Diseño y análisis de pruebas de múltiple opción

María Noel Rodríguez Ayán

2018

UNADEQ

UDELAR FQ FACULTAD DE QUIMICA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

CONTENIDO

MÓDULO I. DISEÑO

UNIDAD 1: FORMULACIÓN Y VALIDACIÓN DE LOS ÍTEMS

Un ítem de múltiple opción está compuesto por un enunciado o raíz, donde se plantea el problema, y una serie de alternativas de respuesta. La pregunta es: ¿cómo saber si una prueba de múltiple opción es adecuada para medir el nivel de conocimiento de los estudiantes? Haladyna, Dowing y Rodriguez (2002) sistematizaron 31 directrices sobre la redacción de los ítems, las cuales hemos traducido y presentamos en este módulo, complementadas con los aportes de otros autores. También recomendamos la lectura de Haladyna (2006), donde describe la problemática de la validez de los tests estandarizados.

Downing (2005) comparó los resultados de aplicar pruebas de conocimiento empleando ítems que no cumplían con alguna de las 31 especificaciones e ítems estándar -aquellos que cumplían con todas las especificaciones- diseñados para evaluar los mismos contenidos. Obtuvo una tasa de falsos negativos (estudiantes incorrectamente calificados como reprobados al emplear los ítems no estándar) de 10-15%. Su conclusión es que el empleo de tales ítems produce una varianza de resultados irrelevante a los fines de la prueba y en consecuencia recomienda que se hagan los esfuerzos necesarios para disponer de preguntas que cumplan con los criterios especificados.

A continuación presentamos algunas pautas generales para la redacción de ítems de múltiple opción, adaptadas de Haladyna y otros (2002): el contenido, el formato, el estilo, redacción del enunciado, redacción de las alternativas, ¿cuántas opciones por ítem?, *ninguna de las anteriores es correcta* y *todas las anteriores son correctas.*

EL CONTENIDO DE LAS PREGUNTAS MÚLTIPLE OPCIÓN

A continuación se presentan una serie de puntos que deben ser tomados en cuenta a la hora de pensar el contenido de un ítem múltiple opción:

1. cada ítem debe reflejar un contenido específico (hecho, concepto, procedimiento, etc).
2. el contenido evaluado en cada ítem debe ser relevante para el aprendizaje; debe evitarse los contenidos triviales.
3. no emplear contenidos literales de textos o materiales didácticos. De esta manera se evita evaluar aprendizajes meramente memorísticos.
4. evitar los ítems concatenados. El contenido de cada ítem debe ser independiente de los ítems restantes.
5. evitar los contenidos excesivamente específicos, así como los excesivamente generales.
6. evitar preguntar sobre contenidos que puedan ser susceptibles de opinión.

7. evitar preguntas que puedan inducir a error, denominadas comúnmente preguntas “engañosas” o “tramposas” (trick items).
8. emplear un vocabulario apropiado para la población que se desea evaluar.

Respecto a la definición de "pregunta tramposa", Haladyna y Downing (1989a) definen a estos ítems como aquellos que, de alguna manera, inducen a error. Cabe preguntarse si ello es accidental o deliberado. Para algunos autores un ítem engañoso es el que ha sido redactado con la intención deliberada de inducir a error. Puede tratarse de preguntas dirigidas a identificar al estudiante que no está completamente alerta, en lugar de a evaluar un conocimiento específico (Ebel, 1979) o preguntas con contenidos triviales, ambiguos o confusos (Hopkins, Stanley y Hopkins, 1990). También el uso de términos que parecen irrelevantes para el contenido del ítem pero que en realidad resultan ser cruciales induce a cometer errores (Thorndike, Cunningham, Thorndike y Hagen, 1991).

Entre los antecedentes solamente hemos encontrado un estudio empírico sobre los ítems tramposos: Roberts (1993). Este autor investigó las concepciones de los estudiantes y docentes universitarios sobre la definición y las características más importantes de las preguntas tramposas, encontrando tres definiciones mayoritarias:

- la intencionalidad del autor de las preguntas
- la ambigüedad del contenido
- un nivel de precisión entre las alternativas más fino que el nivel empleado durante la instrucción.

Estas categorías no son independientes. Por ejemplo, un contenido ambiguo puede provenir de una intención deliberada o de una redacción pobre por falta de experiencia del instructor, etc. Los resultados de Roberts también sugieren que el rendimiento de los estudiantes en ítems engañosos (contenidos triviales, ambiguos, etc.) es en general más pobre que en ítems no engañosos paralelos, diseñados para evaluar el mismo contenido.

EL FORMATO DE LAS PREGUNTAS DE MÚLTIPLE OPCIÓN

1. Evitar el empleo de los denominados "*ítems K*".

Estos son ítems complejos, que contienen una lista de posibles alternativas de respuesta (*alternativas primarias*) y una lista de combinaciones de las respuestas primarias (*alternativas secundarias*). El estudiante debe seleccionar la respuesta de la lista de alternativas secundarias.

Ejemplo de un ítem tipo K (extraído de Suen y McClellan, 2003):

De los siguientes alimentos:

1. Costillas de cerdo

2. Manzanas
3. Chauchas
4. Camarones

(*respuestas primarias, que forman parte del enunciado*)

¿cuál/es debería/n restringirse en una dieta dirigida a personas con problemas de colesterol?

A. 1, 2 y 3
B. Sólo 4
C. 2 y 3
D. 1 y 4

(*respuestas secundarias, de las cuales el estudiante debe seleccionar una*).

El empleo de este tipo de ítems suele tener efectos en las propiedades métricas del instrumento aplicado a la población, tales como disminución del índice de dificultad y del índice de discriminación de los ítems (Albanese, 1993; Dawson-Sanders, Nungester y Downing, 1989; Rodriguez, 1997). La disminución del índice de dificultad puede ser valorada como algo bueno o malo dependiendo del grado de dificultad de la prueba deseado por el autor de la misma. Pero dado el efecto observado sobre el índice de discriminación se desaconseja el uso de estos ítems complejos.

2. Se sugiere ubicar los ítems en formato vertical y no horizontal a fin de ofrecer una visión más compacta del enunciado y de sus alternativas de respuesta.

EL ESTILO DE LOS ÍTEMS

En el estilo de las preguntas múltiple opción se deben tener en cuenta los siguientes puntos:

1. cuando corresponda, se sugiere revisión editorial de las preguntas
2. cuidar errores gramaticales, de puntuación y de ortografía
3. evitar abreviaturas
4. minimizar el tiempo de lectura de cada ítem

REDACCIÓN DEL ENUNCIADO

1. El enunciado debe contener directivas bien claras.

2. Debe contener la idea central objeto de evaluación, así como la información necesaria para responder.

Deben evitarse los denominados "unfocused stems", que son enunciados sin dirección, que no plantean una idea concreta a resolver. Ello da como resultado ítems que requieren del análisis de alternativas heterogéneas, cada una del tipo falso / verdadero, en lugar de alternativa de respuesta para una única pregunta planteada (Downing, Dawson-Sanders, Case y Powell, 1991). ¿Es correcto decir que:

a) la hormona de crecimiento induce la producción de IGFBP3?
b) la proteína de unión al factor de crecimiento similar a la insulina predominante en el suero humano es IGFBP3?
c) múltiples formas de IGFBP derivan de un único gen?

