10

Análisis de items

1. Introducción: Análisis cualitativo y análisis cuantitativo de items

En el proceso de elaboración de un test una fase esencial es la de análisis de los items. Su propósito es garantizar que los items son adecuados de acuerdo con los propósitos del constructor del test. Esta fase, por lo general, es previa al análisis de la fiabilidad y de la validez del test como un todo.

El análisis de los items consiste en efectuar un estudio de la "calidad" de los items que comprende un análisis cualitativo y otro cuantitativo. Por lo general el análisis cualitativo precede al cuantitativo.

El *análisis cualitativo de los items* pretende garantizar que el contenido, el muestreo de contenido y la forma sean adecuados.

1. *Que el contenido del ítem sea el adecuado.* Es un aspecto muy importante y supone que, racionalmente, pueda establecerse con claridad que el ítem mide aquello que tiene que medir; o, al menos, que pueda establecerse que no mide otras cosas distintas aunque relacionadas con aquello que la escala pretende medir. Este análisis es esencial para garantizar una escala razonable y no puede ser sustituido por ningún cálculo o fórmula. Este análisis tiene que ver con la delimitación semántica del ítem, con establecer su significado psicológico, aquello que se supone justifica el que construyamos items.

2. *Que los items formen una muestra adecuada de contenidos.* Es decir, que en conjunto, los contenidos de los items sean una buena muestra de aquello que el test pretende medir. También es esencial para la calidad del test, y tampoco es sustituible por ninguna fórmula o cálculo sobre datos empíricos. Si los items no presentan contenidos adecuados no será posible obtener una muestra adecuada de contenidos. Ahora bien, es posible que un conjunto de items presenten uno a uno un contenido adecuado aunque en conjunto el muestreo sea inaceptable.

Los puntos 1 y 2 constituyen los pilares para el establecimiento de la *validez de contenido*

del test, Este análisis tiene que ver con 1) la delimitación semántica de la variable, es decir, con establecer el significado psicológico del test, aquello que se supone justifica el que construyamos el test, con aclarar aquello que se quiere medir, 2) con que cada ítem desde su formulación mida eso y no otra cosa, es decir, con que el contenido de cada ítem corresponda con el del constructor que se pretende medir y 3) con que los items en su conjunto correspondan adecuadamente a ese contenido, es decir, que en conjunto representen adecuadamente la totalidad del constructo de un modo adecuado a la estructura psicológica del mismo.

3. *Que los aspectos formales del ítem sean adecuados.* Es decir, que los aspectos de redacción, comprensión, etc. garanticen que los sujetos pueden acceder generalmente a una interpretación adecuada de los mismos. Esta es una condición elemental necesaria para que el test satisfaga bien cualquier otra. La ambigüedad, la polisemia, las dobles negaciones, las frases demasiado generales a las que se puede dotar de diferentes significados, etc. impiden esclarecer a qué contenido responde el sujeto, y, por tanto, no es posible sostener la validez de contenido si este permanece poco accesible a la compresión de los sujetos.

El *análisis cuantitativo de los item* pretende garantizar que la distribución, fiabilidad, validez de cada ítem sea adecuada. Estos análisis implican administrar el test a una muestra adecuada y se basan sobre los estadísticos obtenidos de las puntuaciones de los sujetos de esa muestra.

1. *Qué los items presenten una distribución adecuada.* La distribución de frecuencias de una variable expresa cuantos casos han aparecido de cada uno de los valores posibles. No cualquier distribución de puntuaciones puede admitirse para cualquier ítem. Qué distribución deben tener las respuestas a un ítem para que sea aceptable depende esencialmente del tipo de ítem de que se trate y de los propósitos que el constructor del test ha concebido para él. Hay varios indicadores estadísticos a tener en cuenta: la distribución de frecuencias misma, la media, la mediana, la desviación típica, el rango... La media de los items con respuesta verdadera valorados dicotómicamente representa la dificultad de los mismos y es un indicador especialmente importante para estos items que merece una atención especial.

2. *Qué los items sean fiables.* Qué los items sean fiables significa algo muy parecido a qué los tests como un todo sean fiables. El indicador más utilizado

es la homogeneidad del ítem o la correlación del ítem con el total del test, (aunque puede recibir diversos nombres), sobre la que se han elaborado diversas variantes. Este indicador esencialmente refleja si hay una relación lineal entre las puntuaciones de los sujetos en el ítem y las puntuaciones de los sujetos en el test.

3. *Qué los items sean validos.* El significado también es análogo al de validez del test. El indicador más utilizado es la correlación del ítem con uno, o varios, criterios externos, que se denomina índice de validez del ítem. Esta correlación indicará si existe una relación lineal entre el ítem y el criterio.

Muy frecuentemente se designa como "análisis de items" el análisis cuantitativo de items, relegando al análisis cualitativo a un segundo plano u omitiendo por completo su consideración.

Importancia del análisis cualitativo de items

Hay una tendencia a sobrevalorar la importancia del análisis cuantitativo de los items y a infravalorar el análisis cualitativo. Quizás porque el análisis cuantitativo puede "objetivarse" en algunos indicadores característicos. Sin embargo, el análisis cualitativo es *principal* en la definición de aquello que se está midiendo y no puede ser suplido por los análisis cuantitativos.

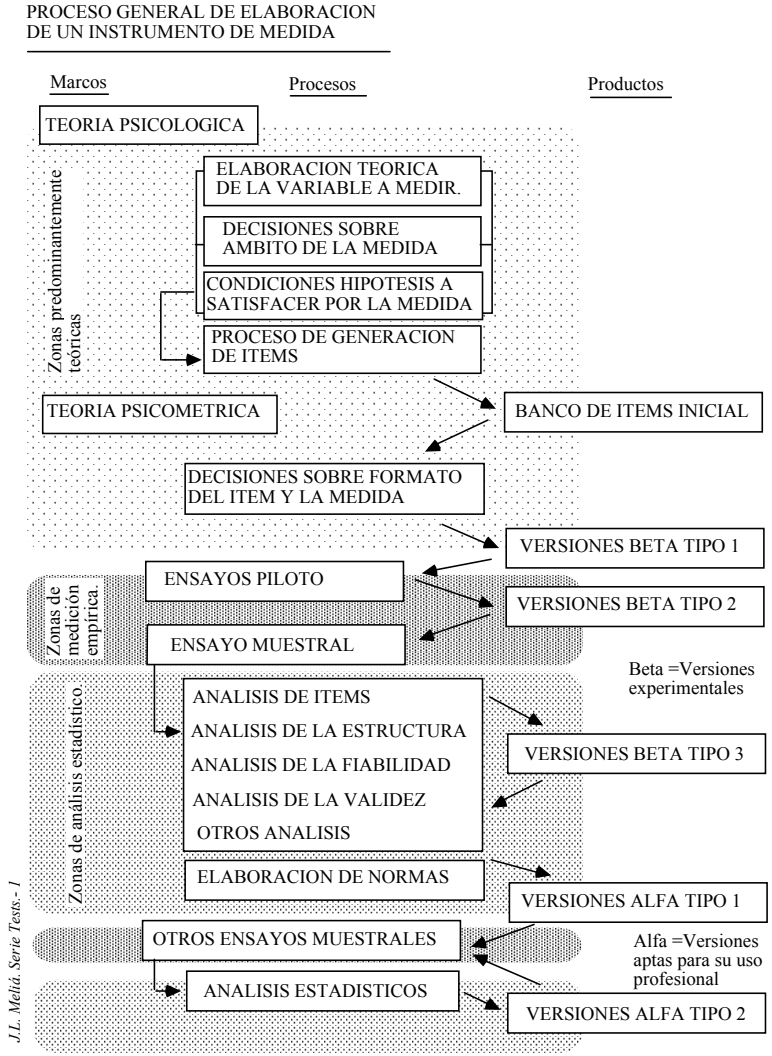
Aunque algunos análisis cuantitativos pueden sugerir ideas acerca del formato, del significado de los items y del muestreo de contenidos contribuyendo a establecer si el análisis cualitativo fue adecuado, lo cierto es que es posible encontrar y diseñar escalas, cuestionarios y tests cuyos análisis cuantitativos produzcan resultados razonables y que, sin embargo, no resistan una discusión cualitativa. Dicho de otro modo, los análisis cuantitativos a que se someten los items y los tests no garantizan que su contenido, muestreo y forma sean adecuados, aunque pueden contribuir a apoyar o no estas afirmaciones.

También es cierto que el más cuidadoso análisis cualitativo no garantizará un resultado positivo de los análisis cuantitativos, aunque por lo general contribuirá poderosamente a ellos. La conclusión es obvia: ambos análisis son imprescindibles y han de complementarse mutuamente, guiados por los propósitos psicológicos del constructor de la prueba.

2. Lugar del análisis de items en el análisis de un instrumento de medida

Aunque teóricamente la fase de análisis de items puede aparecer separada y previa al análisis de la dimensionalidad, a la decisión sobre cómo obtener la puntuación total, y al análisis de fiabilidad y validez de la prueba, lo cierto es que estas cuestiones se implican mutuamente y con frecuencia suponen la realización reiterada de análisis y nuevos análisis introduciendo las variaciones sugeridas por los resultados de los análisis anteriores.

En el proceso general de construcción de un instrumento de medida representado en la figura siguiente, el análisis cualitativo de items ocupa un lugar destacado dentro del proceso de generación de los items que constituye el momento en que, propiamente, se crea el instrumento. El análisis cuantitativo juega un papel principal dentro del conjunto de análisis a que se somete el instrumento cada vez que este se somete a un ensayo muestral.



Más que como un recorrido lineal y secuencial, la construcción de un instrumento de medida puede verse como la confluencia de cuatro procesos cíclicos:

el análisis cualitativo de los items y de la escala como un todo,

el escalamiento de los items,

el análisis (cuantitativo) de los items, y

el análisis (cuantitativo) de la escala como un todo.

En la figura siguiente se refleja esta concepción del proceso de elaboración de un instrumento. Los resultados de cada una de esas fases son relevantes para determinar qué items formarán la prueba final: Cada fase ayuda a tomar decisiones sobre que items excluir, que items añadir y que items modificar. Se comienza, generalmente, por el análisis cualitativo, y a partir de ahí se recorre el esquema en el sentido de la agujas del reloj, es decir, después se escala, se analizan los items y, por último, se analiza la escala como un todo. Pero cada paso puede redefinir que es lo que hay que analizar excluyendo, modificando o incluyendo items, lo que provoca una "vuelta atrás" y el inicio de un nuevo ciclo parcial o total de análisis.

Por ejemplo, para analizar cuantitativamente los items (lo que se conoce en la mayoría de los manuales simplemente como análisis de items), es necesario definir como se obtiene el total de la escala y calcular el total de la escala para cada sujeto. Sin embargo, bastará que el análisis de la fiabilidad o de la validez de los items nos lleve a excluir un solo ítem para tener que volver atrás a calcular un nuevo total y rehacer cuantos análisis cuantitativos se hayan efectuado sobre la escala total o utilizando su puntuación total pues al excluir un solo ítem hemos variado el instrumento total, (y en algunos casos puede, incluso, haber variado o haberse matizado su significado psicológico).

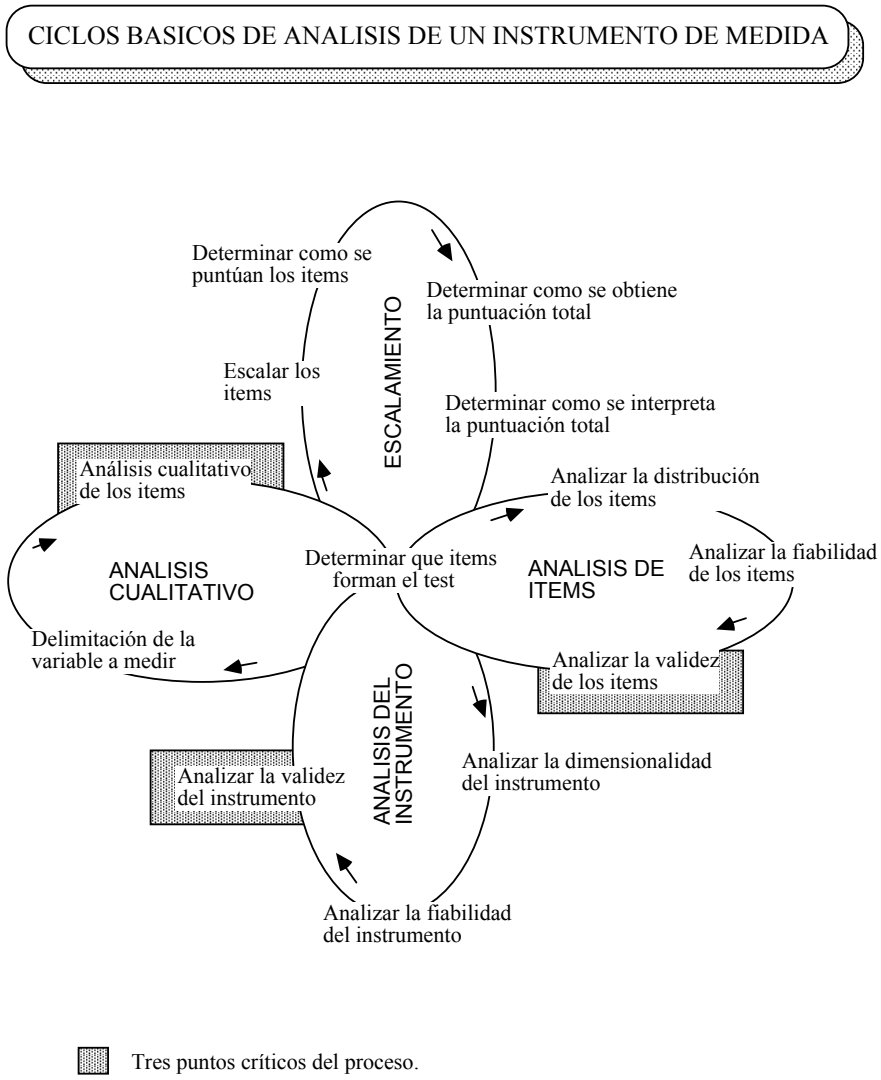
Lo usual es trabajar por etapas. Es decir, se define que estadísticos son de interés. Se calculan y se toman las decisiones más claras acerca de excluir y aceptar items en un determinado instrumento o dimensión de un instrumento. Se vuelven a calcular los totales y los estadísticos relativos a los totales y de nuevo se analizan los items, la dimensionalidad... Se vuelven a tomar decisiones y se sigue adelante hasta encontrar soluciones razonables, lo más adecuadas posible con los datos disponibles.

3. Análisis cuantitativo de los items

Este análisis, como hemos anticipado, tiene tres componentes: el análisis de la distribución de las puntuaciones, el análisis de fiabilidad de los items y el análisis de la validez de los items.

Estos análisis presuponen que se ha administrado la prueba a una muestra adecuada, semejante a aquélla en la que después se piensa utilizar el test y que se ha calculado la puntuación total de cada sujeto en la prueba. Algunos análisis supondrán, además, haber tomado alguna información adicional de los casos, tal como la puntuación en criterios externos al tests.

> Un "**criterio**", o "criterio externo", es una variable distinta al test y medida de forma independiente al test con la que se espera que las puntuaciones en los items del test, en partes del test o en el total del test mantengan determinadas relaciones.
>
> Clásicamente se definía un criterio como una variable con la se esperaba que el test correlacionará (por lo general alto y positivo), contribuyendo a indicar que el test, sus partes o sus items miden con el sentido que se esperaba.



La mayoría de los tests clásicos de inteligencia y aptitudes son tests de lápiz y papel donde los items tienen *una respuesta correcta general que se valora con un punto y otra u otras respuestas incorrectas que se valoran, en el caso más sencillo, con cero puntos*. Para abreviar llamaremos a esta clase de items ***items dicotómicamente valorados***.

La puntuación total se obtiene, en el caso más sencillo y común, como la simple suma de puntos. Es decir, la puntuación total de un sujeto es el número de items que ha acertado.

El otro tipo de items de mayor interés por su frecuencia de uso son aquellos, propios de escalas de actitudes y opiniones, que ofrecen una *escala graduada* de respuestas que, generalmente, muestra en que grado el enunciado del ítem se ajusta al sujeto. Para abreviar llamaremos a estos ***items graduados***.

Los items tipo Likert son un buen ejemplo de items graduados. En estos items, generalmente, no hay una respuesta correcta general para todos los sujetos.

4. Análisis de la distribución de puntuaciones

La distribución de las respuestas que recibe un ítem ha de cumplir las características y particularidades que se esperen de ese ítem según cual sea su función en la prueba. En algunos casos se formulan propósitos e hipótesis específicas para la forma de las distribuciones de las puntuaciones. En otros tan sólo pueden considerarse características generales que hagan al ítem no rechazable para el propósito del test.

