno no podrán exceder **la extensión de 50 páginas (en Word)** salvo excepciones, unos 105.000 caracteres; espacios, referencias, cuadros, gráficas y notas, inclusive.

PRESENTACIÓN DE ORIGINALES

Los textos han de incluir, en formato electrónico, un **resumen** de unas diez líneas y tres palabras clave, no incluidas en el título. Igualmente han de contener el **título**, un ***abstract*** y tres ***keywords*** en inglés.

Respecto a la **manera de citar y a las referencias bibliográficas**, se han de remitir a las utilizadas en este cuaderno.

EVALUACIÓN

La aceptación de originales se rige por el **sistema de evaluación externa por pares**.

Los originales son leídos, en primer lugar, por el **Consejo de Redacción**, que valora la adecuación del texto a las líneas y objetivos de los cuadernos y si cumple los requisitos formales y el contenido científico exigido.

Los originales se someten, en segundo lugar, a la **evaluación de dos expertos** del ámbito disciplinar correspondiente, especialistas en la temática del original. Los autores reciben los comentarios y sugerencias de los evaluadores y la valoración final con las correcciones y cambios oportunos que se han de aplicar antes de ser aceptada su publicación.

Si los cambios exigidos son significativos o afectan a buena parte del texto, el nuevo original se somete a evaluación de dos expertos externos y de un miembro del Consejo de Redacción. El proceso se lleva a cabo como «doble ciego».

Revisores

http://www.ub.edu/ice/llibres/eduuni/Revisores_Octaedro.pdf

BUENAS PRÁCTICAS EN EL USO DE PRUEBAS DE ALTERNATIVA MÚLTIPLE

JORDI RENOM PINSACH
EDUARDO DOVAL DIÉGUEZ