3. Evitar información innecesaria o "window dressing".

4. Siempre que sea posible se debe redactar el enunciado en forma afirmativa.

El uso de frases negativas suele tener un impacto contraproducente, tanto en la comprensión como en el rendimiento (véanse Haladyna y Downing, 1989a, 1989b, para sendas revisiones). En caso de que se empleen formas negativas se sugiere que estas aparezcan destacadas, en negrita y mayúscula, para alertar al lector.

REDACCIÓN DE LAS ALTERNATIVAS

1. Asegurarse de que sólo una de las alternativas sea correcta.
2. Alternar la ubicación de la alternativa correcta en función del número de alternativas. Si se la ubica siempre en la misma posición se puede estar favoreciendo a quienes responden de acuerdo a criterios formales.
3. Cuando corresponda, las alternativas deben ubicarse en orden lógico o numérico.
4. Las alternativas no deben solaparse, deben ser independientes.
5. Todas las alternativas deben tener una estructura gramatical similar, a fin de evitar que se pueda distinguir la alternativa correcta de los distractores siguiendo criterios no técnicos.
6. Las alternativas deben tener todas una longitud similar, por las mismas razones expuestas en el numeral 5.
7. Enunciar las alternativas en forma afirmativa, evitando las negaciones.
8. Evitar el empleo de términos que puedan orientar en la identificación de la alternativa correcta o en la eliminación de distractores.

    Ejemplos de tales términos: determinantes específicos (siempre, nunca, completamente, absolutamente, etc) pues pocas veces se pueden hacer juicios universales y la presencia de estos términos puede ayudar a descartarlos como alternativa correcta; alternativas idénticas o con contenidos muy similares a los del enunciado; inconsistencias gramaticales; alternativas absurdas; etc.

9. Los distractores deben ser tales que puedan parecer aceptables para los estudiantes que no saben tanto; pero deben poder ser descartados por los estudiantes con un alto nivel de conocimiento.
10. En la construcción de los distractores se recomienda hacer uso de los errores típicos cometidos por los estudiantes. Se trata de incluir contenidos que sean verdaderos, pero que no respondan a la pregunta formulada en el enunciado.
11. Evitar el uso de alternativas que apelan al humor, a menos que ello sea compatible con el entorno de aprendizaje. Se desaconseja su uso en entornos formales de evaluación.

ALTERNATIVAS POR ÍTEM

El número de alternativas por ítem puede ser variado, pero actualmente la investigación sugiere que el número más adecuado sería 3.

Rodríguez (2005) condujo un metaanálisis entre publicaciones donde se hubiera evaluado el efecto de variar el número de alternativas (entre 2 y 5 por ítem) sobre las propiedades métricas (ÍNDICE DE DIFICULTAD y de ÍNDICE DE DISCRIMINACIÓN y Fiabilidad del test en su conjunto). Sus resultados sugieren que 3 opciones por ítem parece ser un valor óptimo en la mayoría de los casos: una alternativa correcta y dos distractores plausibles. El empleo de 4 ó 5 alternativas no parece mejorar las propiedades métricas y suele redundar en la inclusión de distractores poco probables.

Por otro lado, el número de opciones está vinculado al tiempo de duración de la prueba, por lo que emplear ítems con 3 opciones permite incrementar el número total de ítems a administrar, lo que constituye una mejora potencial de la cobertura de los contenidos a evaluar.

NINGUNA DE LAS ANTERIORES ES CORRECTA

La alternativa *Ninguna de las anteriores es correcta* debe emplearse con cautela.

Algunos autores consideran que esta alternativa es controvertida porque comúnmente se desaconseja su uso sin evidencia empírica que sustente tal recomendación (Knowles y Welch, 1992). Sin embargo, la revisión de Haladyna y Downing (1989b) muestra que su uso en general disminuye los índices de dificultad y de discriminación del ítem, así como la Fiabilidad de la prueba en su conjunto.
En concordancia con lo anterior, el estudio empírico conducido por Frary (1991) revela que el uso de esta alternativa puede ser efectivo y comparable al de cualquier otro distractor cuando se trata de ítems que, en ausencia de esta opción, resultarían demasiado fáciles (índices de dificultad superiores a 0,6). Tal es el caso de ítems ante los cuales los estudiantes no podrían elaborar la respuesta si se tratara de preguntas

abiertas, pero eventualmente pueden reconocer la forma de completar correctamente el enunciado y así seleccionarlo.

La presencia de *Ninguna de las anteriores* inhibiría estas conductas. Además de que, también motivaría al estudiante a analizar más críticamente las restantes alternativas, en lugar de solamente seleccionar la que le parezca más plausible. Para ítems que, en ausencia de esta opción, presenten niveles de dificultad más elevados Frary no recomienda su inclusión.

La inclusión de *Ninguna de las anteriores* como alternativa correcta abre la posibilidad de premiar una respuesta incorrecta (Gross, 1994). Por ejemplo, ante la pregunta *¿Cuál es la capital del país….?* Un estudiante puede seleccionar *Ninguna de las anteriores* o bien porque sabe cuál es la capital y ésta no figura en la lista o bien porque cree -erróneamente- saber cuál es la capital y esta no figura en la lista. En cualquier caso obtendrá el crédito correspondiente.

Por todo lo anterior sugerimos limitar el uso de esta alternativa.

TODAS LAS ANTERIORES SON CORRECTAS

Evitar la alternativa *Todas las anteriores son correctas*.

El uso de esta alternativa también es controvertido. La revisión de Haladyna y Downing (1998a) muestra que puede aumentar el índice de dificultad y disminuir el de discriminación.

Mueller (1975) analizó los efectos de esta opción, controlando por el efecto de *Ninguna de las anteriores*, y sus resultados indican que cuando *Todas las anteriores* es la opción correcta el ítem resulta extremadamente fácil. En cambio como distractor puede tener un funcionamiento superior al de otros distractores, de carácter sustantivo.

Frary (1991) desaconseja el uso de esta alternativa en cualquier caso. El reconocer un distractor como respuesta incorrecta sería suficiente como para descartar *Todas las anteriores son correctas*. Por otro lado, la identificación de dos respuestas correctas sería suficiente como para seleccionarla. Así, el uso de *Todas las anteriores* puede beneficiar a estudiantes que, teniendo el mismo nivel de conocimiento de sus pares respecto al objeto de evaluación, tengan un nivel superior de desarrollo de sus habilidades lógicas, lo que limitaría la validez de la prueba como instrumento idóneo para evaluar los conocimientos pretendidos.

MÓDULO II. ANÁLISIS

UNIDAD 2. ANÁLISIS FORMAL DE LAS RESPUESTAS

Seguramente todos hemos oído hablar, en mayor o menor grado, de los índices de dificultad y de discriminación de los ítems de una prueba de múltiple opción.

Pero ¿qué son esos índices? ¿De qué dependen? ¿Qué tanto discriminan los ítems entre los estudiantes que "saben" y los que "no saben"? Un mismo ítem ¿puede tener distintos valores de índice de discrminación (o dificultad)?

Y cuando decimos índice de dificultad ... ¿dificultad para quién? ... Y el puntaje total de la prueba ¿qué representa?