La más general de esas características es que ha de *mostrar variabilidad*, el grado y tipo de variabilidad que resulte adecuado.

Si se trata de un ítem graduado, esperamos que, en general, la muestra se distribuya a lo largo de las diversas respuestas posibles, o, al menos, en cierto rango de ellas. Si toda la muestra diera la misma respuesta al ítem éste carecería de interés práctico para determinar la posición de los sujetos de la muestra en la escala bajo consideración.

Esto suele suceder con items de actitud que manifiestan un punto de vista tan extremo, por exceso o defecto, que la respuesta de aceptación o rechazo es unánime. O con items “no comprometidos” o bien aceptables socialmente donde pueden aparecer consistentemente respuestas de aceptación que no ayudan a comprender la posición de los sujetos en la escala.

También sucede este fenómeno con items mal diseñados que, en lugar de averiguar la actitud u opinión de los sujetos en realidad preguntan sobre cuestiones de hecho, para las que la respuesta depende simplemente de estar informado.

Si, en general, las respuestas de los sujetos se describen bien con tres o cuatro categorías, entonces será innecesario utilizar cinco, seis o siete. De este modo el análisis de las respuestas de los sujetos a items graduados puede ayudar a sugerir modificaciones en el número de categorías o en el enunciado de los anclajes verbales. Se denominan *anclajes verbales* a cada uno de los enunciados cortos que explican verbalmente el significado de los grados de respuesta que se pueden dar a un ítem. Por ejemplo, son anclajes verbales las expresiones "1. Muy en desacuerdo", o "3. Bastante". Cada una de ellas intenta dar un referente semántico para el número al que acompañan que, a su vez, indica una posición determinada en la escala de respuestas al ítem.

Un buen análisis cualitativo -racional- de los items, tanto *a priori* -antes de aplicar la escala por primera vez- como *a posteriori* -después de ver resultados de los análisis cuantitativos de los items- contribuirá a paliar o eliminar la mayoría de estos problemas.

Si un ítem dicotómicamente valorado tiene dos opciones de respuesta -acertar o fallar,- pero, de hecho, todo el mundo falla, el ítem no tiene interés para esa muestra. Se dice que el ítem no discrimina entre los sujetos. Es un ítem que no ofrece información y por tanto, en general, para esa muestra no es útil.

5. El análisis de la dificultad

En un ítem con respuesta verdadera valorado dicotómicamente el análisis de la dificultad del ítem constituye el apartado esencial del análisis de la distribución de puntuaciones. En estos items, la distribución de las puntuaciones de los sujetos en el ítem, dado que estas consisten esencialmente en aciertos o fallos, puede ser analizada desde el punto de vista de lo difícil que el ítem resulta a esa muestra de sujetos.

La proporción de sujetos que aciertan un ítem dicotómicamente valorado se conoce como "ID" ***índice de dificultad*****,** o valor "p". Un concepto sencillo pero importante en la aplicación de la teoría clásica de tests a items valorados dicotómicamente que tienen respuesta verdadera.

Equivale a la suma de puntos que los N sujetos han obtenido en el ítem partido por N. Equivale al número de aciertos que se han registrado en ese ítem partido por el número N de sujetos. Este índice, dado que en estos items

cada acierto vale un punto y cada fallo o no respuesta un cero, equivale a la media del ítem.

El ID, o "p" va de 0 a 1, cuando vale 1 significa que el ítem es tan fácil que lo ha acertado todo el mundo. Cuando vale 0 que es tan difícil que nadie lo ha acertado. Si vale 0,5 lo ha acertado la mitad de la muestra y el ítem tiene exactamente una dificultad media.

El ***valor "q"*** representa 1-p, es la proporción de sujetos que no han acertado el ítem.

El valor "q" va de 0 a 1, cuando vale 1 significa que el ítem es tan difícil que no lo ha acertado nadie; cuando vale 0 que es tan fácil que lo ha acertado todo el mundo. Si vale 0,5 lo ha acertado la mitad de la muestra.

Puede observarse que, en realidad, ID=p más que un índice de "dificultad" es un índice de facilidad, dado que, cuanto mayor es p más fácil es el ítem. El valor q puede considerarse que expresa mejor la dificultad del ítem dado que es la proporción de sujetos que no aciertan. Cuanto más difícil es un ítem mayor es el valor q.

La *varianza de un ítem dicotómicamente valorado* puede calcularse con la fórmula general de la varianza, pero también puede demostrarse fácilmente que para items dicotómicamente valorados esa varianza es igual a "p" por "q".

Por ejemplo, un ítem con respuesta verdadera se ha administrado a 250 personas y lo han acertado 87, valorando un acierto con un punto y el fallo o la omisión con cero. Entonces, p=87/250=0,348, siendo q=1-p=0,652. La media es también 87/250=0,348. Es decir, el 34,8% de la muestra ha acertado el ítem. No lo han acertado el 65,2% restante. El I.D.=0,348, indica la proporción de personas que aciertan el ítem. La varianza del ítem es pq=0,226896.

La varianza de un ítem dicotómicamente valorado es máxima cuando p=q=0,5, es decir, cuando la mitad de los sujetos han acertado el ítem. En ese caso la varianza vale el cuadrado de 0,5, es decir, 0,25. Por tanto la varianza máxima de un ítem dicotómicamente valorado es 0,25.

En Yela, (1984; pag. 62 y siguientes) puede verse una detallada explicación intuitiva y también una explicación formal de este tope. El lector puede comprobar inmediatamente que para cualquier valor de p distinto de 0,5 la varianza disminuye.

Sea, por ejemplo, un ítem administrado a 676 personas y acertado por 338, es decir, por la mitad de ellas. p=338/676=0,5. Entonces q=1-p=0,5. Y la varianza es pq=0,25. Se dice que este ítem discrimina máximamente

entre sujetos, es capaz de diferenciar a la mitad de personas de la otra mitad.

Por lo general nos interesan items que tengan gran variabilidad, de modo que muestren las diferencias entre los sujetos, que permitan ver que las personas difieren en la variable que está siendo medida.

Una conclusión de este razonamiento es que, en términos generales estaremos interesados en disponer de items que presenten un ID próximo a 0'50, que son los que más discriminan en términos generales.

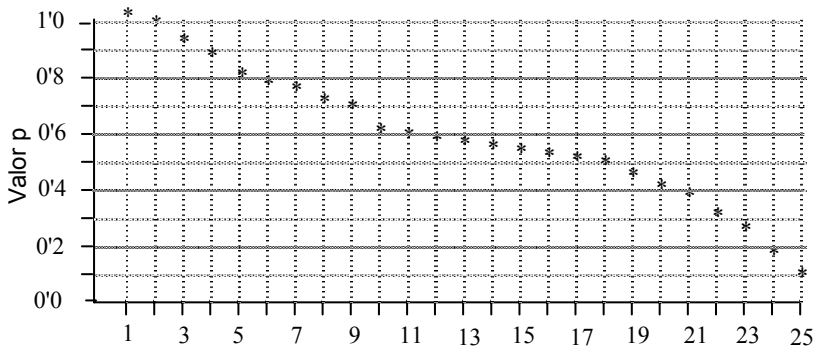
Sin embargo, debe tenerse en cuenta que la dificultad de un ítem depende de la capacidad de la muestra. Si la muestra es "brillante" en la materia evaluada, entonces los sujetos tenderán a acertar más, lo índices p serán altos y los items aparecerán como fáciles. Si por el contrario la muestra no es muy hábil en aquello evaluado, los errores y omisiones serán más frecuentes y los mismos items aparecerán con unos valores p más bajos, es decir, los items aparecerán como más difíciles. De este modo en la Teoría Clásica de Tests la dificultad de un ítem no sólo depende del ítem. Inevitablemente es relativa al comportamiento de la muestra en que se evalúe. Si la muestra es efectivamente una buena muestra de la población bajo estudio este efecto es menos perverso de lo que puede aparecer a primera vista, aunque se ha señalado esta cuestión reiteradamente como uno de los puntos débiles de la Teoría Clásica de Tests.

Esta cuestión nos introduce en la problemática de cómo elegir los items en función de su dificultad, cuestión que se puede estudiar con mayor claridad analizando el perfil de dificultad de un test.

6. El perfil de dificultad y el perfil "q" del test

Se denomina *perfil de dificultad del test* al gráfico que representa el índice de dificultad (ID=p=media) de cada uno de sus items dispuestos correlativamente sobre abscisas, en el orden en que aparecen en el test.

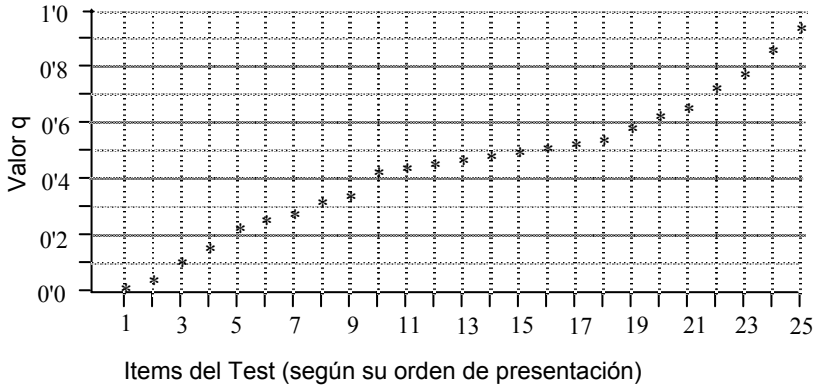
Perfil de dificultad de un test con 25 items.



El gráfico anterior muestra un ejemplo de perfil de dificultad de un test con 25 items. En el ejemplo, los primeros items son muy fáciles y los últimos muy difíciles.

Se denomina ***perfil q de un test*** al gráfico que representa el valor q (q=1-p) de cada uno de los items, dispuestos sobre abscisas en el orden en que se presentan en el test..

Perfil q del mismo test.



El perfil de dificultad y el perfil q, ofrecen la misma información pero en sentido opuesto.

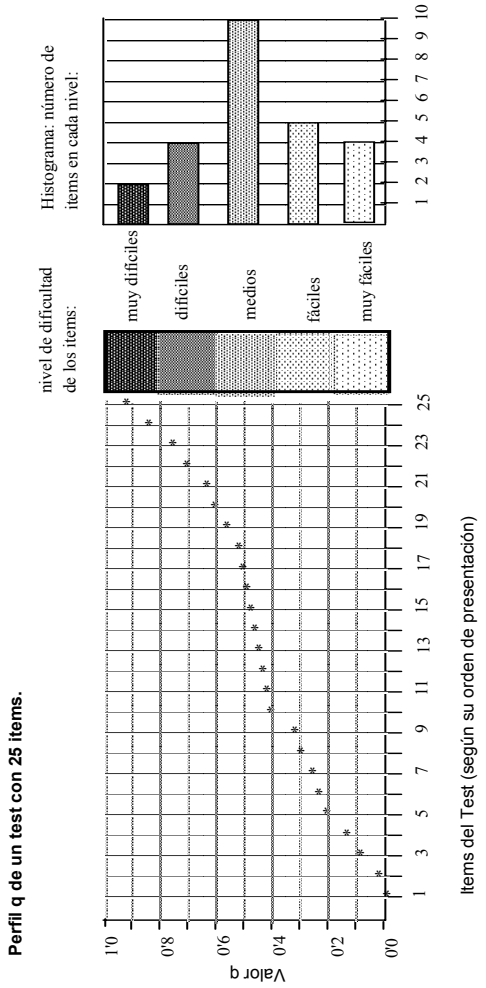
Por lo general en las pruebas de potencia, como puede verse en los perfiles anteriores, los items se ordenan de tal forma que van aumentando su dificultad progresivamente. El primero, los dos primeros, o, a lo sumo, los cuatro primeros items, se denominan "*items de arranque*", y cumplen unas funciones especiales que analizaremos después. El resto de los items presentan paulatinamente más dificultad (menos valor p ó más valor q) y están destinados a medir a los sujetos.

Si el test recorre todo el rango de dificultades, como en el ejemplo anterior, los últimos items serán muy difíciles y sólo

los muy capacitados podrán solucionarlos correctamente (efectos de azar aparte).

Un perfil típico presentará más items en la zona intermedia, cerca de p=0,5 que es, como hemos visto, la zona con más capacidad de discriminar entre personas. En ese caso, el perfil de dificultad y el perfil q reflejarán una meseta en la zona media. La ilustración de la página siguiente muestra un perfil q característico de un test con 25 items diseñado para un propósito general. Puede apreciarse que la mayoría de items están alrededor de un nivel medio de dificultad.

El perfil de dificultad del test (o su perfil q, tanto da) puede ser planificado y estructurado de otros modos, según los propósitos particulares del test.



Por ejemplo, el perfil q puede no alcanzar los valores más altos de dificultad y presentar una meseta en valores bajos, por debajo de q=0,5. Si queremos un test que discrimine con fineza en una muestra de sujetos cuya capacidad deja por debajo de sí aproximadamente a 1/3 de la población general, entonces necesitaremos un test con una amplia meseta en torno a q=0,33 calculado q sobre la población general, y no necesitaremos items por encima de q=0,6 ó q=0,7. Cuando apliquemos ese test a la muestra descrita de sujetos, si recalculamos en esta muestra el perfil q veremos que la meseta se describe ahora en las proximidades de q=0,5.

Si, al revés, estamos interesados en discriminar bien entre los más capaces, pongamos entre sujetos que en la población general dejen por debajo de sí entre el 85 y 95% del grupo normativo, entonces necesitaremos muchos items a ese nivel de dificultad. El perfil q arrancaría con un par de items en un nivel bajo para la muestra general, q=0,20 ó q=0,30 que cumplirían para esta muestra el papel de items de arranque. Inmediatamente el perfil q se situaría por encima de q=0,5 y la mayoría de los items habrían de ubicarse alrededor de q=0,87. Es decir, items muy difíciles, para discriminar con detalle el nivel de capacidad entre sujetos muy capaces en esa variable.

Si después de aplicar la prueba a esta muestra especial de sujetos recalculásemos el perfil q con sus datos, veríamos que la meseta alrededor de q=0,87 en la muestra general significaría para ellos, aproximadamente, una meseta alrededor de q=0,5.

La conclusión de esta discusión es que el nivel de dificultad de los items, para conseguir una descripción lo más precisa posible del nivel de capacidad de los sujetos, ha de estar en consonancia con su nivel de capacidad.

Se obtendrá la máxima discriminación si situamos el nivel de dificultad q de la mayoría de los items alrededor del nivel de capacidad de los sujetos. Dicho de otro modo, para una muestra determinada se conseguirá la mayor discriminación si la meseta del perfil q se sitúa en torno a q=0,5 *para los datos de esa muestra.*

7. La escala relativa de dificultad-capacidad

Los conceptos que presento a continuación sobre una escala relativa de dificultad-capacidad, la probabilidad de acertar un ítem y una función para describir esa probabilidad, van más allá de la Teoría Clásica de Tests y constituyen un intento de reinterpretación singular de

algunos conceptos de la misma, realizada con la ventaja de conocer desarrollos similares en esta linea realizados por la Teoría de la Respuesta al Item.

Todos los conceptos siguientes tienen sus análogos en Teoría de la Respuesta al Item, donde se pueden alcanzar algunos desarrollos que no son posibles desde este nuevo replanteamiento y presentación de conceptos propios de la Teoría Clásica de Tests. En cierto modo este apartado y siguientes pueden verse como un ejercicio didáctico para aproximar algunos conceptos de la TRI desde la TCT.

La idea que subyace a estos conceptos es que la dificultad del ítem y la capacidad del sujeto están relacionadas de forma que, cuanto más capaz es una muestra más fáciles aparecen los items cuando son analizados con los datos de esa muestra. Y al revés, cuanto más fáciles son unos items más capaz tiende a aparecer la muestra evaluada con ellos.

Ni capacidad ni dificultad son conceptos absolutos, tal como los estamos manejando. Ambos se expresan en función de la posición relativa de los sujetos (en percentiles) y de los items (en valores q) en una muestra. Si un sujeto s está en el percentil 65 es que supera al 65% de la muestra. Si un ítem i está en q=0,65 es que no ha sido superado por el 65% de la muestra.

Intuitivamente puede esperarse que, en un test en que los items están ordenados de fáciles (q bajo) a difíciles (q alto) un sujeto comenzará acertando items y seguirá acertando items aproximadamente *hasta que* estos estén por encima de su nivel de capacidad.

Veamos el proceso más de cerca. Si los dos primeros items de arranque son superados por el 100% de la muestra es obvio que nuestro sujeto s los pasará.