Para tratar de dar respuesta a estas preguntas lo invitamos a consultar las secciones que siguen: *Índice de discriminación, Índice de dificultad, Adecuación de los distractores, ¿Se suman los puntos de los ítems?, ¿Se restan puntos por errores? y ¿Cuál es el mejor punto de corte?*

ÍNDICE DE DISCRIMINACIÓN

El índice de discriminación que aquí se presenta es el definido en la Teoría Clásica de los Tests (TCT). Es muy importante tenerlo presente, dado que cuando el marco de trabajo es la Teoría de Respuesta al Ítem (TRI), si bien se emplea la misma denominación, su definición e interpretación son muy distintas. En el marco de la TCT el índice de discriminación depende de la población en que se aplica la prueba. Por lo tanto no es una propiedad del ítem, sino una resultante de la interacción ítem-sujetos. Ésta es una gran diferencia con el índice que se define según la TRI, el cual resulta independiente de las personas y es una verdadera propiedad del ítem. Puede consultarse el artículo de Hambleton y Jones (1993) para una comparación entre ambas teorías.

Hay distintas aproximaciones clásicas para obtener una medida de este indicador. Por un lado está el índice de discriminación basado en correlaciones. Este es el coeficiente de correlación de Pearson entre los puntajes de los sujetos en el ítem y en la prueba. Es una medida de qué tanto discrimina el ítem entre aquellos estudiantes que obtienen puntajes altos y bajos en la prueba. Hay que tener en cuenta que el índice así definido resulta ser una medida "contaminada", pues el puntaje total resulta de la suma de los puntajes parciales. Por tal razón se acostumbra estimar también el índice de correlación ítem-total corregido. Este no es más que el coeficiente de correlación de Pearson entre el puntaje en el ítem y el puntaje en los restantes ítems de la misma dimensión.

Existen diferentes fórmulas para el coeficiente de correlación de Pearson, según la naturaleza de las variables que se desean correlacionar (ítem y prueba). Estas variables

pueden ser dicotómicas, politómicas o dicotomizadas (variables continuas que se dicotomizan según algún criterio). El índice de discriminación que resulta de la aplicación directa de la correlación de Pearson entre las dos variables (ítem y prueba) recibe denominaciones diferentes según el caso. Cuando el resultado de la prueba es una variable continua (por ejemplo, la suma de los puntajes parciales de cada ítem) el coeficiente de correlación se denomina *biserial-puntual*. Cuando el resultado de la prueba es una variable dicotómica (por ejemplo, aprueba o pierde) el coeficiente se denomina *Phi*. Pueden consultar el desarrollo para la estimación del coeficiente de correlación de Pearson, así como un ejemplo de aplicación que facilitará su comprensión, en el material preparado por Andrés Abella (2014) para este curso.

Coeficiente de correlación biserial-puntual entre la variable X (puntaje en la prueba) y la variable Y (ítem dicotómico):

$$\rho_{X,Y} = \frac{\mu_P - \mu_X}{\sigma_X}\sqrt{\frac{p}{q}}$$

μ_p es la media del puntaje total X entre los sujetos que aciertan el ítem

μ_x es la media del puntaje total X entre todos los sujetos

σ_X es la desviación estándar del puntaje total X

p es la proporción de sujetos que aciertan el ítem

$q = (1 - p)$

Respecto a los valores del coeficiente *biserial-puntual*, puesto que corresponden al caso en que una de las variables es dicotómica, muchos autores afirman que no puede alcanzar el valor máximo teórico de 1, que es el máximo que puede adoptar el coeficiente de correlación de Pearson entre dos variables continuas, cuando una de ellas se puede expresar como transformación lineal de la otra. Puesto que una variable dicotómica no puede expresarse como transformación lineal de una variable continua, según Nunnally (1978) el máximo para la correlación *biserial-puntual* sería de 0.798 y correspondería al caso en que la proporción de aciertos en el ítem es igual a la proporción de errores, es decir, p = q = 0.5. Este razonamiento, sin embargo, ha sido objeto de controversia. Los interesados pueden consultar a Hunter y Schmidt (2004) para una discusión sobre el tema.

Para corregir el coeficiente de correlación *biserial-puntal* debido a la contaminación existen distintos procedimientos, sobre los cuales no se entrará en detalle. El más común es restar al puntaje total *X* el puntaje en el ítem dicotómico *Y* y calcular el coeficiente de correlación *biserial-puntal* entre la variable *Z = X - Y* (puntaje total corregido) y el ítem *Y*.

$$\rho_{Z,Y} = \frac{\mu_P - \mu_X - q}{\sqrt{\sigma_X^2 + pq - 2p(\mu_P - \mu_X)}}\sqrt{\frac{p}{q}}$$

Existen otros coeficientes de correlación, que no veremos en este curso, los cuales no surgen de la aplicación directa del coeficiente de Pearson, y que se deben emplear cuando una variable continua se dicotomiza (lo cual es diferente a que una variable sea dicotómica por naturaleza). Dichos coeficientes deben ser *estimados* en base a consideraciones sobre las distribuciones de las variables que se desea correlacionar: coeficiente de correlación *biserial* (entre una variable continua y una variable continua que es dicotomizada) y correlación *tetracórica* (entre dos variables continuas que son dicotomizadas). Los interesados pueden consultar sobre estos dos coeficientes en cualquier texto de estadística.

Otra propuesta para evaluar la capacidad discriminatoria de un ítem es el índice de discriminación basado en proporciones. En este caso se divide la población en dos grupos extremos según el puntaje obtenido en la prueba y para cada grupo se estima la proporción de aciertos en el ítem y se calcula la diferencia.

$$D = P_S - P_I$$

siendo P_S y P_I la proporción de aciertos en el ítem en las subpoblaciones superior e inferior respectivamente. Se pueden emplear distintos criterios para definir los grupos extremos: cuartiles superior e inferior, 27% superior e inferior, etc.

Ebel y Frisbie (1986) sugieren el siguiente criterio general para clasificar a los ítems según el índice de discriminación D:

>= 0.40 excelente. El ítem se conserva.

entre 0.30 y 0.39 bueno. El ítem se puede conservar, pero también puede ser revisado para mejorar.

entre 0.20 y 0.29 regular. El ítem debe ser revisado.

entre 0.10 y 0.19 pobre. El ítem se descarta o se revisa en profundidad.

< 0.10 pésimo. El ítem se descarta.

Las medidas según nos basemos en correlaciones o en proporciones no necesariamente serán coincidentes y no existe una regla para emplear uno u otro criterio. Se trata de definiciones distintas que intentan aproximarse numéricamente al concepto de "ítem que discrimina entre estudiantes con puntajes altos y estudiantes con puntajes bajos".

ÍNDICE DE DIFICULTAD

El índice de dificultad que aquí se presenta es el definido en la Teoría Clásica de los Tests (TCT). Es muy importante tenerlo presente, dado que cuando el marco de trabajo es la Teoría de Respuesta al Ítem (TRI), si bien se emplea la misma denominación, su definición e interpretación son muy distintas. En el marco de la TCT el índice de dificultad depende de la población en que se aplica la prueba. Por lo tanto no es una propiedad del ítem, sino una resultante de la interacción ítem-sujetos. Ésta es una gran diferencia con el índice que se define según la TRI, el cual resulta independiente de las personas y es una verdadera propiedad del ítem. Puede consultarse el artículo de Hambleton y Jones (1993) para una comparación entre ambas teorías.

El índice de dificultad clásico es la proporción o fracción de sujetos que acierta el ítem. Así definido, este índice resulta ser más bien una medida de la *facilidad* del ítem, puesto que el índice aumenta a medida que aumenta la proporción de aciertos.