Si un ítem es tan fácil que solo lo falla el 20% de sujetos menos capaz de la muestra, entonces nuestro sujeto s, que es tan o más capaz que el 65% de la muestra, razonablemente lo acertará. Decimos que la capacidad del sujeto 65% es mayor que el valor q 20% (la dificultad) del ítem.

Esta clase de razonamientos presupone que sujetos e items son ordenables, lo que implica aunque sólo sea aproximadamente que la mayoría de sujetos que aciertan 65 items, por ejemplo, tienden a acertar los mismos 65 y a fallar los mismos 35. Y desde el punto de vista de los items, que si al item 10, por ejemplo, lo aciertan el 65% de la muestra y al 9 el 60% este último 60% está contenido, al menos aproximadamente, en el 65% anterior. Si sujetos e items no son ordenables el concepto de escala subyacente no se sostiene.

En el otro extremo, si un ítem es tan difícil que lo falla el 90% de la muestra, es decir, que solo lo acierta el 10% de sujetos mejores, entonces sólo ese 10% de sujetos que lo hacen mejor que el 90% restante lo acertará. Si nuestro sujeto solo es mejor que el 65% peor de la muestra, entonces, razonablemente, -dejando a un lado el azar y la

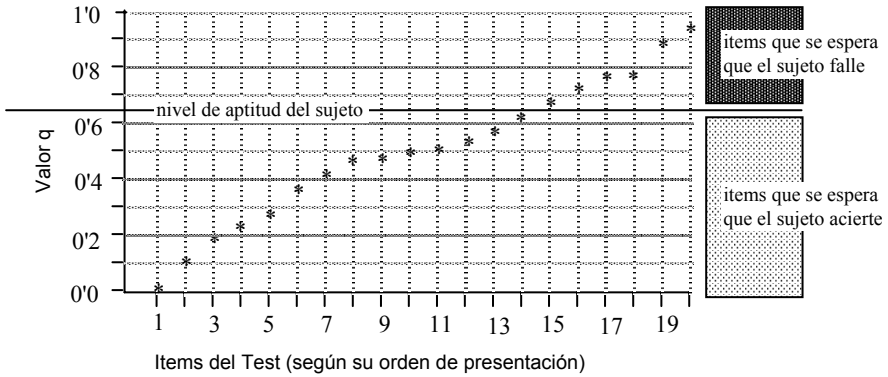
inspiración- no acertará ese ítem. Decimos que la dificultad del ítem 90% es superior a la capacidad del sujeto 65%.

Vamos al punto crítico: Si un ítem es tan difícil que lo falla el 65% menos capaz de la muestra, entonces hará falta un sujeto que deje por debajo de sí al menos al 65% menos capaz de la muestra (percentil 65 aproximadamente) para acertarlo. De otro modo, un sujeto que deje por debajo de sí al 65% menos capaz de la muestra estará en condición de acertar, aproximadamente, todos los items que estén por debajo de su nivel de capacidad, es decir, todos los items que supera ese 65% menos capaz. Por eso se espera que un sujeto en nivel 65% acierte items de hasta un nivel 65% (valor q), aproximadamente.

En términos prácticos, se espera que, aproximadamente, un sujeto supere todos los items cuyo valor q este por debajo de su nivel k, siendo k el percentil que ocupa su puntuación total.

El gráfico siguiente muestra los razonamientos anteriores sobre el perfil q del test de 20 items que había presentado antes.

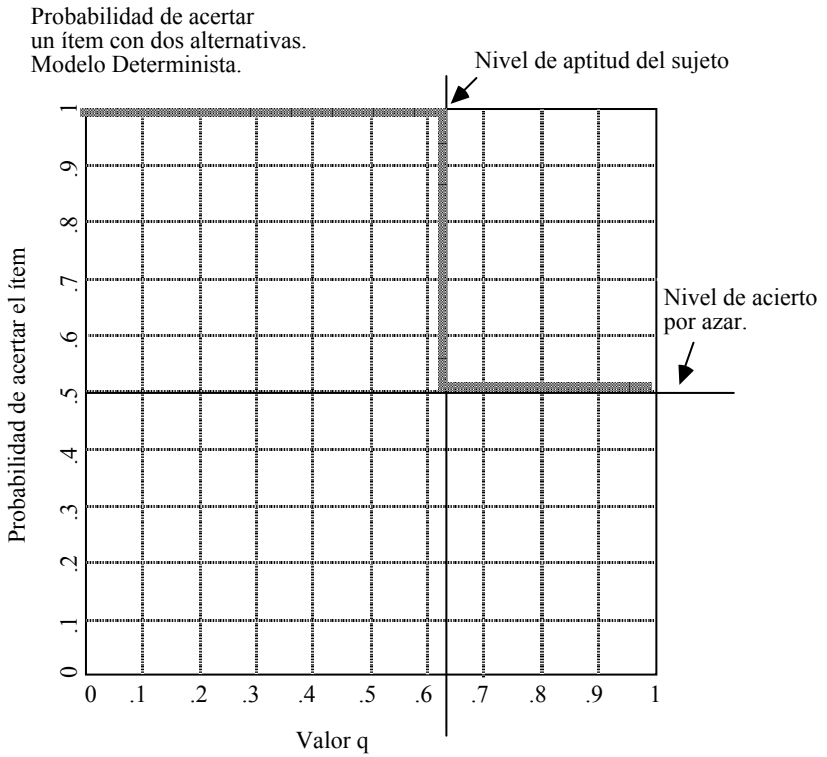
Perfil q de un test con 20 items.



En realidad, esperamos variaciones aleatorias sobre este pronóstico razonable, de modo que, naturalmente, la persona fallará algunos items por debajo de su nivel de capacidad (es decir, items teóricamente fáciles para él) y quizás acierte algún o algunos items por encima de su nivel de capacidad (es decir, items teóricamente muy difíciles para él).

8. La probabilidad de acertar un ítem

Si los items formaran un escalograma perfecto entonces el nivel k de aptitud o capacidad del sujeto determinaría completamente la probabilidad de acertar un ítem.



Los items formarían un escalograma perfecto si que un sujeto acertara un ítem significara, inevitablemente, haber acertado todos los items más fáciles a ese.

En ese caso, todo ítem cuyo valor q estuviese por debajo de ese nivel k sería acertado.

Si el sujeto contestase al azar el resto de los items y estos tuviesen solo dos opciones de respuesta (como en una prueba verdadero-falso) entonces todos los items cuyo q fuera mayor que k tendrían una probabilidad de 0,5 de ser acertados, como muestra el gráfico siguiente.

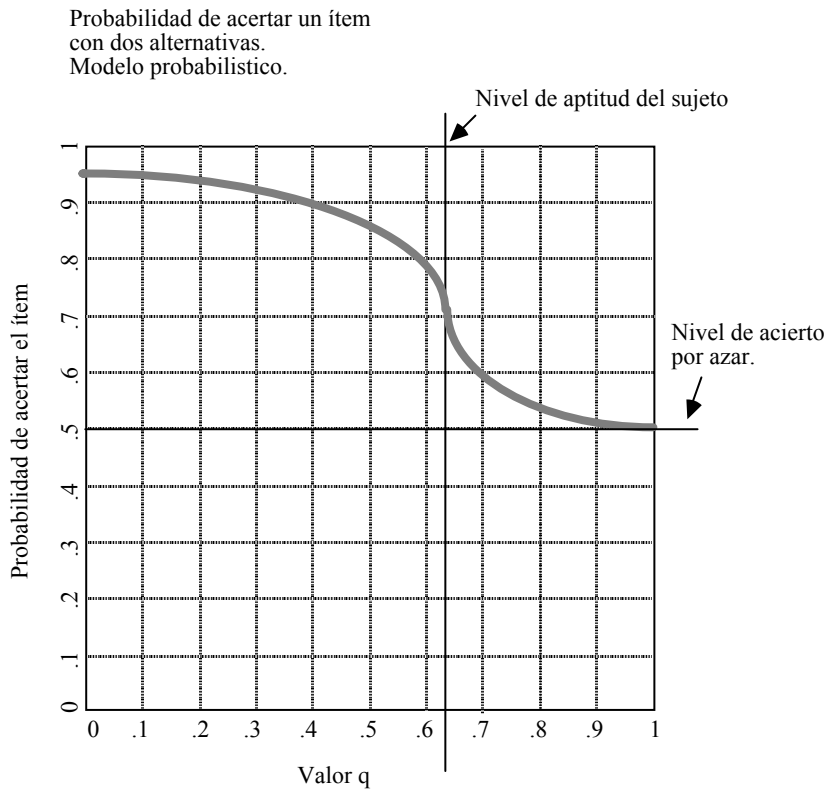
Si el sujeto no contestase ningún ítem que no supiese, la probabilidad de acertar esos items difíciles para el sujeto sería entonces 0.

Sin embargo, como hemos puesto de manifiesto, no puede sostenerse un modelo determinista como éste.

Precisamente, si hemos de aceptar que toda medición es una aproximación más o menos inexacta a lo medido, o, en términos clásicos más precisos, que toda medición conlleva un error de medida, entonces es más razonable proponer un modelo probabilístico para explicar las respuestas de los sujetos a los tests.

No puede, de hecho, sostenerse un modelo determinista, como el de la figura anterior. En la realidad, los sujetos fallan items teóricamente fáciles para ellos y aciertan otros supuestamente difíciles para ellos, considerada su capacidad en función de su posición relativa en la muestra.

Los datos reales contradicen, en general, un modelo determinista.



Para explicar que los sujetos fallen items fáciles para ellos (algunas veces) y acierten items teóricamente difíciles para ellos (algunas veces) puede proponerse un modelo probabilístico como el que muestra la figura siguiente.

Puede apreciarse que incluso para items tan fáciles que q=0 no se asegura que el sujeto acertará (la probabilidad de acierto a lo largo de los niveles de dificultad no parte de 1). Según nos vamos acercando al nivel tope de capacidad del sujeto su probabilidad de acertar va descendiendo, pero no acaba totalmente al llegar al punto de su nivel de capacidad. En lugar de esto, disminuye admitiendo que existe una probabilidad de acertar items más allá de su nivel teórico de capacidad.

9. Una función que describe la probabilidad de acertar el ítem en función de su dificultad

La curva del gráfico anterior refleja la probabilidad que un sujeto de una capacidad determinada k tiene de acertar diversos items en función del valor q de esos items.

Una curva como la del gráfico anterior puede denominarse como *curva característica del sujeto* de nivel k.

El significado de la curva es que, conocido el nivel k del sujeto, puede pronosticarse razonablemente el número de aciertos que producirá en items de cada nivel q. Es decir, la

curva característica de un sujeto permite pronosticar sus resultados en el test en función de la dificultad de los items.

¿Cómo podría estudiarse empíricamente si el modelo descrito por la curva del gráfico anterior es adecuado? Es relativamente sencillo, y existen varias alternativas.

Si se dispone de una persona cuya capacidad k es conocida (en el ejemplo del gráfico esa capacidad es aproximadamente 0,65 ó 65% en términos de porcentajes) con disponer de una muestra adecuada de items de cada nivel q podemos establecer cual es la proporción de items que esa persona resuelve en cada nivel q. Esa proporción podría utilizarse para estimar la probabilidad de un sujeto de capacidad k de resolver items de cada nivel q. Es decir, esa proporción para cada nivel q habría de describir una curva semejante a la del gráfico anterior.

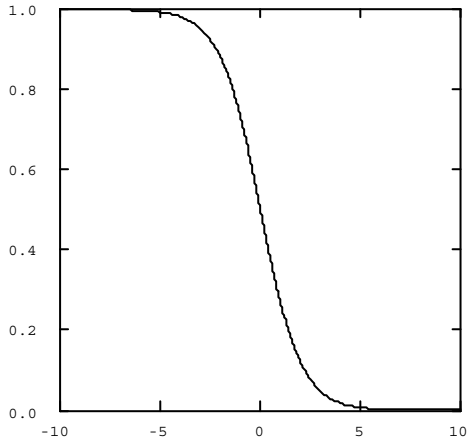
Otra opción es disponer de uno (o muy pocos items) de cada nivel q, y disponer de una muestra adecuada de N personas con el mismo nivel k. La proporción de personas que acertarán los items de cada nivel q podría utilizarse para estimar la probabilidad de un sujeto de nivel k de acertar items de cada nivel q.

La curva del gráfico anterior resultará familiar a quien conozca la curva que traza una función de curva normal acumulada, o, con otra función más sencilla, una función logística.

Por ejemplo, la siguiente función logística:

$$P = \frac{1}{1 + EXP(X)}$$

permite describir la siguiente curva, escalada para X entre -10 y +10:.



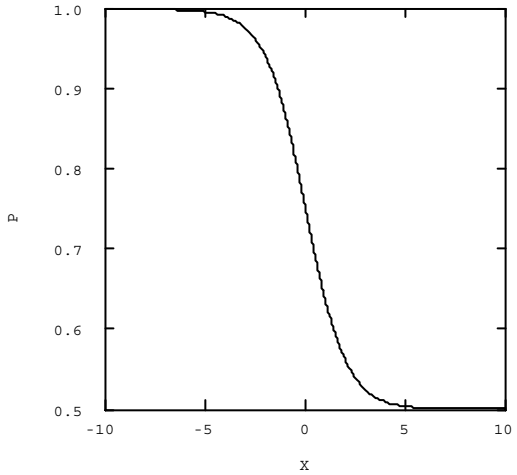
En la escala de dificultad X (eje de abscisas) el valor medio 0 representa la capacidad del sujeto. Esta función es muy sencilla, define la probabilidad P (ordenadas) en función de la dificultad X (abscisas), pero asume que la probabilidad de acertar para un ítem muy difícil (en el gráfico para X aproximadamente mayor a 5) es 0.

Si asumimos que los items pueden ser acertados por azar con una probabilidad igual a 0,5, se le puede añadir un parámetro a la función para determinar que el nivel mínimo

de probabilidad esté en 0,5, aproximándonos a la curva teórica que describíamos inicialmente. Llamaremos a este parámetro de suelo de función *parámetro c*, (por analogía con la Teoría de la respuesta al ítem). La función quedaría así:

$$P = 1 - (0'5 - \frac{1 - 0'5}{1 + EXP(X)})$$

La función anterior es un ejemplo de un tipo de función logística de un parámetro. La curva descrita por la ecuación anterior sería esta:

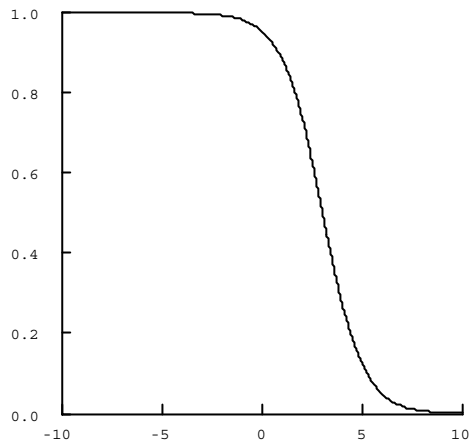


(Obsérvese que en la gráfica anterior el eje de ordenadas comienza ahora en 0,5).

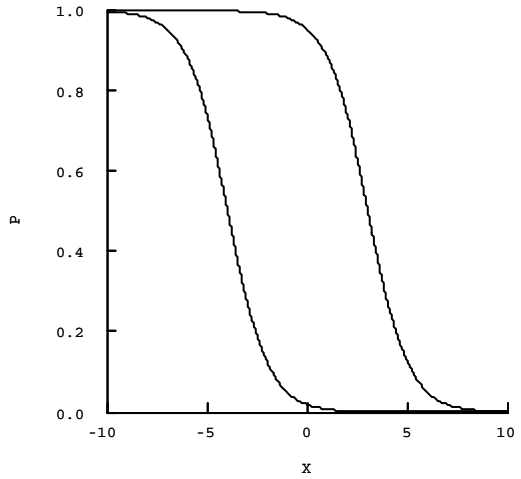
La función sigue "centrando" el punto de inflexión -que representa el nivel k del sujeto en esta escala- en 0. Si deseamos poder reflejar que diversos sujetos pueden tener diversas capacidades k y por tanto diversos puntos de inflexión necesitamos un parámetro más en la función logística. Lo denominaremos *parámetro b*, (por analogía con la Teoría de la respuesta al ítem). Este parámetro b significa la localización de la curva a lo largo del eje de abscisas. En esta curva característica de un sujeto de capacidad k el parámetro b representa precisamente ese nivel de capacidad k, escalado en el gráfico, por comodidad, entre -10 y +10. (Podría reescalarse entre 0 y 1, como nuestra escala inicial de valores q, sin perdida de generalidad).

En la función siguiente no se ha introducido un parámetro c, por sencillez, pero se ha introducido un parámetro b=-3.