Los índices de dificultad y de discriminación no son independientes entre sí. Un caso bastante intuitivo es el de ítems con índices de dificultad extremos: muy bajos (0.10 o menos) o muy altos (0.90 o más). En estos casos el porcentaje de acierto en estos ítems habrá sido de 10% o de 90% respectivamente. Es de esperar que un ítem de estas características no discrimine adecuadamente, ya que el 10% de los estudiantes acertó en el primer caso y no acertó en el segundo, con lo cual el puntaje en el ítem prácticamente no presenta variabilidad.

El valor “óptimo” para el índice de dificultad de un ítem dependerá de los objetivos de la prueba. Es común considerar que el valor óptimo debe ubicarse en el punto medio entre la probabilidad de acertar al azar y 1 (100% de aciertos).

Lord (1952) estudió cómo se relacionan el índice de dificultad, el número de alternativas de los ítems y la fiabilidad de la prueba. Sus resultados muestran que la fiabilidad es máxima cuando: a) la variabilidad del índice de dificultad es mínima, o sea, cuando todos los ítems de la prueba tienen una dificultad similar (siendo el caso extremo cuando todos los ítems tienen exactamente el mismo índice de dificultad) y b) cuando el índice de dificultad es ligeramente más alto que el punto medio entre la probabilidad de acertar al azar y 1.

En la siguiente tabla se resumen los resultados de Lord (1952). La tabla debe interpretarse teniendo en cuenta que el autor consideró una prueba formada por ítems de

igual índice de dificultad y de iguales correlaciones inter-ítem (es decir, para cada par de ítems i,j la correlación es una constante).

k	I_{dif} como punto medio del intervalo (1/k, 1)	I_{dif} que maximiza la fiabilidad de la prueba
5	.660	.694
4	.625	.742
3	.667	.770
2	.750	.846

k= número de alternativas por ítem; I_{dif} = índice de dificultad

ADECUACIÓN DE LOS DISTRACTORES

El análisis de los índices de discriminación y de dificultad de cada ítem orienta sobre qué preguntas habría que eliminar, modificar o mantener en futuras pruebas. Pero también interesa comparar el rendimiento diferencial de los estudiantes en cada alternativa (correcta y distractores).

En una situación ideal la mayoría de los estudiantes habrá elegido la alternativa correcta y las restantes alternativas habrán sido seleccionadas por un pequeño número de estudiantes. Si esto no se cumple puede ser simplemente porque la mayoría desconoce la respuesta. Pero también puede ser que la pregunta no haya sido formulada en términos lo suficientemente claros, o que las alternativas presumiblemente incorrectas contengan elementos correctos (alternativas "parcialmente" correctas). Se aconseja revisar especialmente la redacción de aquellas preguntas que hayan tenido un bajo porcentaje de acierto.

Otro aspecto importante a tener en cuenta es que el número de estudiantes que selecciona un distractor debe ser aproximadamente el mismo para todos los distractores. Una alternativa incorrecta poco seleccionada en relación con las demás puede ser indicio de una opción poco plausible, que es fácilmente descartable. Por otro lado, una alternativa incorrecta muy seleccionada respecto a las otras quizás contenga elementos correctos. En ambos casos se aconseja una revisión de la redacción de tales ítems.

Una estrategia útil para el análisis de los distractores es estimar su índice de discriminación. Se recodifican las respuestas de los estudiantes de manera de asignar el valor 1 al distractor en estudio y 0 al resto de los distractores y a la alternativa correcta y se estima el índice de discriminación (puede ser tanto el basado en correlaciones como el basado en proporciones o ambos). De esta manera se obtendrá una medida de qué tanto el distractor discrimina entre estudiantes de alto y bajo rendimiento. Se espera que cuanto

más alto es el rendimiento menor sea el grado de adhesión a determinado distractor, por lo que se esperan índices de discriminación negativos.

Otra estrategia útil es comparar, para cada ítem, la distribución de frecuencias de los distractores y de la alternativa correcta entre estudiantes de alto y bajo rendimiento. Se espera que la mayoría de los primeros haya elegido la alternativa correcta; para los segundos, en cambio, se espera que tanto los distractores como la opción correcta tengan porcentajes similares de adhesión.

¿SE SUMAN LOS PUNTOS DE LOS ÍTEMS?

Pese a que sumar el puntaje parcial obtenido en cada pregunta para obtener el puntaje total es una práctica común, ello sólo tiene sentido si el conjunto de preguntas cuyo resultado estamos sumando depende de un rasgo concreto del estudiante (dimensión). Para ello es necesario tener claro previamente qué rasgo o rasgos se pretende evaluar a través de la prueba. Y luego elaborar una prueba cuyo resultado sirva para medir el nivel de rasgo. En otras palabras, **una prueba válida para los objetivos de evaluación pretendidos**.

Para la validación del contenido de la prueba se recomienda recurrir al juicio de expertos. Un primer panel de expertos (p.e., un grupo de docentes) define los rasgos a evaluar (objetos de evaluación) y elabora las preguntas. Un segundo panel de expertos (otro grupo de docentes) evalúa si las preguntas constituyen una representación adecuada del rasgo a evaluar. En caso de discordancia entre los grupos las diferencias deben resolverse mediante discusiones conjuntas. De tal manera al final se dispondrá de un conjunto de ítems válido para evaluar los rasgos pretendidos.

Una vez identificadas las dimensiones y los ítems que las componen, se puede estimar la **consistencia interna de cada dimensión**. Esto es, la correlación media entre los ítems que la componen. Este estadístico es una medida indirecta de la fiabilidad de la dimensión (o de la prueba, si ésta fuera unidimensional) y oscila entre 0 y 1. Cuando los ítems efectivamente se nuclean en torno a una única dimensión, cabe esperar un rendimiento similar en cada uno de ellos y en la dimensión (consistencia interna próxima a 1). Valores bajos de consistencia interna indican que debe revisarse tanto la redacción de los ítems como la correspondencia item-dimensión (juicio de expertos).

Si todos los ítems evalúan la misma dimensión, entonces tiene sentido pensar en términos de rendimiento global en la prueba. Si así no fuera y la prueba fuera multidimensional, se suman los puntajes de los subconjuntos de ítems que definen cada dimensión; pero no tendría sentido un puntaje global.

¿SE RESTAN PUNTOS POR ERRORES?

Muchas veces las pruebas de múltiple opción incluyen la indicación al estudiante de que a los efectos del puntaje contestar mal es equivalente a no contestar. Ello es así cuando el puntaje en la prueba resulta únicamente de contabilizar el número de aciertos *("number right scoring"*).

En ocasiones se prefiere corregir tomando en consideración algo más que el número de aciertos ("*formula scoring*"). Para ello puede elegirse o bien penalizar los errores o bien "premiar" las omisiones. Ambas formas de corrección están vinculadas con el tema de Corrección por los aciertos al azar y se detallan en el apartado correspondiente.

Respecto a cuál de los dos criterios seguir, *number right scoring* o *formula scoring*, existen trabajos que sustentan tanto una como otra alternativa. Recomendamos la lectura de algunos artículos clásicos donde se discute el tema y que siguen vigentes en la actualidad: Lord (1975), en favor de la corrección, y Glass y Wiley (1964), en favor de considerar únicamente los aciertos. También recomendamos el artículo de Frary (1988).

¿CUÁL ES EL MEJOR PUNTO DE CORTE PARA EL NIVEL MÍNIMO DE SUFICIENCIA?

En realidad no hay un "mejor punto de corte" sino que lo que hay son distintos criterios de selección del punto de corte.