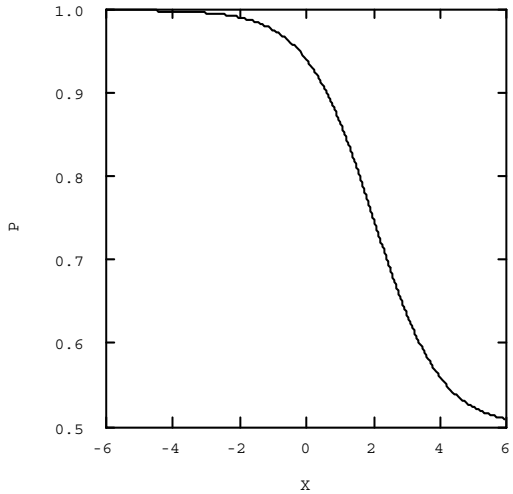
La función resultante es P=1/(1+EXP(X-3), y su gráfica:



En la gráfica siguiente se ha representado la curva de un sujeto con b=-3 y, además, la de otro sujeto con b=4 (cuya función sería P=1/(1+EXP(X+4))) para que se aprecie el efecto de este *parámetro b*. (Las curvas están centradas sobre -3 y +4)

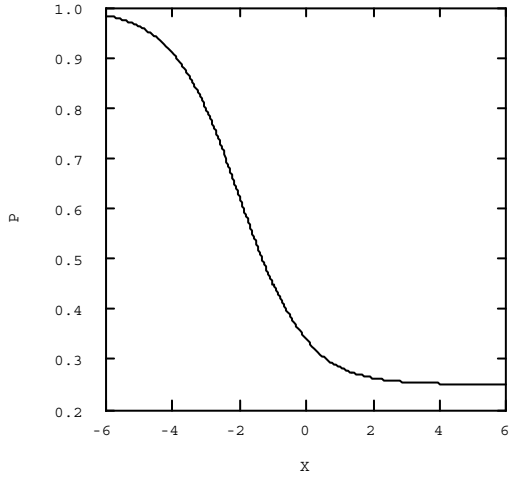


Combinando ambos parámetros, c y b, podemos obtener funciones como esta P=1-(0.5-((1-0.5)/(1+EXP(X-2)))) que producen curvas así (ver expresión general en página siguiente):



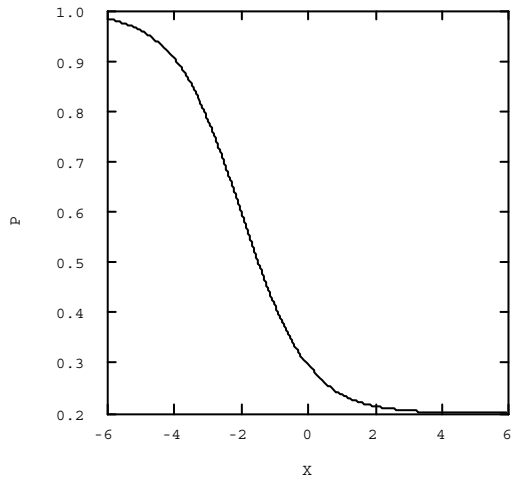
En el ejemplo anterior tenemos una función logística de dos parámetros con c=0,5 y b=2. El parámetro c coloca el suelo de la función en 0,5 y el parámetro b coloca el punto de inflexión en X=2, es decir, se supone un sujeto con k=2 (expresando k en la escala centrada en 0 que estamos usando en los gráficos).

Para una función P=1-(0.75-((1-0.25)/(1+EXP(X-(-2))))) tendríamos una curva (expresión general en página siguiente):



El ejemplo anterior puede representar la curva para un ítem de 4 alternativas, donde la probabilidad de acertar al azar es 0,25. El sujeto tiene una capacidad -2.

Para P=1-(0.8-((1-0.20)/(1+EXP(X-(-2))))) tendríamos:



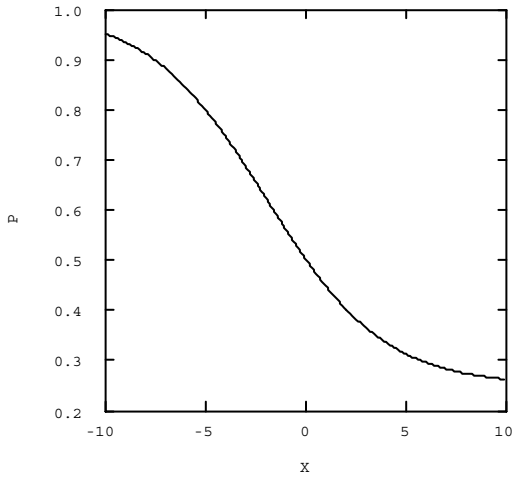
En el caso anterior se trata de un sujeto de capacidad -2 con una probabilidad de acertar por azar de 1 entre 5 (es decir, c=0,2).

Las funciones anteriores son casos particulares de la función logística de dos parámetros, y responden a la fórmula general:

$$P = 1 - ((1 - c) - \frac{1 - c}{1 + EXP(X - b)})$$

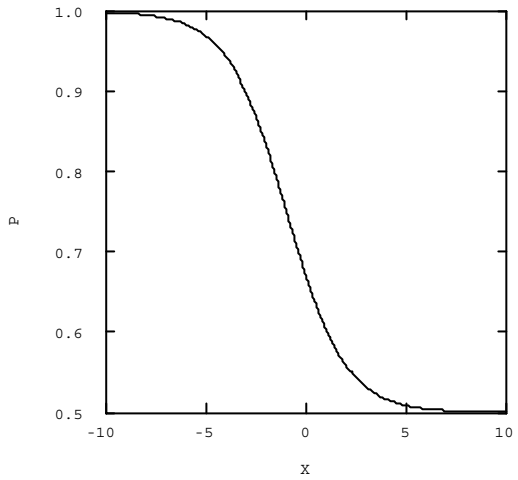
Se puede complicar más la función, introduciendo un tercer *parámetro a* (también por analogía con la Teoría de la Respuesta al ítem). El parámetro a permite reflejar diversas inclinaciones de la curva, es decir, si la probabilidad de fallar decrece muy aprisa o muy despacio en función de la escala de dificultad X (abscisas).

Por ejemplo con a=0,2, c=0,25 y b=-2 la función logística P=1-(0.75-((1-0.25)/(1+EXP(1.7*0.2*(X-(-2))))))
representaría una curva:



(En la fórmula anterior 1,7 es una constante auxiliar que facilita el ajuste de la fórmula).

Si se compara el ejemplo anterior con el que produce la función P=1-(0.5-((1-0.5)/(1+EXP(1.7*0.4*(X-(-1)))))) podrá apreciarse como un parámetro a=0,4 produce una curva más inclinada que el parámetro a=0,2 del ejemplo anterior.



Cuanto más inclinada sea la curva mejor podríamos predecir las respuestas del sujeto a partir de la dificultad de los items.

Si consideramos simultáneamente los parámetros a(inclinación), b(nivel k) y c(nivel de acierto por azar) dispondremos de una función logística de tres parámetros, análoga a la que Lord(1980) ha postulado desde la tradición de la Teoría de la respuesta al ítem.

Para nuestro caso esta función logística de tres parámetros puede escribirse así:

$$P = 1 - ((1 - c) - \frac{1 - c}{1 + EXP(1'7 \cdot a \cdot (X - b))})$$

Esta familia de funciones es adecuada para representar la relación dificultad/capacidad que estamos analizando.

Para un sujeto dado, ¿cómo podríamos obtener empíricamente una aproximación al valor de los parámetros a, b y c que describen su curva? Los métodos empíricos que hemos sugerido antes permitirían una aproximación a esta estimación desde un enfoque operativo fuera de la Teoría de la respuesta al ítem (aunque no podemos decir desde un enfoque clásico puesto que todos los métodos, razonamientos y fórmulas que he propuesto en estos apartados no forman parte desde luego del "corpus" de la teoría clásica). La Teoría de la respuesta al ítem por su parte se ha ocupado en elaborar sofisticados métodos de estimación de estos parámetros que caen fuera del objeto de este texto.

10. Perfil de dificultad del ítem en función de la capacidad de los sujetos

El perfil de dificultad del ítem en función de la capacidad de los sujetos es una gráfica que señala para cada puntuación total en la prueba cual es el ID obtenido. Hay un perfil de este tipo para cada ítem.

Se obtiene del siguiente modo: Supongamos que hemos administrado una prueba objetiva de 20 items a 250

estudiantes. Cada estudiante puede obtener una puntuación entre 0 y 20. (A efectos de este análisis la puntuación total de un estudiante la obtenemos como simple suma de sus aciertos, sin contar errores ni omisiones).

Agrupamos a los estudiantes en submuestras, según su total. Por ejemplo, hacemos un grupo con aquellos que tienen puntuaciones totales de 1 a 4, otro con los que tienen totales de 5 a 9 etc.

Supongamos que los sujetos se distribuyen en esos grupos conforme a la siguiente tabla:

X	Frec.
1-4	25
5-8	50
9-12	100
13-16	50
17-20	25

La tabla se lee del siguiente modo: 25 sujetos han obtenido una puntuación entre 1 y 4; 50 sujetos han obtenido un total entre 5 y 8, etc.

Supongamos ahora un ítem i en el que los sujetos hayan obtenido los siguientes resultados:

X	Frec.	Aciertan	D del grupo:
1-4	25	5	0'2
5-8	50	20	0'4
9-12	100	60	0'6
13-16	50	40	0'8
17-20	25	22	0'88
Totales:	250	147	0'58

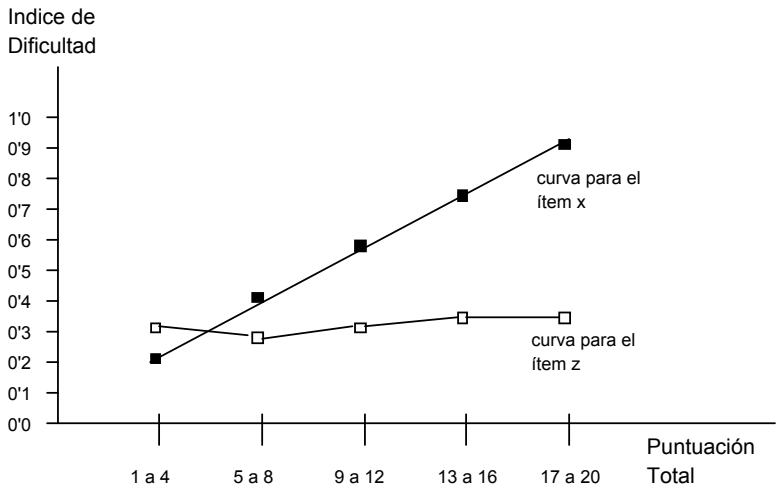
La tabla se lee así: De los 25 sujetos que han obtenido un total en la prueba entre 1 y 4 puntos, solo 5 han acertado el ítem, es decir, para ese grupo el ID es 5/25=0'2.

Como puede apreciarse, la proporción de sujetos que aciertan el ítem del ejemplo crece según sea la puntuación total, de modo que el ítem es más acertado por los mejores en el test y menos acertado por los peores en el test. Aunque no es necesario un crecimiento lineal del ID como el del ejemplo, los items razonables funcionan aproximadamente de este modo: a mayor nivel de capacidad de los sujetos más fácil resulta el ítem.

La gráfica de perfil de los totales de un ítem se obtiene representando el ID obtenido para cada grupo establecido según el total.

En la gráfica siguiente se representa el perfil de dificultad de dos items a través de 5 grupos de sujetos determinados en función de su capacidad, según el total del test.

Perfil de Indices de Dificultad para grupos formados según su total en el test.



■ Perfil de Indices de Dificultad que obtiene un ítem x hipótetico en cada uno de los grupos de sujetos con diferente total en la prueba. Este ítem muestra un comportamiento razonable: Cuanto mayor es la capacidad de los sujetos (mayor total han obtenido), más proporción de sujetos aciertan el ítem.

□ Perfil de un ítem z que no funciona adecuadamente y debe ser revisado. Obsérvese que este ítem lo acierta la misma proporción de sujetos en cada grupo, lo que no es razonable. Tanto el grupo de los sujetos muy capaces como los poco capaces obtienen los mismos resultados en el ítem. Aproximadamente un 30% aciertan el ítem independientemente de su nivel de capacidad, por tanto el ítem no refleja la capacidad de los sujetos.

Un modo más riguroso de tratar el eje de abscisas de la gráfica anterior es definir una escala de capacidad donde los sujetos se ubican en función de su posición k' en la muestra.

Una gráfica de este tipo presenta en su eje de abscisas la escala de capacidad relativa k, (escalada de 0 a 100, ó de 0 a 1, usualmente), y en ordenadas la dificultad del ítem, expresada por su valor p (ID) (escalado de 0 a 1, usualmente) y la denominaremos *"perfil p del ítem a través de la capacidad*

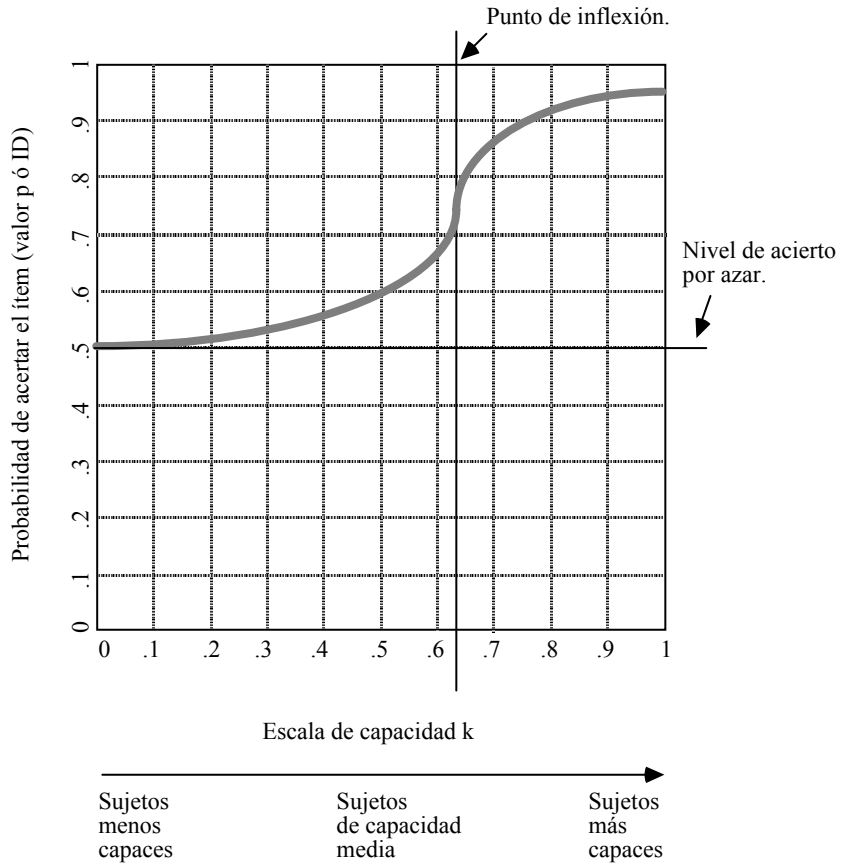
Una gráfica que presente en abscisas la escala de capacidad relativa k y en ordenadas el valor q del ítem, forma un *"perfil q del ítem a través de la capacidad*.

Para realizar una gráfica así (perfil p o q del ítem) hay que disponer de suficientes sujetos en cada nivel k, lo que no siempre es posible, especialmente con muestras o grupos normativos locales que suelen ser relativamente pequeños.

Posiblemente para comenzar a tener una aproximación a una descripción razonable del ID del ítem en cada punto de la escala de capacidad (sujetos que ocupan la misma posición percentil) necesitaríamos al menos unos 10 sujetos por punto. Si los sujetos se distribuyeran homogéneamente a lo largo de la escala de aptitud, con una escala de 100 puntos, una muestra de unos 1000 sujetos sería suficiente. Sin embargo, los sujetos no se distribuyen homogéneamente en la mayoría de las variables de interés psicológico. Más bien en la mayoría de variables, las distribuciones, cuando N es muy grande, suelen presentar muchos casos concentrados en una zona media más popular y relativamente pocos casos en las colas. En estas condiciones, si la escala de capacidad se define sobre puntuaciones directas, como en la gráfica anterior, una muestra de 1000 sujetos produciría una excelente descripción de los valores centrales y una más pobre descripción de las colas. Sin embargo, si el eje de abscisas representa una escala de capacidad k donde cada nivel k es un percentil, (supuesto un test con suficiente número de items y suficiente variabilidad) entonces, queda garantizado que de cada valor k a cada valor k hay exactamente un 1% de la muestra. En términos de 1000 sujetos de k a k hay 10 sujetos, lo que permite una descripción más adecuada.

Un modo más realista (aunque menos preciso) de efectuar el análisis cuando no se dispone de tantos sujetos consiste en agrupar los sujetos en submuestras en función de su nivel k, de modo que en cada intervalo definido en k aparezcan los suficientes sujetos (al menos unos 10).

Probabilidad de acertar un ítem concreto con dos alternativas para sujetos con diferentes grados de capacidad. Modelo probabilistico.

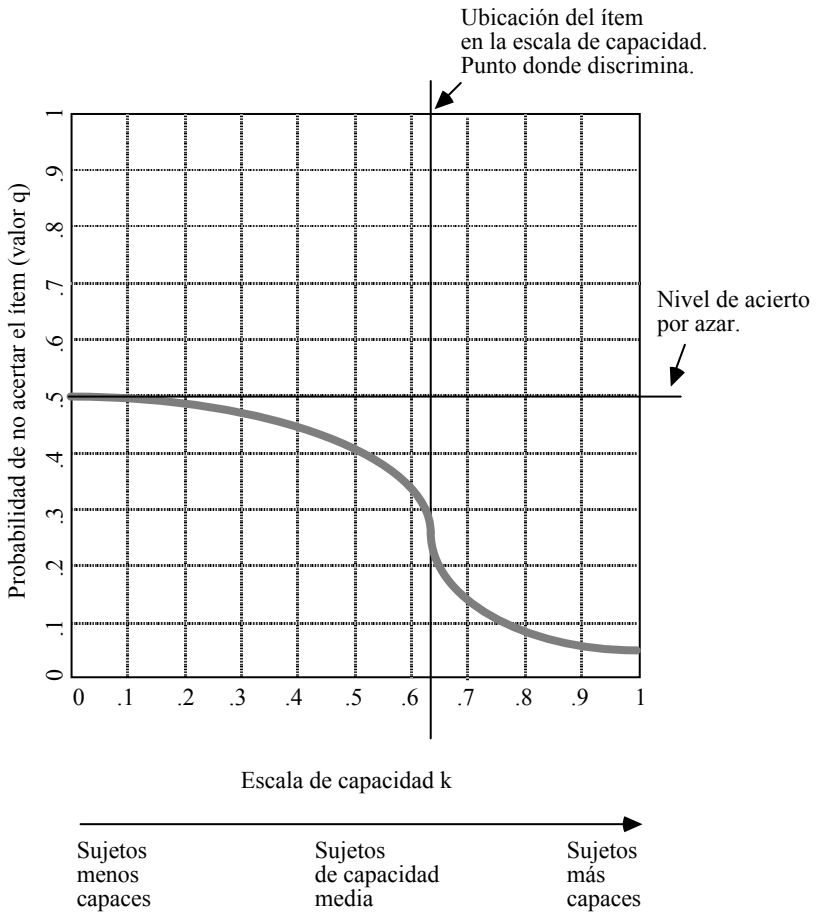


El gráfico anterior muestra la relación entre la probabilidad de acertar el ítem (valor p ó ID) y el nivel de capacidad de los sujetos descrito por su posición k en el grupo.

Se presenta un modelo probabilístico donde la probabilidad de acertar el ítem crece en función de la capacidad de los sujetos. La curva crece lentamente presentando un punto de inflexión allí donde la capacidad de los sujetos empieza a estar por encima de la dificultad del ítem.

Esta misma relación puede expresarse en función del valor q del ítem en cada nivel de capacidad, como se muestra en el gráfico siguiente.

Probabilidad de no acertar un ítem concreto con dos alternativas para sujetos de diferente nivel de capacidad.
Modelo probabílistico.



Las funciones logísticas pueden utilizarse también aquí para describir muy adecuadamente la probabilidad de acertar (o no acertar) un ítem en función de la capacidad de los sujetos.

De modo análogo al que nos ha permitido construir y completar paulatinamente una función adecuada para describir la curva característica del sujeto es posible elaborar una función logística que describa la *curva característica del ítem*.

En la Teoría de la Respuesta al Item, la curva característica del ítem refleja la probabilidad de acertar un ítem dado (ordenadas) en función del nivel de capacidad de los sujetos (abscisas). En la presentación anterior la estimación de la aptitud se ha restringido a los medios limitados de la Teoría Clásica de Tests, pero introduciendo los conceptos de la TRI a efectos didácticos. La TRI inicia sus razonamientos, precisamente, a partir de la consideración de esta curva característica del ítem. La curva característica del ítem es un concepto central de la TRI; sin embargo la TCT tradicionalmente no ha centrado su atención en este tipo de curvas.

Sería posible empíricamente describir si un modelo como el anterior (o en su caso qué otro modelo de curva) ajusta a los datos de una muestra. Como una aproximación desde la TCT, sin estimar la aptitud latente como en TRI, bastaría

con ubicar a los sujetos por su nivel k y obtener los valores p (ó q) del ítem para cada nivel k.

La TRI ha desarrollado poderosos métodos de estimación de la aptitud latente de los sujetos y de los parámetros a, b y c que describen una función logística capaz de representar la curva característica de cada ítem.

Hay un aspecto verdaderamente esencial en TRI de las curvas que describen a sujetos y a items en el que, intencionadamente, no vamos a entrar aquí. Ese aspecto se refiere a la constancia de esas curvas a través de muestras de sujetos y de muestras de items que pertenecen a una misma población. Si esas curvas pueden describirse de modo relativamente independiente de la muestra de sujetos o de la muestra de items, lo cual puede ser objeto de contraste empírico, entonces los parámetros que describen al sujeto o al ítem pueden considerarse, en su sentido amplio, generalizables a través de muestras de items o sujetos, una quimera que fue difícilmente perseguida fuera de la Teoría de la Respuesta al Item.

11. ¿Por qué en los tests clásicos de inteligencia o aptitudes no se escalan los items?

Para contestar esta cuestión será conveniente recapitular ordenadamente una serie de razonamientos que hemos visto anteriormente a lo largo de diversos apartados del capítulo.

La Teoría Clásica de Test fue pensada básicamente para tests de aptitudes o rendimiento donde los items tienen respuesta verdadera y se puntúan concediendo un punto cuando el sujeto acierta, obteniendo el total como suma de puntos. El prototipo de test para el que esta pensada la TCT es el test clásico de inteligencia administrado en formato de lápiz y papel, un test de fondo o casi totalmente de fondo con numerosos items que constituyen una clase de problemas más o menos semejantes entre sí.

En la elaboración de estos tests clásicamente no se escalan los items, o al menos, no se utiliza ningún valor de escala en la corrección de la prueba.

En el esquema más sencillo de puntuación, los items se valoran con un punto cuando el sujeto acierta y con 0 puntos cuando el sujeto falla o no lo intenta. La puntuación total de un sujeto se determina como la simple suma de puntos que ha obtenido (número de items que ha acertado).

El principal estadístico de un ítem de este tipo con respuesta verdadera es su índice de dificultad (o media, o "p") que indica que proporción de sujetos de una muestra adecuada han superado el ítem.

El valor q = 1-p (la proporción de sujetos que no aciertan el ítem o auténtica *dificultad* del ítem) podría considerarse por sí un valor de escala razonable para cada ítem. Característicamente un test clásico de inteligencia o aptitud es una secuencia de items con ID decreciente (o lo que es lo mismo con q creciente).

Si se escala los items por su valor q la escala va desde 0 items tan fáciles que todo el mundo acierta, hasta 1, items tan difíciles que nadie acierta. Una q de 0'75 significa un ítem razonablemente difícil que han fallado 3 de cada 4 personas.

Se denomina *perfil de dificultad del test* al gráfico que representa el ID o valor p de cada uno de sus items dispuestos ordenadamente sobre abscisas.

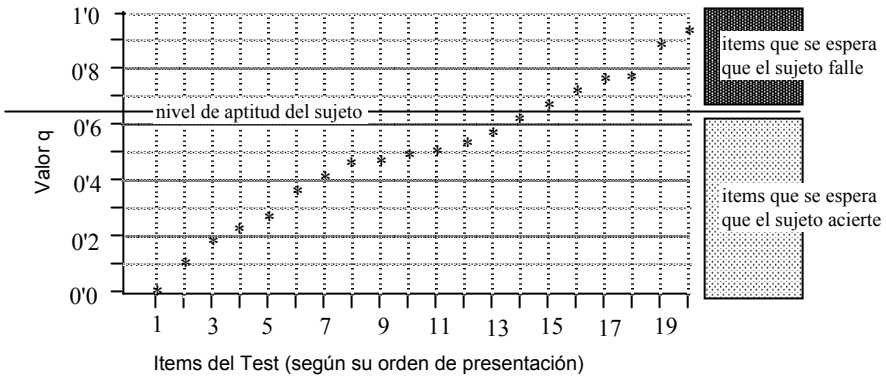
Se denomina *perfil q de un test* al gráfico que representa el valor q de cada uno de sus items dispuestos ordenadamente sobre abscisas.

La concepción del constructor del test es que el sujeto comienza acertando los items y sigue acertándolos hasta el límite de su nivel de capacidad. Hasta el punto que su nivel de capacidad es inferior al valor q del siguiente ítem. A partir de ahí (aproximadamente) comienza a fallar.

Con este modelo de razonamiento podría considerarse una estimación adecuada del nivel del sujeto el valor q del último ítem que acierta (o, en una concepción más sofisticada, el valor q del último ítem que acierta no por azar). Es decir, el valor q de aquel ítem acertado con valor q más alto podría interpretarse como la ubicación del sujeto en la escala de capacidad.

Este modo de razonar tiene algunas ventajas. Por ejemplo ubica a los items (las dificultades: q) y a los sujetos (los percentiles P_k) en la misma escala de capacidad/dificultad.

Perfil q de un test con 20 items.



Si la capacidad del sujeto, por ejemplo, deja por debajo de si al 65% de los sujetos de su muestra de referencia

(percentil 65) entonces es razonable suponer que acertará, aproximadamente, los items cuyo valor q este por debajo de 0'65. (Ver tabla de ejemplo siguiente).

Si se asume un modelo determinista (se aciertan todos los items por debajo del nivel de capacidad y se fallan todos los items por encima del nivel de capacidad) entonces la relación es exacta, de modo que podría ubicarse al sujeto en su muestra conociendo únicamente el valor q del ítem con valor q más alto de aquellos que ha contestado el sujeto. O bien al revés, si se conoce a que percentil corresponde su puntuación total queda determinado que valor q es capaz de superar.

Si se sigue este esquema de razonamiento bastará conocer el valor k (porcentaje de sujetos que la puntuación total del sujeto iguala o deja por debajo de sí) asociado a la puntuación total del sujeto para ubicarlo sobre la dimensión.

La cuestión es que q es un valor que ubica al ítem respecto a la muestra de sujetos y k ubica a un sujeto determinado también respecto a la muestra de sujetos. Si las respuestas a todos los items son perfectamente consistentes con la posición del sujeto y de los items respecto a la muestra de sujetos entonces la información de ambos valores ha de ser consistente.

Dado que el valor k es relativo a la cantidad de sujetos que aciertan los items e independiente de como se puntúen los items, entonces, desde esta perspectiva es indiferente para ubicar a los sujetos en la dimensión valorar los items asignando a cada acierto un punto o valorar los items asignando a cada acierto el valor de escala (por ejemplo q) del ítem. Se utilice el método de valoración que se utilice el sujeto obtendrá el mismo valor k (posición en su grupo).

Relación entre el valor q de una secuencia de items ordenados en función de q y el porcentaje de sujetos que aciertan hasta ese ítem bajo supuesto de escalograma en una muestra hipotética.

Nº de ítem o puntuación total en el test(*)	I.D. del ítem o valor p. Proporción de sujetos que aciertan ese ítem	Valor q = 1 - p Proporción de sujetos que no aciertan ese ítem	% de sujetos que tienen esa puntuación total en el test	% de sujetos que aciertan hasta ese ítem o Percentil para ese total en el test
1	1'00	0'00	5	5
2	0'95	0'05	5	10
3	0'90	0'10	3	13
4	0'87	0'13	8	21
5	0'79	0'21	4	25
6	0'75	0'25	3	28
7	0'72	0'28	3	31
8	0'69	0'31	2	33
9	0'67	0'33	8	41
10	0'59	0'41	1	42
11	0'58	0'42	2	44
12	0'56	0'44	1	45
13	0'55	0'45	2	47
14	0'53	0'47	3	50
15	0'50	0'50	1	51
16	0'49	0'51	1	52

17	0'48	0'52	2	54
18	0'46	0'54	4	58
19	0'42	0'58	4	62
20	0'38	0'62	3	65
21	0'35	0'65	8	73
22	0'27	0'73	6	79
23	0'21	0'79	6	85
24	0'15	0'85	8	93
25	0'07	0'93	7	100
			100	

(*) El supuesto de escalograma implica una concepción determinista donde cada sujeto acierta items hasta su nivel aptitud. Acierta todos los items hasta su nivel de aptitud y falla todos los items más allá de su nivel de aptitud. Bajo este supuesto obtener una puntuación X significa haber acertado precisamente los X primeros items, dado que estos están ordenados por su dificultad. Si eso es así entonces el valor q del ítem X determina que porcentaje de la muestra contestará hasta el ítem anterior X-1; dicho de otro modo si un sujeto en el test tiene una puntuación X correspondiente al percentil k acertara todos los items con valor q<k. Bajo un modelo probabilístico cabe esperar que suceda solo una aproximación a este principio.

Por cierto que, por esta misma razón, también resulta indiferente si las valoraciones de cada ítem se suman o se promedian para formar la puntuación total del sujeto en la escala (supuesto que no haya omisiones o que éstas se traten como errores). Para la determinación de k también es indiferente si los errores penalizan o no (supuesto que no hayan omisiones o que se traten sistemáticamente como errores), y, por tanto, resulta indiferente si se aplica o no una fórmula de corrección del azar (de nuevo bajo el supuesto de que no haya omisiones o que éstas se traten sistemáticamente como errores).

Paradójicamente es estrictamente indiferente si se valoran los aciertos con un 1 o con el valor q *si* se asume un modelo determinista, pero no es indiferente si se asume un modelo probabilista (es decir un modelo que asume error de medida y por tanto que el sujeto puede acertar items por encima de su nivel de capacidad y fallarlos por debajo de su nivel de capacidad) en cuyo caso qué items concretos ha acertado el sujeto y qué dificultad o valor q presentan no es indiferente.

El modelo de la teoría clásica de tests es claramente probabilista, pero se comporta en este asunto como si fuera determinista. Puede permitirse ignorar los valores de escala de los items en la medida en que los trata como si fueran un escalograma, pero, en primer lugar, es bien sabido en la práctica que generalmente no son, ni de lejos, un escalograma, y, en segundo lugar, este proceder es contrario al supuesto generador del modelo que establece que toda puntuación empírica contiene un componente de error aleatorio.

He de prevenir que todas estas consideraciones, desde el concepto de escala de dificultad-capacidad hasta aquí son fruto de aplicar TRI y escalas Guttman al análisis clásico de la dificultad. No conozco ningún estudio empírico que haya asumido este enfoque y, por tanto, su utilidad debería ser

contrastada antes de ser admitida, aunque, en mi opinión, los razonamientos que sustentan este enfoque no son más cuestionables que el método generalmente aceptado de ubicar a los sujetos y a los items sobre un grupo normativo.

12. Items de arranque

Como excepción a los principios anteriores acerca del nivel de dificultad recomendable para un conjunto de items, es común que el primero o los dos primeros items de pruebas de inteligencia y aptitud sean tan fáciles que todo el mundo los acierte. Estos items cumplen un papel distinto en las pruebas al de discriminar entre los sujetos. Estos *items de arranque* sirven para:

1. Afianzar al sujeto en el manejo físico del test (cuadernillo, hoja de respuestas, modo de marcar la respuesta, posición relativa de las sucesivas preguntas en el test y en la hoja de respuestas, etc.).

2. Permiten verificar que se han entendido las instrucciones básicas y los ejemplos previos. En tests colectivos permiten al psicólogo que administra los tests disponer de unos segundos para pasar junto a cada sujeto y comprobar, antes de que sea tarde, que se está contestando en el lugar indicado y del modo indicado, es decir, que se han entendido las instrucciones generales.

3. Animar a la persona que contesta el test a seguir adelante. Probablemente sería frustrante no poder acertar ni un ítem. Además la persona expuesta a esta situación no sabría si no acierta porque no alcanza a obtener la respuesta o porque no ha comprendido bien como debe hacer la tarea.

13. El análisis de los distractores

En un ítem con respuesta verdadera que ofrece g alternativas se denomina distractor a cada una de las g-1 alternativas que no son la respuesta correcta.

Los distractores también han de ser analizados y han de cumplir una serie de características formales y empíricas para que sean admisibles. La calidad del ítem depende fuertemente, también, de la calidad de los distractores.