Lo más común es recurrir al juicio de expertos disciplinares. Éstos sugieren el número mínimo de preguntas que deben contestarse satisfactoriamente para cada dimensión. El criterio resultante lógicamente dependerá de la muestra de expertos seleccionada, por lo que se recomienda hacer una validación cruzada, similarmente a lo comentado en ¿Se suman los puntos?

Otra forma de hacerlo es adoptar como referencia, en lugar del juicio de los expertos, el rendimiento en dicha prueba de otro grupo de estudiantes especialmente seleccionado. Por ejemplo, el rendimiento mínimo a exigir puede definirse como el rendimiento medio alcanzado por un grupo de estudiantes cuyo nivel de competencia sea claramente inferior al del grupo que se desea examinar.

Existen otros métodos que se centran en cada ítem y no en el análisis global de la dimensión (métodos de Nedelsky, Angoff y Ebel).

UNIDAD 3. QUÉ HACER Y QUÉ NO HACER CON LOS ACIERTOS AL AZAR

El procedimiento a seguir con los aciertos al azar ha sido muy discutido y existe aún hoy controversia sobre distintos aspectos: si deben tenerse en cuenta, en tal caso cómo debe

corregirse su efecto, qué probabilidad tiene un estudiante de acertar por azar, etc. En los siguientes enlaces comentamos tres aspectos que nos parecen esenciales.

¿INCIDEN EN EL RESULTADO?

Cuando alguien contesta una prueba de múltiple opción conoce algunas respuestas y desconoce otras. Entre estas últimas puede acertar algunas por azar. Por lo tanto, **un primer efecto del azar** en el resultado de una prueba es la introducción de un sesgo en el puntaje: el puntaje obtenido será el correspondiente a las preguntas cuya respuesta el sujeto cree conocer (más allá de que las haya contestado correctamente o no) más el puntaje correspondiente a los aciertos por azar. Este efecto no siempre será el mismo, como veremos en ¿De qué depende el efecto de los aciertos por azar?

Consideremos una prueba de múltiple opción de n ítems. El resultado parcial de cada ítem **contestado al azar** es una variable de Bernouilli, que puede adoptar solamente dos valores, acierto o error, con probabilidades p y q respectivamente. Supongamos que todos los ítems son contestados únicamente por azar, independientemente unos de otros. El resultado del total de la prueba **contestada al azar** será suma de *n* variables de Bernouilli independientes:

$$X = x_1 + x_2 + \ldots + x_n$$

Si asignamos 1 a los aciertos y 0 a los errores, X representa el número total de aciertos al azar. X es una variable aleatoria con distribución binomial, con parámetros: $\mu = np$ y $\sigma^2 = npq$.

El valor medio del número de aciertos al azar es np, pero habrá personas que acierten más y otras menos. Por lo tanto, para dos personas con un mismo nivel de conocimientos, el efecto del azar contribuye a que no obtengan el mismo puntaje. Esto significa que el azar no sólo introduce un sesgo en la estimación del nivel de conocimientos: **un segundo efecto del azar** es que aumenta la imprecisión de las medidas.

¿DE QUÉ DEPENDE EL EFECTO DE LOS ACIERTOS POR AZAR?

Si el total de aciertos por azar X es una variable con distribución binomial, la probabilidad de que X adopte un valor particular X0 viene dada por la ecuación:

$$P(X = X_0) = C_{x_0}^{n} p^{x_0} q^{(n-x_0)}$$

donde n es el número total de ítems de la prueba y p y q las probabilidades respectivas de acierto y error por azar.

En la Figura 1 se muestra cómo impacta el número de ítems en el número de aciertos por azar, comparativamente para dos pruebas de 10 y 50 ítems, con 4 opciones de respuesta.

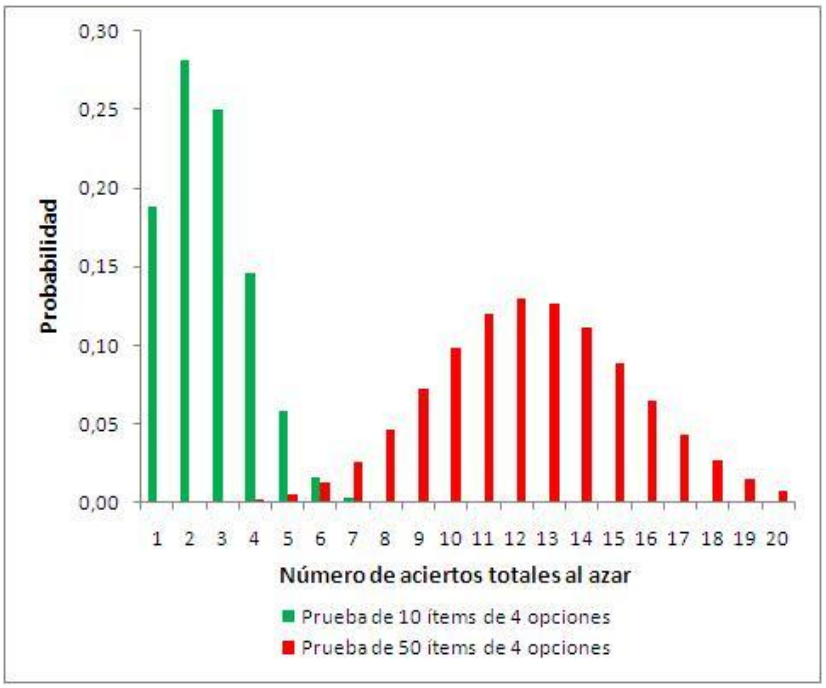

Figura 1. Probabilidad de acierto al azar según el número de ítems

Para ambas el número medio de aciertos al azar representa el 25% de la prueba. Sin embargo las situaciones son distintas. En el caso de la prueba de 10 ítems el número más probable de aciertos por azar es 2 y la probabilidad asociada a dicho número de aciertos es de 28%. El total de aciertos por azar oscilará prácticamente entre 1 y 6 (a partir de 7 la probabilidad es inferior a 1%). En el caso de la prueba de 50 ítems el número más probable de aciertos es 12 y la probabilidad asociada es de 13%. El total de aciertos por azar en este caso oscilará más probablemente entre 6 y 19 (fuera de dicho rango la probabilidad disminuye por debajo de 1%).

Ello quiere decir que si bien en ambos casos el valor medio de los aciertos esperados por azar representa el mismo porcentaje de la prueba, la variabilidad posible en torno a dicho valor medio disminuye a medida que aumenta el total de ítems.

El otro factor que impacta en el efecto de los aciertos por azar es la cantidad k de alternativas por ítem, que es la inversa de la probabilidad p de acertar por azar.

De lo anterior se desprende que a mayor número de ítems y a mayor número de alternativas por ítem menor será el efecto de los aciertos al azar en el resultado. Sin

embargo esto tiene sus limitaciones. Respecto al total de ítems, hay que tener en cuenta que la prueba debe ser factible de ser realizada en el tiempo previsto y no se puede agotar a los participantes con una prueba excesivamente larga. El número óptimo de opciones por ítem se discute en el apartado correspondiente.

¿SE PUEDE CORREGIR EL EFECTO DE LOS ACIERTOS POR AZAR?

Como se mencionó anteriormente, existen dos "escuelas" respecto a cómo corregir las pruebas de múltiple opción: "*number right scoring*", que toma en cuenta solamente el número de aciertos y "*formula scoring*", que efectúa correcciones debido a los aciertos al azar, distinguiendo entre los errores y las omisiones. En este apartado comentaremos esta última.