La característica más obvia es que han de ser falsos. Clara y expresamente falsos si se trata de escoger la respuesta verdadera. (O verdaderos: clara y expresamente verdaderos si el enfoque del ítem es escoger la respuesta falsa de entre un conjunto de verdaderas.) Y, por supuesto, no han de llevar consigo indicios formales o semánticos de que son falsos, ni estar contenido en otro ítem o en otra parte del mismo ítem la información que permita

descartarlos. Preferentemente serán independientes entre sí (si no lo son la asignación de una penalización razonablemente por posibles aciertos al azar puede ser compleja).

A parte de las características anteriores, que pueden considerarse obvias, los distractores han de cumplir las siguientes cualidades generales:

1. Desde el punto de vista formal y de contenido, han de presentar el mismo aspecto, longitud, formato y características que la respuesta correcta. Su contenido ha de ser verosímil, razonable y cualitativamente semejante al de la respuesta correcta.

Algunas características formales como longitud en número de lineas y palabras y semejanza de contenido son fácilmente contrastables.

2. Han de ser atractivos para las personas con menor nivel de capacidad y no-atractivos, como respuesta, para los sujetos capaces.

Esto puede contrastarse empíricamente después de administrar la prueba, estudiando para cada alternativa de respuesta qué proporción de sujetos de cada nivel de capacidad la han acertado. Cuanto mayor sea el nivel de capacidad de los sujetos (abscisas) menos proporción de sujetos (ordenadas) tiene que haber escogido un distractor dado. Para cada distractor la probabilidad de que sea escogido como respuesta por un sujeto ha de disminuir conforme aumenta la capacidad de los sujetos.

A veces un distractor aparece como muy atractivo (elegido) para los sujetos muy capaces. Ello probablemente significa que estos sujetos están analizando la cuestión a un nivel distinto, más profundo o más sofisticado que el resto del grupo, debido a que saben más o a que son más capaces, lo que les lleva a dar una respuesta original distinta a la que se asume como convencional. Si su explicación es razonable debe modificarse la prueba o corregirse el distractor como correcto para ese grupo.

3. Los distractores han de ser equiprobables entre sí. Esto quiere decir que cada uno de los distractores debería tener la misma probabilidad de ser escogido por un sujeto que no sepa la respuesta correcta.

Esto también puede contrastarse empíricamente. Si los distractores son equiprobables entonces cada uno de ellos habrá recibido aproximadamente el mismo número de elecciones.

Romper la equiprobabilidad de los distractores es uno de los defectos más típicos de las pruebas objetivas. En estos casos encontramos algún o algunos distractores que son fácilmente desechables por las personas que contestan la prueba, o, simplemente, que son más fácilmente desechables que los otros. Esto tiene como consecuencia la reducción de errores al contestar al azar. La conclusión de este fenómeno es que la capacidad absoluta de los

sujetos queda sobrevalorada debido a que aumentan los aciertos por azar.

Yela (1984; Cap. 18, pag. 94) ha presentado un procedimiento práctico para analizar el funcionamiento de los distractores en los items que resulta sencillo y útil al tiempo. Consiste en dividir a los sujetos en dos grupos partiendo por la mediana en el total del test. Esto permite disponer de un grupo de sujetos más capacitados y menos capacitados en la variable medida. Después se calcula, para cada ítem y separadamente para cada grupo la proporción de elecciones que recibe cada alternativa. En el grupo superior la alternativa correcta debe haber sido más elegida que en el inferior. Las alternativas incorrectas deben haber sido más elegidas en el grupo inferior y además, en cada grupo, cada alternativa incorrecta debe haber sido elegida por aproximadamente el mismo número de sujetos. Si se desea un estudio a grandes rasgos del funcionamiento de los items este enfoque puede ser suficiente. Para un estudio detallado podemos clasificar la muestra en 10 grupos de aptitud (sujetos hasta el percentil 10 en la puntuación total, del 10 hasta el 20, etc.) y estudiar la evolución de la proporción de respuestas que recibe cada alternativa a lo largo de esta escala de aptitud, como antes sugeríamos.

Un ítem de verdadero-falso, o, lo que es prácticamente equivalente, un ítem con dos alternativas de respuesta presenta un solo distractor (la alternativa falsa).

En el caso de los items de verdadero-falso, que presentan un solo enunciado ante el que el sujeto debe pronunciarse, podría decirse que los items falsos actúan como distractores y desde luego comparten con los genuinos distractores algunas características. Así, los items falsos han de ser atractivos para los menos capaces y nada o poco atractivos para los más capaces. También han de ser formalmente y en cuanto al contenido razonablemente semejantes a los verdaderos. Sin embargo, no se puede decir que han de ser equiprobables entre si (entre los items falsos) dado que, legítimamente, diversos items pueden tener diverso nivel de dificultad.

Desde un punto de vista más pedagógico que psicométrico, por lo que se refiere a items de pruebas objetivas para evaluar conocimientos o destrezas, debe decirse que los items han de ser relevantes en su contenido, muestrear adecuadamente el campo a medir y movilizar las capacidades, aptitudes, conocimientos y habilidades superiores que se consideren más importantes -en la medida de lo posible-. Las pruebas objetivas no son adecuadas para medir cualquier aptitud o habilidad en cualquier contexto y otros métodos pueden presentar otras ventajas estimables.

Construir buenos items con respuesta verdadera con g alternativas de las cuales g-1 son buenos distractores no es fácil en muchos contextos. Exige un esfuerzo adicional por cada distractor más que se añada. Cuantos más buenos distractores se añadan menor es la probabilidad de que la persona que contesta la prueba pueda acertar al azar. Cuanto menos dependa la respuesta del azar y más del conocimiento, capacidad o habilidad medida tanto mejor. Sin embargo, es preferible plantear un ítem con menos distractores si la calidad de estos comienza a disminuir.

Un distractor de mala calidad introduce factores de distorsión en la respuesta. El más conocido es el incremento de la probabilidad de acierto al azar -un efecto contraproducente-, con el agravante de que ese incremento de la probabilidad no es conocido a priori, lo que hace que sepamos menos en qué grado las respuestas del sujeto pueden ser fruto del azar. Además los malos distractores pueden añadir otros problemas como información que contamina otros items, cansancio adicional, confusión, irritación para el sujeto que contesta la prueba, etc. cuyos efectos conjuntos no se pueden cuantificar a priori (antes de administrar la prueba).

Este problema de perdida de calidad de los distractores es especialmente notorio en las pruebas objetivas para evaluar rendimientos académicos donde con frecuencia hay que elaborar h pruebas nominalmente paralelas. Desde luego una ventaja patente de los items de verdadero-falso es que este problema se disuelve dado que no hay que construir distractores propiamente, aunque, como ya he mencionado, los items falsos deben cumplir algunas de las propiedades principales de un buen distractor.

Paradójicamente cuando se trata de pruebas objetivas para medir rendimiento académico algunos estudiantes prefieren items de alternativas, y cuantas más alternativas mejor. Digo paradójicamente porque se supone que cuantas más alternativas más difícil ha de resultar la prueba porque más difícil habría de resultar acertar por azar. Sin embargo, como bien sabe cualquier persona habituada a este tipo de pruebas esto no es así. En realidad, se produce el fenómeno contrario *aunque los distractores estén bien construidos. La razón es que aun los distractores bien construidos pueden ser descartados debido a conocimiento parcial del sujeto, lo que en conjunto aumenta la probabilidad de acierto por respuesta al azar y reduce el grado en que la respuesta final depende exclusivamente del conocimiento del sujeto sobre la alternativa correcta.* No he visto expresado ni analizado este efecto para distractores bien construidos en ningún estudio, pero un sencillo razonamiento creo que podrá mostrar las bases de mi punto de vista.

Un distractor puede o bien contener información adicional a la respuesta correcta o no contenerla. (No la contiene cuando la negación del distractor implica inmediatamente la aceptación de la respuesta correcta, como en una prueba

verdadero-falso). Si no la contiene, o estamos ante una prueba verdadero-falso, -donde esto es legítimo,- o estamos ante un ítem de 3 o más alternativas mal construído puesto que los otros distractores pierden su condición de equiprobables. Ahora bien, y esta es la paradoja, si contienen información adicional, específica, esta información puede ser conocida en su falsedad y descartada por la persona que contesta, *independientemente* de su conocimiento de la información de la respuesta correcta. Si no existe este conocimiento parcial de los distractores entonces éstos permanecen equiprobables para un sujeto que no conoce la respuesta correcta. Pero si existe este conocimiento parcial -como suele suceder para un sujeto determinado en al menos algunos items de cualquier prueba con 3 o más alternativas- entonces los distractores aun estando bien construídos para el sujeto que posee el conocimiento parcial, dejan de ser equiprobables, lo que aumenta la probabilidad de acierto en una cuantía desconocida y para un número de items desconocido. Conclusión: las pruebas de tres o más alternativas introducen un elemento distorsionador en la probabilidad de acierto no atribuible a un defecto técnico en la construcción de los items. (Por supuesto, a este efecto habría que añadir, en la práctica, los defectos técnicos que frecuentemente sí tienen estas pruebas).

Un problema adicional de añadir más distractores -que si ha sido muy analizado en la teoría clásica- es que incrementan el tiempo que la persona que contesta requiere para leer la prueba, y, si son buenos distractores, puede incrementar también el tiempo necesario para elegir la respuesta. Como en la mayoría de circunstancias prácticas el tiempo real de que se dispone para contestar la prueba tiene unos límites concretos, se plantea la cuestión de si añadir más distractores por ítem es o no un buen modo de aprovecharlo.

Esta cuestión del análisis de los distractores y de sus cualidades lleva a plantear cuál es el número óptimo de alternativas de respuesta (distractores más respuesta correcta) en un test o prueba objetiva con respuesta verdadera.

Lord (1980, Cap. 7, pags. 106-127) ha dedicado un capítulo de su influyente obra a revisar la cuestión de cual es el número de alternativas más adecuado. Diversas investigaciones citadas por Lord y las nuevas propias indagaciones que introduce este autor, muestran que, en general, el número de alternativas debería estar entre 2 y un máximo de 4. Parece que 3 puede ser un número óptimo para muchos propósitos, con *muy ligera* ventaja sobre los items de dos alternativas y algo más de ventaja sobre los de cuatro alternativas. Se desprende que, aunque sólo sea por economía de tiempo de los sujetos que contestan la prueba y del constructor de la prueba, los esquemas más adecuados son 2 ó 3 alternativas.

En términos prácticos, sin embargo, las diferencias entre dos o tres alternativas son tan mínimas que resultan despreciables. Cualquier ligero incremento en la calidad de la prueba será más relevante para su resultado final que pasar de dos a tres alternativas. En ese contexto parece más adecuado centrar el esfuerzo en elaborar buenos items que muestreen bien el dominio a medir que ocuparse de elaborar buenos distractores que minimicen la posibilidad de que el sujeto acierte al azar. Si a estos estudios empíricos y matemáticos se añaden las consideraciones lógicas que hacia anteriormente parece desprenderse que, con las limitaciones que tiene cualquier prueba objetiva, un modelo de prueba de verdadero falso puede ser aceptablemente recomendable como tipo de prueba.

(En cualquier caso nada más lejos de mi intención que mostrar entusiasmo por nuestros procedimientos habituales de medición psicológica o académica).

14. El índice de discriminación basado en p

Si se clasifican los sujetos en submuestras por su nivel de capacidad el índice de dificultad (valor p) debe ir aumentando -para items bien construidos- según aumenta la capacidad de los sujetos. Es lo que hemos denominado "perfil de dificultad del ítem en función de la capacidad de los sujetos", o, colocando la escala k en abscisas, un "perfil p del ítem a través de la capacidad".

El principio de estas gráficas es que p debe crecer a medida que aumenta la capacidad. Es decir, si el subíndice representa cada sucesivo nivel k de capacidad hasta un nivel máximo w:

$$p_1 < p_2 < p_3 < \ldots < p_w$$

La teoría clásica elaboró un estadístico muy sencillo e intuitivo basado en este principio que se nombra como "índice de discriminación" en algunos textos. Aquí lo denominaremos *índice de discriminación basado en p* para evitar confusiones pues la palabra discriminación, como después comentaré, se ha usado con diversos -aunque congruentes- significados en diversos manuales de teoría clásica.

Si dividimos la muestra por la mediana en función de su puntuación en el total del test (es decir, los que son más capaces y los que son menos capaces según el test), y calculamos separadamente p en cada uno de estos grupos, entonces el estadístico:

$$D = p_s - p_i$$

expresa la diferencia entre la proporción de sujetos que aciertan el ítem en el grupo superior y la proporción de sujetos que aciertan el ítem en el grupo inferior. Si el test y el ítem son razonables entonces la proporción de sujetos

que aciertan el ítem en el grupo superior (más competentes) debe ser mayor que la proporción de sujetos que lo aciertan en el grupo inferior.

Ebel (1965; citado en Crocker y Algina 1986, pag 315) dio unas orientaciones prácticas para interpretar el estadístico D que resultarán muy útiles,(a falta de poder efectuar con D contrastes de hipótesis porque, al parecer, no se conoce su distribución muestral). Según Ebel, si:

-D es mayor que 0,39 el ítem funciona muy bien;

-D esta entre 0,3 y 0,39 ó no hace falta tocar el ítem o bastará con algún ligero retoque del mismo,

-D está entre 0,2 y 0,29 el ítem quizás podría aprovecharse si se revisa para conseguir aumentar D,

-D es menor que 0,2 el ítem no sirve en absoluto y para que sirva ha de ser totalmente revisado (lo que en la práctica supondrá probablemente desecharlo y construir uno nuevo).

Si el grupo superior esta formado por el 27% de sujetos con puntuaciones más altas en el test y el grupo inferior por el 27% inferior la diferencia D puede ser más notoria y estable (es menos pronosticable si un sujeto de capacidad media acertará o no un ítem dado que un sujeto con capacidad marcadamente inferior o superior). Este método de trabajo, con solo el 27% superior e inferior fue sugerido por Kelley (1939) y aunque una de sus principales virtudes (evitar trabajo en cálculos a mano) ya hace mucho que dejo de tener razón de ser, para este caso concreto todavía tiene sentido su uso por la razón expuesta.

Usos de la palabra discriminación. Existe una pluralidad de usos de la palabra discriminación en diversos manuales. Por ejemplo, al "índice de discriminación basado en p" Crocker y Algina (1986; pag. 314) lo denominan "índice de discriminación", sin apellidos, y lo distinguen de los "índices correlacionales de la discriminación del ítem", que abarcan la correlación ítem-test y la correlación ítem-criterio, con un concepto amplio de discriminación que abarca la homogeneidad y la validez de los ítems. Santisteban (1990, cap. 15, pág. 337 y siguientes) mantiene una nomenclatura consistente con la de Crocker y Algina (1986). Santisteban (1990; pag. 335) también utiliza la denominación "índice de discriminación" para referirse a la correlación entre "las respuestas dadas a ese ítem con respecto a las dadas a los otros items del test". Yela (1984, Cap. 18), aunque pone las bases acerca de como estudiar el índice de discriminación basado en p, no lo menciona y denomina a la correlación entre el ítem y el test "homogeneidad" del ítem, enfatizando su relación con la fiabilidad, que es la denominación que he seguido aquí. He preferido hacer este comentario sobre la pluralidad de denominaciones para evitar que el lector pueda confundirse al estudiar diferentes manuales y pensar que esta variedad de denominaciones se debe a algún error o desinformación. Simplemente no hay en este punto una nomenclatura unificada: las mismas etiquetas se han usado

para distintas cosas y una misma cosa ha recibido diferentes etiquetas. Una vez establecida esta innecesaria discordia léxica no conozco ningún argumento fuerte para decidirse por un sistema de denominación particular.

El índice de discriminación basado en p puede verse todavía como parte del estudio de la distribución de los items con respuesta verdadera a través de su dificultad, pero, a la vez, enlaza con el estudio de la relación entre el ítem y el total del test que abordaremos a continuación.

15. La fiabilidad de los items referida a los items

Un ítem es la unidad mínima de medición con sentido completo. Es, metafóricamente, la molécula del test o la frase completa en el discurso. La unidad mínima de reactivo que conserva las propiedades del test o la unidad mínima del test que conserva un significado pleno. Por eso el análisis de los items es esencial como constituyentes del test. En la medida en que cada ítem capta una porción de información valiosa y distinta a la de otro ítem, resulta indicado un análisis de *cada* ítem por sí.

El modo clásico de abordar esta cuestión ha heredado de la tradición de Gulliksen (1950; Cap. 21) y Lord y Novick (1968) un énfasis en el análisis de items subordinado al análisis del test. Es decir, se analizan los items en la medida en que repercuten sobre los parámetros del test (su total, su media, su varianza, su fiabilidad y su validez). Es difícil pensar que cualquier propiedad de los items no influya en el test porque el test no es más que una colección de items estructurados para algún fin. Aunque eso si, pueden haber índices de análisis de los items que no contribuyan o no este claro como contribuyan a los parámetros del test.