El efecto de los aciertos al azar en el resultado de una prueba ya fue analizado y se vio que afecta no sólo el puntaje obtenido sino también la precisión de las medidas. Ello sugiere que sería necesario efectuar correcciones para, cuando menos, atenuar este efecto.

La fórmula clásica de corrección se basa en los siguientes supuestos:

- el estudiante o bien conoce o bien desconoce la respuesta
- el estudiante que conoce la respuesta acierta
- el estudiante que desconoce la respuesta contesta al azar, con una probabilidad $1/k$ de acertar

Lo anterior implica que todos los errores provendrán de intentos fallidos de acertar por azar. Por lo tanto, conociendo el número de errores es posible calcular cuántas preguntas fueron contestadas al azar y de ellas cuántas resultaron correctas. Se deduce fácilmente que:

$$Puntaje\ corregido\ = N^{o}\ aciertos - N^{o}\ errores\ /\ (k-1)$$

Otra forma de corregir es:

$$Puntaje\ corregido\ = N^{o}\ aciertos + N^{o}\ omisiones\ /\ k$$

La segunda forma de corrección se basa en los mismos supuestos de la primera. En lugar de penalizar los errores favorece a quienes dejan las respuestas sin contestar otorgando 1/k puntos por cada omisión, que es lo que promedialmente habrían obtenido si en lugar de dejar la pregunta en blanco la hubieran contestado al azar. Si se elige esta forma de corrección también es necesario corregir el punto de corte establecido para el mínimo de suficiencia.

Ambas formas de corrección correlacionan perfectamente (Ebel, 1979). La segunda refuerza entre los estudiantes la conducta de no contestar al azar: al dejar una pregunta en blanco se obtendrá una pequeña ganancia segura ($1/k$ puntos), en tanto que al contestar al azar la ganancia tal vez sería mayor (1 punto) pero también más incierta (probabilidad $1/k$).

Debe tenerse presente la escasa plausibilidad de los supuestos en que se sustentan estas fórmulas. En general los estudiantes pueden tener un "conocimiento parcial" del tema, lo que invalidaría los supuestos 1 y 3. Por otra parte, el segundo supuesto está sujeto a que los ítems estén confeccionados con gran rigurosidad metodológica y hayan sido piloteados previamente.

UNIDAD 4. ANÁLISIS FORMAL CLÁSICO DE UNA PRUEBA DE MÚLTIPLE OPCIÓN

Una vez implementada la prueba se hace necesario un análisis formal de los ítems a partir de las respuestas, cuyas características generales en el marco clásico se muestran en la Unidad 2. Este análisis resulta particularmente necesario cuando se está piloteando una prueba y se requiere de insumos que orienten al equipo docente en su decisión de mantener, eliminar o revisar ítems, con miras a obtener una versión mejorada de la prueba.

Pero hay dos aspectos esenciales a considerar previo a efectuar el análisis. Primero, los estadísticos que surgen del análisis clásico deben ser entendidos en el contexto en que se realiza la prueba y siendo conscientes de las limitantes que existen. Algunas de estas, adaptadas de Mehrens y Lehmann (1973), son:

- el análisis formal clásico es siempre tentativo, ya que los valores de los índices que se suelen calcular (dificultad y discriminación) dependen de la población en que se aplica la prueba. Si una misma prueba se vuelve a aplicar, es necesario repetir el análisis formal con los nuevos datos.
- el índice de discriminación de un ítem no siempre es una medida de la calidad del mismo. Por ejemplo, un ítem que resulta muy fácil (o muy difícil) no tendrá una buena capacidad para discriminar, pues la mayoría de los estudiantes habrá elegido la respuesta correcta (o algún distractor). Sin embargo, el docente puede considerar necesaria su inclusión a fin de lograr un muestreo adecuado de los contenidos a evaluar.
- no debe confundirse el análisis formal de los ítems con el análisis de la validez de su contenido. Para que los ítems sean válidos se requiere de un criterio externo a los datos (juicio de expertos), que determine qué tanto esos ítems están evaluando lo que se pretende evaluar. El análisis de la validez debe hacerse antes de

implementar la prueba, comparando el contenido de los ítems con los objetivos de aprendizaje pretendidos (véase Sartori y Pasini (2007) para los distintos tipos de validez que existen). En cambio el análisis formal es posterior a la implementación de la prueba y arroja información sobre cómo se comportan los estudiantes frente a los ítems.

El segundo aspecto a tener en cuenta es que no existe una sola forma de efectuar el análisis formal. Aunque los índices son definiciones precisas, dicho análisis no es un protocolo de acción luego de aplicar una prueba de múltiple opción. El análisis será útil en tanto sea entendido como un conjunto de herramientas que puede ayudar a los docentes a mejorar sus prácticas de enseñanza y de evaluación. Ejemplos de preguntas que el docente pretende contestar a través de un análisis formal pueden ser:

- ¿Este ítem resulta demasiado fácil?
- ¿Qué tanto predice este ítem el rendimiento del estudiante en la prueba?
- ¿Hay distractores problemáticos?
- ¿Este ítem resulta inadecuado para algún conjunto de estudiantes?
- Etc.

Y en función de las respuestas, determinar las pautas de acción a seguir. Las anteriores son ejemplos de preguntas, pero es cada docente el que determina las preguntas que desea contestar y para lo cual se vale de un análisis formal.

Habiendo aclarado lo anterior abajo les dejamos algunas recomendaciones muy esquemáticas para la revisión de los ítems, que pueden facilitar el trabajo de quienes se inician en la temática. Recuerden que se trata de ejemplos solamente. Son ustedes los que decidirán cómo conducir sus propios análisis.

RECOMENDACIONES A SEGUIR PARA INCLUIR / ELIMINAR ÍTEMS EN UNA PRUEBA DE MÚLTIPLE OPCIÓN

Estas recomendaciones están adaptadas de Santisteban y Alvarado (2001).

1. Aplicar la prueba de n ítems a una muestra de sujetos similar a la población objetivo en que habrá de aplicarse la prueba final. Estimar los índices de discriminación y de dificultad de cada ítem, así como la consistencia interna de la prueba (ejecución piloto).
2. Identificar los ítems con índices de discriminación muy bajos (inferiores a 0,2) y eliminarlos: valores tan bajos revelan una correlación muy débil.
3. Identificar los ítems con índices de dificultad extrema (por debajo de 0,1 ó por encima de 0,9) y eliminarlos. Valores muy bajos pueden indicar problemas de

contenido del ítem; valores muy altos indican que el ítem es demasiado sencillo y proporciona poca información acerca de los estudiantes.

4. Luego de la eliminación primaria, ordenar los ítems restantes en función de sus índices de discriminación en orden decreciente y seleccionar los primeros n/2. Para éstos, estimar la consistencia interna de la nueva prueba que conforman, así como sus nuevos índices de discriminación y dificultad. También se puede graficar la distribución de los resultados.
5. Si se desea aumentar (o disminuir) el puntaje promedio obtenido será necesario intercambiar algunos ítems. Por ejemplo, remplazar ítems con valores bajos (altos) de índice de dificultad por otros con valores más altos (bajos). De este modo la prueba contendrá ítems más sencillos (difíciles), aumentando (disminuyendo) la media del puntaje total.
6. Para evitar o minimizar el impacto del intercambio de ítems sobre la fiabilidad de la prueba, en lo posible se debe comenzar por remplazar los ítems con valores de índice de discriminación más bajos.