Una exposición de la relación de los índices más importantes referidos a los items y los parámetros del test puede verse en Gulliksen (1950, Cap. 21).

Aunque este no es el modo usual de enfocar la cuestión en los manuales de teoría clásica, lo cierto es que podríamos estudiar la cuestión de la fiabilidad de los items análogamente a como hemos estudiado la cuestión de la fiabilidad del test como un todo.

En mi opinión, desarrollando esta linea de razonamiento pueden establecerse las siguientes definiciones:

Estabilidad del ítem a través del tiempo: Se define como la correlación entre las puntuaciones en el ítem en tiempo 1 para una muestra y las puntuaciones en el ítem en tiempo 2 para esa misma muestra siempre que no hayan cambiado sujetos o condiciones. Este concepto es análogo al de fiabilidad test-retest para la prueba total.

Fiabilidad entre items paralelos. Se define como la correlación para una misma muestra entre las puntuaciones del ítem y otro ítem paralelo. Muchos tests clásicos parecen estar construidos para obtener items paralelos. Si la paralelidad pudiera satisfacerse -cuestión compleja a la que ya dedicamos mucha atención anteriormente- la correlación entre items paralelos sería desde luego la fiabilidad del ítem. (Un concepto por cierto que, como hemos visto, presuponen algunos métodos y procedimientos de análisis de la fiabilidad del test total, pero que, sorprendentemente, es usualmente ignorado en los capítulos de análisis de items).

16. La homogeneidad de los items

El enfoque tradicional sin embargo no está orientado al ítem sino al test, y, consecuentemente, enfoca esta cuestión mediante el análisis de la relación entre ítem y total del test.

Se define como *índice de homogeneidad del ítem* a la correlación del ítem con el total del test.

Se trata de la *correlación de Pearson* entre las puntuaciones que los sujetos de una muestra obtienen en un ítem y las puntuaciones de esos mismo sujetos en el total del test, generalmente definido este total como la suma de los items. (Algunos manuales denominan a esta correlación índice de discriminación).

Homogeneidad corregida. Como el total es la suma de los items, entonces contiene también al ítem cuya correlación con el total estamos calculando. Esto, por supuesto, produce una correlación sobrestimada. Esa sobrestimación puede ser irrelevante cuando el número de items es alto (por ejemplo más 20) pero será importante en pequeños cuestionarios con unos pocos items. En ese caso puede optarse por calcular la correlación del ítem con el total excluido el ítem. Esto significa que para estudiar la homogeneidad de cada ítem hay que calcular cada vez un total distinto (todos los demás items menos ese ítem en particular) -Esto resultará tedioso hasta con un paquete estadístico.-

Si no se dispone de un programa o paquete estadístico que calcule directamente la homogeneidad corregida lo más práctico es aplicar la siguiente fórmula de la homogeneidad corregida:

$$r_{x\,(y-x)} = \frac{r_{xy}\,s_y - s_x}{\sqrt{s_y^2 + s_x^2 - 2\,r_{xy}\,s_y\,s_x}}$$

La fórmula anterior expresa la correlación entre el ítem X y el total Y menos el ítem X, es decir, la correlación de X con (Y-X) sin necesidad de elaborar un total del test distinto para cada ítem.

En Yela (1984; pag. 79), por ejemplo, puede seguirse la obtención de esta fórmula a partir de la fórmula general del coeficiente de correlación de Pearson entre X y (Y-X).

En realidad la correlación ítem-test, dado que el ítem es parte del test es una cuestión teóricamente engorrosa. Si se utiliza siempre el total con todos los items, efectivamente puede decirse que evaluamos la relación de cada ítem con el total real del test (con todos sus items) pero sobrestimamos la homogeneidad. Si "corregimos la homogeneidad" entonces en realidad no calculamos la relación de cada ítem con el total del test, sino con una aproximación a ese total fruto de excluir el ítem. Vistas así las cosas la corrección de la homogeneidad que se trasmite de manual en manual es solo una pseudosolución. Que yo sepa esta paradoja ni tiene solución ni tiene demasiada importancia.

Se denomina *índice de fiabilidad de un ítem* a la homogeneidad del ítem multiplicada por la desviación típica del ítem. Este índice considera simultáneamente el grado de variabilidad del ítem (y recordemos que cuanto mayor es su variabilidad más discrimina entre sujetos) y su relación con el total de la escala.

La desviación típica del total de un test (elaborado como suma de los items) es igual a la suma de los índices de fiabilidad de sus items (Gulliksen, 1950). Si los índices de fiabilidad son mayores mayor será la varianza del total del test.

Interpretación del índice de homogeneidad. La homogeneidad de un ítem expresa en que grado el test puede considerarse una función lineal del ítem. Cuanto mayor es esta correlación más consistente es el ítem con el resultado del test.

Si la homogeneidad es alta la puntuación del ítem permite predecir bien la puntuación en el test. Si la homogeneidad del ítem es alta, el ítem, por tanto, permite discriminar (diferenciar) entre sujetos con puntuaciones altas y bajas en el test. Si el ítem no está relacionado con el total del test (homogeneidad baja) entonces no permitirá pronosticar (diferenciar) bien entre sujetos con puntuaciones altas y bajas en el total del test.

En términos de discriminación, cuanto mayor es la homogeneidad de un ítem mayor capacidad tiene el ítem para discriminar, para distinguir, entre los que puntúan alto y los que puntúan bajo en el test. Desde este enfoque, tradicionalmente, se recomienda que los items presenten al menos cierta homogeneidad, de modo que los items pueden ser seleccionados para formar parte del test, junto a otros criterios, en función de su homogeneidad. O, mejor todavía, teniendo en cuenta simultáneamente su homogeneidad y su dispersión, en función de su índice de

fiabilidad. Por lo que respecta a este criterio tienden a preferirse aquellos ítems cuyo índice de fiabilidad es mayor.

Si un ítem no presenta ninguna relación con el total del test (en cuyo caso su homogeneidad estará próxima a 0) puede cuestionarse que este ítem forme parte del test. Puede cuestionarse que tenga sentido sumar este ítem a los demás para formar un todo que es el test.

En el otro extremo, si un ítem correlacionara 1 con el total del test, entonces la información que da el test es completamente redundante con la que da el ítem, bastaría el ítem para conocer con certeza el total del test.

Desde el punto de vista del contenido cuanto más homogéneo, semejante, parecido sea el contenido de los items más alta será su homogeneidad. Si producimos items muy semejantes probablemente estamos aumentando la homogeneidad. Por otra parte, si en el análisis de un proceso psicológico mezclamos items sobre causas y efectos, si estas causas y efectos aparecen indisolublemente asociados en la realidad, también estaremos aumentando la homogeneidad aunque desde el punto de vista del análisis cualitativo psicológico del contenido se trate de cosas bien distintas.

Por supuesto puede ser pernicioso para una medida aumentar la homogeneidad a costa de reducir los contenidos psicológicos que se muestrean. Si todos los items son variantes de uno sólo la homogeneidad probablemente será altísima, y, en la misma medida, el test absurdo, poco válido o irrelevante. Muestrear *bien* un campo de contenidos psicológicos (sean conocimientos, capacidades, habilidades, aptitudes, personalidad o actitudes) implica un cierto grado de heterogeneidad. Implica tanta heterogeneidad como partes, porciones, dimensiones, factores o procesos distintos de interés podamos distinguir en la variable a medir.

A veces se razona apuntando que la homogeneidad no es un fin en sí misma, sino en la medida en que contribuye a la validez del test. Mi opinión personal es que la homogeneidad del ítem es secundaria. Si hemos de estar interesados esencialmente en la validez del test entonces lo prioritario es la validez del test misma -y por tanto la validez de los items, especialmente en la medida en que aportan algo más, algo nuevo, a la capacidad predictiva del test-. Podemos analizar esa validez directamente. Esto no debe interpretarse en el sentido de que cuando se hace análisis de items no deba calcularse la homogeneidad para comprender como opera cada ítem. Debe interpretarse en el sentido de que, por lo general, la validez es preferible a la homogeneidad y, a veces, a pesar de la pérdida de homogeneidad.

17. La validez de los items

Se denomina *índice de validez de un ítem* a su coeficiente de correlación de Pearson con un criterio externo al test.

Expresa el grado de relación lineal entre el ítem y el criterio. En general es propósito del constructor del test disponer de items validos, que correlacionen con el criterio o criterios en el modo esperable por hipótesis.

Gulliksen (1950) para ser consistentes con la definición anterior de índice de fiabilidad de un ítem, denomina índice de validez de un ítem al producto de esta correlación entre ítem y criterio con la desviación típica del ítem. Santisteban,(1990, Cap. 15, pag.343) parece seguir la denominación de Gulliksen. Aquí he preferido seguir la que utiliza Yela (1984, Cap. 18, pag. 80).

18. Síntesis de las relaciones entre los parámetros de los items y los parámetros del test

Gulliksen (1950, Cap. 21, Pag. 380 y siguientes) ha presentado sistemáticamente la relación entre los parámetros del ítem y los del test. No entraré en la exposición de las deducciones algebraicas que llevan a esas relaciones, basadas en la definición del test como un compuesto aditivo de los items.

Media del test. La media $\bar{Y}$ del test es igual a la suma de las medias $\bar{X}_j$ de los items, y, en el caso de ítem dicotómicamente valorados, a la suma de los valores p de los items.

$$\bar{Y} = \sum \bar{X}_j = \sum p_j$$

Desviación típica del test. Es igual a la suma de los índices de fiabilidad. Dicho de otro modo, la desviación típica del total del test s_y es igual a la suma de los productos de los índices de homogeneidad r_{xy} por las desviaciones típicas de los items s_x.

$$s_y = \sum r_{xy} s_x$$

Fiabilidad (consistencia interna) del test. Como aplicación directa de la igualdad anterior puede reescribirse sustituyendo la varianza en el total del test por la suma de los índices de fiabilidad al cuadrado.

$$r_{XX} = \frac{n}{n-1}\left(1 - \frac{\sum s_x^2}{(\sum r_{xy} s_x)^2}\right)$$

Donde n es el número de items, y s_x^2 la varianza de los items.

Coeficiente de validez del test. Se denomina así a la correlación entre el total del test y un criterio. Gulliksen (1950, pag.382) mostró que la correlación entre el total del test Y y un criterio C es igual a:

$$r_{YC} = \frac{\sum r_{xc} s_x}{\sum r_{xy} s_x}$$

Donde:

r_{xc} es la correlación de cada ítem con el criterio

r_{xy} es la correlación de cada ítem con el test

s_x es la desviación típica de cada ítem.

La fórmula anterior supone que cuanto mayores sean las valideces de los items mayor será la validez del test. Sin embargo, cuanto mayores sean las homogeneidades de los items la validez será menor.

Como la validez del test es, a su vez, una correlación y, por tanto, no puede ser mayor que 1, se desprende que el numerador ha de ser menor que el denominador. Es decir, que las valideces serán menores que las homogeneidades.

Se obtendría una validez del test perfecta si las valideces llegan a ser iguales que las homogeneidades.

Esta síntesis sigue la ofrecida por Gulliksen (1950; pag. 389), con solo alguna variación en la nomenclatura.

19. Corrección de la respuesta al azar

Uno de los inconvenientes de las pruebas objetivas y tests, con items con respuesta correcta, donde la tarea del sujeto consiste simplemente en identificar qué es lo verdadero, consiste en que los sujetos pueden acertar algún o algunos items al azar.

¿Cuántos items pueden acertarse por azar? Suponiendo que los sujetos (condición a) no sepan absolutamente nada y (condición b) contesten rigurosamente al azar entre las k alternativas, entonces puede esperarse que los sujetos acierten un ítem por cada k items.

Por ejemplo, si la prueba tiene 4 alternativas (una de las cuales es correcta y las otras tres distractores) un sujeto acertara al azar uno de cada 4 items. Ello significa que si todos los items fueran contestados al azar, en promedio, el

sujeto habría de acertar un 25% aproximadamente. Es decir, para un sujeto que no sabe nada la puntuación media que obtendrá contestando por azar una prueba objetiva de 4 alternativas es 2'5 (escalada la puntuación de 0 a 10).

Por ejemplo, con una prueba de 3 alternativas contestada al azar se espera un acierto del 33'33% y una puntuación de 3'3 en escala de 0 a 10.

Por ejemplo, con una prueba de verdadero-falso, la puntuación media de una muestra de sujetos que no supieran nada y la contestaran rigurosamente al azar sería 5, en escala de 0 a 10. Es decir, por azar se espera un acierto del 50% porque si el ítem tiene dos posibilidades (verdadero o falso, k=2) existe un 50% de probabilidades de acertar contestando al azar.

En síntesis puede decirse que cumplidas las condiciones a y b antes enunciadas el número de aciertos esperados por azar es 1/k.

Por supuesto esto solo es cierto en promedio. Para una persona concreta a la que sonría la fortuna, el acierto por azar podría ser del 100% -debería dedicar su tiempo a hacer loterías y no desperdiciarlo en pruebas objetivas, antes de que cambie la racha.- Para otro, desafortunado, el acierto por azar podría ser del 0%.

Evidentemente las condiciones a y b no se cumplen de modo general ni completamente para las muestras de sujetos a los que administramos pruebas objetivas o tests susceptibles de ser acertados por azar. Lo usual es que nos encontremos con puntuaciones en las que hay un cierto número de aciertos y un cierto número de errores. Y también *omisiones*, es decir, preguntas no contestadas. ¿Cómo saber cuantos de los aciertos se deben al azar? En rigor nunca lo sabemos con certeza. Posiblemente algunos aciertos sean fruto del azar y otros del conocimiento, habilidad o aptitud de la persona que contesta la prueba. Los aciertos totales pueden clasificarse en dos grupos: los debidos al azar y los debidos a la aptitud.

$$A_{totales} = A_{azar} + A_{aptitud}$$

Parece necesario utilizar alguna estimación de cuántos aciertos pueden deberse al azar. Si se cumplen las condiciones a y b, entonces, puede utilizarse el número de errores que el sujeto ha cometido, para obtener una estimación de cuanto acierto podría deberse al azar.

En una prueba de verdadero falso (k=2), cumplidas las condiciones a y b, por cada error habrá un acierto por azar. Es decir, si tuviéramos que estimar los aciertos debidos a azar a partir de los errores diríamos que por cada error se habrá obtenido, en promedio, un acierto por azar.

En una prueba de 3 alternativas (k=3), por cada 2 errores habrá un acierto por azar.

En una prueba de 4 alternativas (k=4), por cada 3 errores habrá un acierto atribuible al azar

En general por cada k-1 errores habrá un acierto por azar. Por tanto, si en una prueba vemos E errores ¿cuantos aciertos por azar se habrán producido? Depende del número de alternativas, cada k-1 errores supondrá un acierto. Por tanto un estimador del número de aciertos debidos al azar es el número de errores que de hecho observamos dividido por k-1, porque por cada k-1 errores toca un acierto por azar.

$$A_{azar} = \frac{E}{k-1}$$

Si deseamos que el total de la prueba refleje la aptitud del sujeto, y no el azar, entonces deberíamos descontar del número total de aciertos observados aquellos debidos al azar:

$$A_{aptitud} = A_{totales} - A_{azar}$$

Como los aciertos por azar podemos estimarlos a partir de los errores resulta:

$$\boxed{A_{aptitud} = A_{totales} - \frac{E}{k-1}}$$

La fórmula anterior estima los aciertos por azar a partir del número de errores y los descuenta de los aciertos totales, ofreciendo una mejor estimación de los "puntos" debidos a la aptitud. Esta es la conocida fórmula de corrección del azar que se suele aplicar en pruebas objetivas.

La fórmula de corrección del acierto por azar es correcta *si se cumplen* una serie de condiciones prácticas:

1. Que se trate de una prueba con alternativas de respuesta conocidas y prefijadas entre las que el sujeto ha de escoger la respuesta correcta (Por supuesto k>1 y la corrección solo tiene sentido para items con respuesta verdadera).

La fórmula de corrección del acierto por azar no puede aplicarse donde no puede haber acierto por azar. La fórmula de corrección del acierto por azar no puede aplicarse, por ejemplo, a pruebas de ensayo libre, a pruebas basadas en preguntas cortas abiertas, a pruebas basadas en recordar y escribir, etc. Por ejemplo, no puede aplicarse a una prueba objetiva donde cada ítem es un hueco en un párrafo de un texto donde hay que escribir la palabra correcta.