No olvidar que estos son apenas lineamientos muy generales y que es imprescindible tener en cuenta el contenido de los ítems para un correcto análisis. Es muy frecuente tener que balancear los distintos factores y llegar a una solución de compromiso.

MATERIAL COMPLEMENTARIO

ORIGEN Y USO DE LAS PRUEBAS DE MÚLTIPLE OPCIÓN

El primer test de inteligencia fue desarrollado a principios de siglo XX por los psicólogos franceses Alfred Binet y Theodore Simon, como un instrumento de medida objetiva, capaz de identificar distinto grado de retardo mental (Binet y Simon, 1905). La primera versión fue revisada y expandida y los resultados de su implementación fueron tan exitosos que se logró un reconocimiento institucional del papel de los tests en los procedimientos de diagnóstico (Navas, 1999).

Los tests de aplicación colectiva se introdujeron por vez primera durante la I Guerra Mundial, ante la necesidad de Estados Unidos de reclutar personal en cantidad y de manera rápida. Así se desarrollaron el test Alfa -prueba verbal - y el test Beta - prueba no verbal dirigda a personas con problemas de alfabetización. El éxito de su aplicación determinó que una vez finalizada la guerra se continuara con el uso de estos tests, extendiendo su uso a la sociedad civil (áreas educativa, industrial, clínica) no sólo con fines de diagnóstico sino también para la evaluación de personas normales (Navas, 1999).

Este tipo de prueba, pese a su uso generalizado, ha recibido numerosas críticas desde el inicio, más allá de las que resultan de las limitaciones inherentes a cualquier instrumento de medida. La controversia acerca del cociente intelectual determinó que los tests de inteligencia fueran - y continúen siendo - muy cuestionados, especialmente cuando se emplean con fines de selección o de distribución de recursos en ámbitos laborales o educativos, en los cuales la capacidad del test como predictor de la conducta posterior de la persona puede ser muy discutible (Graham y Lilly, 1984; Jensen, 1980; Wigdor y Gardner, 1982).

Otras críticas provienen de la interpretación que se hace de los resultados obtenidos con los tests y de su potencial uso como instrumentos de discriminación social (Gould, 1981). El abordaje sistemático a este problema surge en la década de los '60 en Estados Unidos, a partir del interés de los investigadores en desarrollar procedimientos estadísticos para determinar si las diferencias de rendimiento entre distintos grupos (hombres y mujeres, grupos étnicos, etc.) en determinados ítems de una prueba son realmente atribuibles a diferentes niveles de capacidad o si provienen de la existencia de un sesgo en dichas preguntas, favorable a los sujetos de uno de los grupos. Así surge como área de investigación el Funcionamiento Diferencial del Ítem, también conocido como DIF por sus siglas en inglés (Differential Item Functioning).

En general, según Muñiz (1998) y Navas (1999), las principales críticas provienen del mal uso de las pruebas más que de las características del instrumento. Y el mal uso puede deberse a cuestiones éticas (p.e., utilizar el instrumento con fines de discriminación social)

o, más comúnmente, a la falta de información o desconocimiento (p.e., aplicar el test en una población que no es la adecuada).

FIABILIDAD

La fiabilidad es un concepto exclusivamente clásico y hace referencia siempre a la totalidad de una prueba unidimensional. Cuando una prueba está compuesta por mútliples dimensiones debe estimarse la fiabilidad para cada una de ellas separadamente.

Se entiende por fiabilidad de una prueba unidimensional el grado de precisión – no de exactitud – con que esta mide el atributo deseado. A grandes rasgos, el concepto de fiabilidad puede entenderse de la siguiente manera. Supongamos que se desea medir el nivel de comprensión lectora de una cierta población y para ello se diseña una prueba de múltiple opción. Por simplicidad supongamos que todos los ítems se califican con 0 (error) o 1 (acierto) y que el puntaje total X es la suma de los puntajes parciales. El puntaje total X presentará una cierta dispersión o variabilidad, dado que no todos los estudiantes habrán obtenido el mismo puntaje. La pregunta es ¿qué tanto la variabilidad de X representa la variabilidad de la comprensión lectora de los estudiantes? O lo que es lo mismo ¿qué tanto de la variabilidad observada en el puntaje X es atribuible a la variabilidad de la comprensión lectora de los estudiantes?

En un extremo encontramos que toda la variabilidad observada en X se debe a la variabilidad en comprensión lectora: entonces X es una “medida perfecta”, sin error, y la fiabilidad de la prueba es 1. En el otro extremo encontramos que no hay variabilidad de la comprensión lectora (todos los alumnos tienen exactamente el mismo nivel de comprensión lectora) y sin embargo hay variabilidad en el puntaje X. En este caso la variabilidad de X no puede ser atribuida en ningún grado a la variabilidad de la comprensión lectora; por lo tanto la variación en X se debe enteramente a errores aleatorios inherentes al proceso de medición; la fiabilidad de la prueba es 0.

En la realidad no sucede ninguna de las dos cosas, la fiabilidad está comprendida entre 0 y 1; cuanto más próxima a 1 menor es el error de medida. En términos prácticos, la fiabilidad se suele estimar a partir de las correlaciones entre los ítems que componen la prueba.

No debe confundirse fiabilidad con validez de contenido. La fiabilidad es una medida del error con que se mide el atributo a través del puntaje total en una prueba; se estima a partir de los datos empíricos. La validez de contenido se refiere a qué tan representativa es la prueba del universo de posibles pruebas para evaluar los contenidos pretendidos; requiere del juicio de especialistas en la disciplina.

TEORÍA CLÁSICA DE LOS TESTS Y TEORÍA DE RESPUESTA AL ÍTEM

TEORÍA CLÁSICA DE LOS TESTS

La Teoría Clásica de los Tests (TCT) se ubica a comienzos del siglo XX con los primeros trabajos realizados por Charles Spearman (1904, 1907, 1913). Básicamente se trataba de encontrar un modelo estadístico que fundamentara de manera adecuada los puntajes de los sujetos en los tests y que a su vez permitiera la estimación de los errores de medida. La TCT es, ante todo, un modelo de trabajo. Postula un modelo matemático muy sencillo, que vincula el rasgo o atributo que se desea evaluar con la medida empírica, a través de una relación lineal. La TCT toma al test global como unidad de análisis, su interés se centra en el puntaje total de los sujetos que realizan el test, sin perjuicio de que además permite la estimación de algunos parámetros propios de los ítems.

Supuesto de la TCT: $X = V + e$

X es una medida empírica y corresponde al puntaje total obtenido en la prueba; V es una magnitud teórica denominada "puntaje verdadero" (V se define como la esperanza matemática de X); e representa el error aleatorio en la medida.

Los supuestos subyacentes al modelo clásico son:

- Los puntajes verdaderos V de los sujetos en un test y sus correspondientes errores de medida e no están correlacionados: $\rho(V, e) = 0$.
- Los errores de medida de los sujetos en un test i no están correlacionados con los errores de medida de los mismos sujetos en otro test paralelo k: $\rho(e_i, e_k) = 0$
- Dos tests i, k se denominan paralelos si sus errores tienen la misma varianza y si los sujetos tienen los mismos puntajes verdaderos en cada uno: $\sigma^2(e_i) = \sigma^2(e_k)$ y $V_i = V_k$

La debilidad de los supuestos de la TCT y su escasa plausibilidad real constituyen al mismo tiempo su fortaleza que es la sencillez de los tratamientos matemáticos requeridos

para calcular los estadísticos. Una aguda crítica a la TCT puede leerse en Lumsden (1976).

TEORÍA DE RESPUESTA AL ÍTEM

Las deficiencias de la TCT fueron un factor desencadenante de modelos alternativos y hacia fines de la década de los ’60 surgen nuevas perspectivas en torno al abordaje de los tests. Así comienza una nueva línea de trabajo complementaria, conocida hoy como Teoría de Respuesta al Ítem (TRI), inicialmente denominada Teoría del Rasgo Latente. Este marco de trabajo cobra hegemonía en la década de los ’80, pero ello no supone el fin del enfoque clásico. El modelo clásico sigue siendo apropiado en numerosas situaciones en las que la sofisticación de los procedimientos de la TRI no permite su aplicación con eficacia.

A diferencia de la TCT, que postula un único modelo, bajo el nombre TRI se agrupa a una serie de modelos matemáticos, tanto para ítems dicotómicos como para politómicos. Los modelos se distinguen por su formulación matemática (modelos de Lord, Birnbaum, Rasch, etc.), pero todos procuran vincular la probabilidad de obtener determinada respuesta en el ítem i en función del nivel de aptitud (atributo) de los sujetos y de ciertos parámetros de dicho ítem. A diferencia de la TCT, la TRI no se enfoca en el test como un todo sino en cada ítem separadamente. Por lo tanto existen tantas curvas de probabilidad como ítems componen un test. En todos los modelos la probabilidad de acierto del ítem i para sujetos con nivel de aptitud θ se representa como $P_i(\theta)$ y los parámetros del ítem se representan como a_i (índice de discriminación), b_i (índice de dificultad) y en algunos casos c_i (índice de conjetura).

Los índices de discriminación a_i y dificultad b_i, aunque tienen la misma denominación de los índices clásicos, en el marco de la TRI tienen otras definiciones e interpretaciones. El índice de dificultad es un índice de posición del ítem en una escala de aptitud: describe qué tanto nivel de aptitud se requiere para contestar el ítem correctamente. El índice de discriminación i indica hasta qué punto un ítem permite diferenciar entre sujetos de aptitud inferior y superior a la posición del ítem: refleja la tasa de cambio en la probabilidad de éxito según se aumenta la aptitud de los sujetos. Estos índices son propios del ítem y no dependen de la población en que se aplican. Ello supone una gran ventaja frente a la TCT, pues permite la confección de bancos de ítems estandarizados que pueden ser aplicados en cualquier población, dado que sus propiedades son invariantes.

Supuestos de la TRI:

- unidimensionalidad del test: hay una única aptitud que se pone en juego para contestar el test

- ausencia de factores de velocidad: cuando los sujetos fallan al contestar NUNCA se debe a falta de tiempo sino a limitaciones en sus conocimientos
- independencia local: las respuestas a los ítems por parte de sujetos con el mismo nivel de aptitud son estadísticamente independientes. Es decir, la respuesta a un ítem no está afectada por las respuestas al resto de los ítems.

A diferencia de la TCT, la TRI parte de supuestos más reales, pero como contrapartida requiere tratamientos matemáticos mucho más sofisticados debido a la complejidad de los modelos propuestos. Un excelente material online para estudiar la TRI es el libro de Baker (2001).

REFERENCIAS

Abella, A. (2014). Índice de discriminación. Material para el curso "Diseño y corrección de pruebas de múltiple opción, 2013".

Albanese, M. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28–33.

Baker, F. (2001). *The basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Binet, A., y Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. *L'Année psychologique, 11*, 191–336.

Dawson-Saunders, B., Nungester, R.J. y Downing, S.M. (1989). A comparison of single best answer multiple-choice items (A-type) and complex multiple-choice items (K-type). *Proceedings of the Twenty-Eighth Annual Conference on Research in Medical Education*, 161–166.

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

Downing, S.M., Dawson-Saunders, B., Case, S.M. y Powell, R.D. (1991, Abril). *The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II characteristics*. Ponencia presentada en la Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Ebel, R. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. y Frisbie, D.A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Frary, R.B. (1988). Formula scoring of multiple choice tests (correction for guessing). *Educational measurement: Issues and practice, 7*(2), 33-38.

Frary, R.B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education, 4*, 115–124.

Glass, G.V. y Wiley, D.E. (1964). Formula scoring and test reliability. *Journal of Educational Measurement, 1*(1), 43-49.

Gould, S.J. (1981). *The mismeasure of man*. New York: Norton.

Graham, J. y Lilly, R.S. (1984). *Psychological testing*. Englewood Cliffs, NJ: Prentice-Hall

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items. *Evaluation and the Health Professions, 17*(1), 123-126.

Haladyna, T. (2006). Perils of Standardized Achievement Testing. *Educational Horizons*, 30-43.

Haladyna, T. M., y Downing, S. M. (1989a). A taxonomy of multiple-choice item writing rules.*Applied Measurement in Education, 2*(1), 37-50.

Haladyna, T. M. y Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 2*(1), 51-78.

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

Hambleton, R.K. y Jones, R.W. (1993). Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational measurement: Issues and practices, 12*(3), 38-47.

Hopkins, K., Stanley, J., y Hopkins, B. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hunter, J.E. y Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.

Jensen, A.R. (1980). *Bias in mental testing*. New York: The Free Press.

Knowles, S.L. y Welch, C.A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using none-of-the-above. *Educational and Psychological Measurement*, *52*, 571-577.

Lord, F.M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, *17*(2), 181-194.

Lord, F.M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*(1), 7-11.

Lumsden, J. (1976). Test Theory. *Annual Review of Psychology*, 251-280.

Mehrens, W.A. y Lehmann, I.J. (1973). *Measurement and evaluation in Education and Psychology*. New York: Holt, Rinehart and Winston.

Mueller, D.J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational and Psychological Measurement*, *35*, 135-41.

Muñiz, J. (1998). *Teoría clásica de los tests*. Madrid: Pirámide.

Navas, M.J. (1999). Un siglo usando tests. *Revista Electrónica de Metodología Aplicada, 4*(2), 1-11.

Nunnally, J. C. (1978). *Psychometric theory*. Nwy YorK : McGraw-Hill.

Roberts, D. M. (1993). An empirical study on the nature of trick test questions. *Journal of Educational Measurement, 30*, 331–344.

Rodriguez, M. C. (1997, Abril). *The art and science of item writing: A meta-analysis of multiple-choice item format effects*. Ponencia presentada en la Annual Meeting of the American Educational Research Association.

Rodriguez, M.C. (2005). Three options are optimal for muliple-choice items: A meta-analysis of 80 years of research. *Educational measurement: Issues and practice*, 3-13.

Santisteban, C. y Alvarado, J. (2001). *Modelos psicométricos*. Madrid: UNED Ediciones.

Sartori, R. y Pasini, M. (2007). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity, 41*, 359–374.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1907). Demonstration and formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1913). Corelations of sums and differences. *British Journal of Psychology, 5*, 417-426.

Suen, H.K., y McClellan, S. (2003). Item construction principles and techniques. En N. Huang (Ed.), *Encyclopedia of vocational and technological education* (Vol 1, pp. 777-798). Taipei: ROC Ministry of Education.

Thorndike, R. M., Cunningham, G., Thorndike, R. L., y Hagen, E. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York: Macmillan.

Wigdor, A.K. y Garner, R. (1982). *Ability testing: Uses, consequences and controversies*.Washington, DC: National Academy Press.

Por favor no imprima si no es necesario. Cuidar el medioambiente es responsabilidad de TODOS.