2. Que la probabilidad de acierto por azar sea conocida y sustancial.

La fórmula de corrección del acierto por azar no puede aplicarse donde la probabilidad del acierto por azar es remotísima o simplemente desconocida. Si le pedimos a alguien que adivine y escriba los nombres científicos de 10 flores distintas, de los que se muestra la ilustración, la probabilidad de acierto por azar es remotísima y desconocida. Si se equivoca al intentar contestar no procede descontar aciertos debidos al azar. Si le pedimos a alguien que no conozca nuestro número de documento de identidad que intente adivinarlo, la probabilidad de acertar

es conocida (fijado el número de cifras) y remotísima. No procede descontar aciertos por azar en función de sus errores.

Algunos items mezclan contenidos de alternativas entre sí, condicionan de diversos modos unas alternativas a otras, etc. Estos procedimientos introducen generalmente incertidumbre sobre la probabilidad de acierto al azar, y, son desaconsejables si vamos a utilizar corrección del acierto por azar.

3. Que las personas que contesten la prueba estén motivadas para intentar acertar al azar y que, de hecho, lo intenten.

Si estamos en una prueba objetiva tipo examen, parece que las personas pueden estar motivadas a intentar obtener su mejor puntuación. Lo mismo sucede en un contexto de selección y clasificación de personal. Sin embargo, si estamos en una situación de gabinete, examinando la inteligencia de un sujeto, posiblemente el sujeto pueda no estar motivado para intentar "ganar" puntos mediante el azar. En este último caso no procede aplicar la corrección.

La fórmula es razonable para corregir los posibles aciertos al azar; su propósito no es sobrecorregir la incapacidad, la ignorancia o el desconocimiento. Por ejemplo no tiene sentido "penalizar" contando las omisiones como si fueran errores. Si un sujeto omite contestar una pregunta es obvio que no ha jugado al azar en esa pregunta ¿qué razón hay para utilizar las omisiones como un estimador de los aciertos que el sujeto pueda haber obtenido por azar?

Tampoco me parece legítimo técnicamente obligar al sujeto a contestar al azar y después penalizar el azar. Es como obligar a todo el mundo a jugar a la lotería. ¿Por qué si el propósito de la prueba ha de ser medir capacidades, aptitudes, habilidades...? Creo que este proceder es contradictorio. Puede verse como un modo probabilístico de penalizar la omisión. (A favor de estos usos a veces se argumenta que, a juicio del constructor de la prueba, por razones ajenas al proceso de medición, el examinado *debería saber* aquellos contenidos para superarla. En ese caso lo más sencillo y correcto es situar el nivel de "aprobado" en un porcentaje de aciertos que garantice que no pueda aprobarse con omisiones.)

Las condiciones anteriores son *condición necesaria* para que tenga sentido aplicar la corrección de la respuesta al azar. Pero, aun en el caso ideal la fórmula no garantiza un tratamiento "justo" de una persona en particular.

Para comenzar, desde un punto de vista ético, resulta imprescindible que los sujetos conozcan antes de la prueba las condiciones bajo las que son examinados. Es decir han de conocer si se va a corregir la respuesta al azar y con qué fórmula. (En caso contrario las especulaciones sobre como se corregirá la prueba introducirían un elemento divergente que podría ser esencial para explicar el comportamiento de los sujetos en la prueba y por tanto sus puntuaciones. Un

elemento espurio ajeno a la capacidad que queramos medir.)

Pero, cuando los sujetos conocen la fórmula de corrección esta modifica su conducta y es razonable pensar que se vuelven más cautos y que, en general, menos de sus aciertos y menos de sus errores se deben al azar.

No todos los *errores* se deben al azar. Puede afirmarse que el número total de errores puede descomponerse en errores debidos al azar y equivocaciones.

$$E_{totales} = E_{azar} + E_{equivocaciones}$$

Una *equivocación* es un error legítimo, donde no se ha jugado al azar.

Algunos de mis colegas discreparán de este planteamiento y preferirán juzgar las equivocaciones como errores que también implican un componente de azar, pero yo no comparto este punto de vista. Creo que hay casos claros en que un error es al azar, casos con cierto componente de azar (después hablaremos de esta gradualidad) y casos donde el error es equivocación claramente. Por ejemplo, se presenta un ítem en una prueba objetiva verdadero-falso de estadística donde a partir de unos datos la persona examinada ha de calcular el coeficiente de correlación de Pearson y decir si la respuesta que ofrece el ítem es la correcta (que, efectivamente, supongamos, lo es). Una persona en papel adjunto al examen muestra haber conocido, recordado y aplicado paso a paso la fórmula adecuada. Pero en un paso determinado comete un error mecánico de cálculo que le lleva a invertir la posición de dos cifras en un número. Conclusión el sujeto, que domina la materia, contesta, convencido, que el ítem es falso cuando no lo es. En mi opinión no puede sostenerse que esto sea un error debido a azar y por tanto no debería descontarse por un error de este estilo, aunque, de acuerdo al patrón de valoración de la prueba, el ítem debiera valorarse con cero puntos. Puede argumentarse que fue el azar el que llevo a invertir los números. Efectivamente, pero este azar no puede sostenerse que vaya a ayudar a acertar. Si el lector examina el caso con cuidado observará que la cuestión es todavía más compleja, pues si el ítem hubiese sido formulado como falso el sujeto hubiera acertado aun equivocándose (despreciamos la probabilidad de que por error encontrara el mismo resultado que el resultado erróneo del ítem).

En mi opinión, si fuera posible, habría que aplicar la corrección de la respuesta al azar estimando,los aciertos por azar solo a partir de los errores por azar.

$$A_{aptitud} = A_{totales} - \frac{E_{azar}}{k-1}$$

Si un sujeto no juega al azar pero se le aplica la corrección a sus equivocaciones puede argumentarse que la fórmula sobrecorrige a la baja la puntuación del sujeto. El problema es cómo distinguir los errores debidos a azar y las equivocaciones legítimas. En muchas situaciones reales

esto puede ser prácticamente muy difícil o francamente imposible.

Un mecanismo posible, aunque poco riguroso, consiste en introducir correcciones de la corrección que estimen, aun por algún procedimiento subjetivo, un margen para las equivocaciones.

Si calculo el coeficiente de correlación de Pearson "a mano" con datos de 10 casos y dos variables de dos cifras cada una ¿cuántas veces llegaré a un error de cálculo después de repasar dos veces debido a errores mecánicos? ¿una vez de cada 10, una de cada 25...? Esa probabilidad podría tenerse en cuenta para suavizar la fórmula de corrección al azar.

Una sugerencia para intentar mejorar la capacidad de la fórmula de corrección del azar de reflejar la puntuación debida realmente a capacidad, y no a otras cosas, consiste en tener en cuenta el nivel de "riesgo" que parece haber asumido la persona al contestar. Puede utilizarse el número de omisiones para estimar ese nivel de riesgo.

Por ejemplo, si una persona, en una prueba verdadero-falso de 50 items acierta 25 y falla 25 es bastante razonable pensar que ha jugado al azar. Se aplica la fórmula y el resultado es cero puntos. Esto parece bastante razonable, aunque en rigor nunca sabemos exactamente si los errores son o no equivocaciones.

Pero si ante esa misma prueba otra persona ha acertado 25 items y ha fallado 1 parece poco razonable pensar que de los 26 contestados sabía solo 24 (que acertó), y jugó al azar con dos de los que no sabía, acertando uno y fallando el otro (Este modelo un poco absurdo en este caso es el que sustenta la fórmula de corrección del azar). Para este caso parece más razonable pensar que la persona creía saber 26 y contesto 26, pero en uno se equivocó. Sin pretensión de acertar por azar. Paradójicamente una persona que sabe todo un conjunto de items y solo contesta los que sabe solo puede equivocarse y, por tanto, solo puede ser perjudicada con la fórmula de corrección al azar.

Estos razonamientos insinúan que la fórmula podría ser corregida en la dirección de que, existiendo condiciones de contenido que se presten a la aparición de equivocaciones, los errores deberían pesar razonablemente más a medida que creciera el número de errores, hasta igualar el peso concedido al error por la corrección al azar en el error enésimo. Un modo sencillo e imperfecto de hacer esto es evitar que los h primeros errores descuenten, donde h depende del tamaño de la prueba y de la probabilidad estimada de equivocaciones.

Otro problema importante es el del conocimiento parcial de la respuesta correcta o de los distractores. Este conocimiento parcial introduce un "riesgo parcial de error". Es decir, altera, de hecho, la probabilidad de acertar por

azar de un modo, a priori, impredictible (que puede ser distinto para cada ítem y para cada sujeto).

La aplicación de este principio y una calculadora al lado (aunque el examen no sea de estadística ni de psicometría) para calcular el punto óptimo de riesgo a asumir "ayuda" mucho a muchos. Supóngase una prueba de 5 alternativas con 30 items y en la que se aplica la fórmula de corrección al azar. Primero hay que clasificar los items por el grado de probabilidad de acierto. Para ello hay que establecer el grado de certeza (subjetivo) con que se conoce cada alternativa. Por ejemplo, es posible que después de un primer análisis se esté seguro del acierto en 12 items. Perfecto. Desde luego estos hay que contestarlos, pero, hay que advertir que dada la naturaleza de la prueba incluso en lo que consideramos certeza subjetiva puede haber algún error, y, además, necesitamos todavía "tres puntos limpios más". Podemos estimar la probabilidad de error en ítems de los que estamos "seguros" en un 10%, de modo que en 12 items "seguros" puede haber unos dos errores, que exigen uno o dos aciertos más, más tres puntos más que nos faltaban, en total necesitamos contestar entre 4 y 6 items más para conseguir 15 puntos "limpios", una vez descontados los errores. Ahora viene lo más interesante de la tarea. Hay que analizar cuidadosamente cada uno de los items no seguros y aprovechar bien el conocimiento parcial de los mismos. Hay que juzgar el grado de conocimiento de cada alternativa. Por ejemplo, en un ítem podemos desconocer la respuesta segura, pero estar seguros de que no es la alternativa a, ni la c ni la d. Eso aumenta las probabilidades de acertar hasta un 50%, comparativamente, si nos equivocamos solo perdemos un 25% de punto. Este desequilibrio entre el coste del error y la probabilidad de acierto vuelve la apuesta probabilísticamente injusta a nuestro favor y en él radica la clave de la ganancia debida a conocimiento parcial. En principio es "rentable" contestar al azar todo ítem donde la probabilidad de acierto sea mayor que la penalización por error. Obsérvese que si hay diez items así, por azar acertaríamos 5, pero la penalización por los otros cinco errores sólo sería igual a 1'25, la ganancia es de 3'75 -aplíquese la fórmula-. Un punto esencial para conseguir este efecto es atribuir correctamente las probabilidades subjetivas y jerarquizar bien los items desde el más probable al menos probable de ser acertado. Después se van contestando de fácil a difícil hasta detenerse en el punto donde el riesgo de sufrir una penalización iguala la probabilidad de acertar por azar. Este proceso se realiza de forma más o menos intuitiva al contestar pruebas objetivas. Si se realiza de forma sistemática probablemente (todo son probabilidades) se mejorarán los resultados. Por supuesto estas dificultades, desde el punto de vista de que la puntuación refleje la verdadera aptitud (y no una mezcla de esta con la habilidad para manejar probabilidades) contribuyen a recomendar las pruebas de k=2 donde estos problemas, aunque existen, están en su expresión mínima.

En síntesis, la fórmula de corrección del azar es perfectamente correcta si se cumplen perfectamente sus supuestos y condiciones de aplicación. Como estas, cuando se cumplen, solo se cumplen imperfectamente, cuando se utiliza se hará bien en tener en cuenta sus limitaciones y, si procede, intentar paliarlas -que no resolverlas-.

La experiencia práctica insinúa otras muchas dificultades. Por ejemplo, las pruebas objetivas corregidas con esta fórmula tienden a concentrar los resultados en una franja alrededor de alguna zona central, (sea en torno al valor medio que sea,) con cierta independencia del esfuerzo de estudio de los examinados. Para aquellos para quienes el mundo es un edificio hecho de curvas normales esto se explica, ¿cómo no?, por el hecho de que las aptitudes, conocimientos, etc. tienden a distribuir normalmente. Sin embargo, muchos estudiantes interpretan legítimamente que con este sistema de evaluación estudiando relativamente poco casi "aprueban" o "aprueban" y estudiando relativamente mucho "casi obtienen notable" o simplemente "aprueban" -especialmente, pero no exclusivamente, con las pruebas de alternativas al uso, medianamente mal diseñadas-. Aún conociendo muy bien toda la materia es muy difícil acertar todo (el azar solo puede perjudicar); aun no sabiendo nada es muy difícil fallar todo (el azar solo puede beneficiar). Parece que el método "expulsa", o al menos contribuye a "expulsar", los resultados de los topes de la escala (aplíquese aquí el concepto de regresión a la media que hemos visto muchas páginas atrás). Para algunos estudiantes la afirmación anterior no rige, y obtienen malas puntuaciones altamente improbables por azar (mejorarían con la ayuda de un dado o una moneda no lastrada). Para estos estudiantes desafortunados en sus puntuaciones se aplica fuertemente el principio de que "los distractores han de ser máximamente atractivos para los menos capaces y mínimamente atractivos para los capaces". Un principio que parece lógico, pero que atenta obviamente contra la atribución de los errores al azar. Un distractor no tendría que ser atractivo para un sujeto "capaz", pero para otro "no-capaz" precisamente no tendría que ser más atractivo que la respuesta correcta. Dudo mucho que se pueda ser tan sutil al elaborar buenos distractores.

Creo que el *fenómeno de "expulsión de las puntuaciones de las colas"* en las pruebas objetivas sucede no solo por razones técnicas, sino también por otras psicológicas o pedagógicas, como la incapacidad de las pruebas objetivas para detectar la genialidad, la creatividad, la estupidez o las patologías psíquicas más allá de los límites de la escala. (Por ejemplo, el sistema actual de evaluación masivo mediante pruebas objetivas en las licenciaturas de psicología (y en otras, claro) tiende a garantizar que un sujeto psicótico (y desde luego uno neurótico) medianamente aceptable estudiante, -no hace falta más,- se graduará, y con un poco de suerte, llegará al grado de doctor (En ese periodo de 4 a 7 años de estudios puede que nadie haya hablado personalmente con él; la habilidad

básica tangible que se la habrá requerido es ser capaz de hacer cruces sobre un papel y, desde luego, a ningún estudiante -ni profesor, dicho sea de paso- se le evalúa jamás su salud mental. Puede que hasta mencionar estos casos sea sancionable según el código deontológico de la profesión -que en ningún lado dice que es obligación del psicólogo contribuir a defender a la sociedad de desaguisados patológicos conducidos por colegas-). Por supuesto, más allá del tema de la corrección del acierto por azar quiero enfatizar que, en mi opinión, las pruebas objetivas no son ni el único ni el mejor procedimiento de medición para *todos* los objetivos psicológicos o pedagógicos. Estas son reflexiones subjetivas que tienen un mal apaño en condiciones de necesidades de evaluación masiva de colectivos de personas, tal como sucede en nuestras universidades, pero también en psicología educativa o en psicología industrial. La conclusión no es, obviamente, que las pruebas objetivas "son malas" o que "deben abandonarse". La conclusión es que sirven para ciertas cosas (y eso cuando están bien hechas) pero no para otras. Lo mismo cabe decir de cualquier test: hay que tener bien claras sus limitaciones.

Cuando se realiza el análisis de la dificultad, se puede utilizar el *índice de dificultad corregido el acierto por azar*, ó ID corregido, ó "p corregido". Basta descontar en la fórmula de ID aquellos aciertos atribuibles al azar bajo las condiciones que hemos analizado:

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

$$ID_{corregido} = \frac{A - \frac{E}{k-1}}{N}$$

Aunque esta fórmula quizás es más realista bajo las condiciones analizadas, no suele utilizarse. Desde luego no puede incluirse este "p corregido" en las fórmulas de varianza, correlación etc. que utilizan p.

Ejemplos

Indice de discriminación basado en p

Tenemos un test de aptitudes y estamos interesados en obtener el índice de discriminación basado en p para el ítem 1. Para ello, calculamos la puntuación total del test que obtiene cada persona de una muestra de N=300, y obtenemos el valor de la mediana en esa puntuación total.

Utilizando la mediana como punto de corte descomponemos la muestra total en dos submuestras, inferior y superior, de 150 personas cada una.

En la submuestra inferior aciertan el ítem 25, mientras que en la submuestra superior lo aciertan 125. Calculad el índice de discriminación basado en p.

Solución:

$$p_i = \frac{25}{150}$$

$$p_s = \frac{125}{150}$$

$$p_i = \frac{25}{150} \quad p_s = \frac{125}{150} \qquad D = \frac{125}{150} - \frac{25}{150} = \frac{100}{150} = 0'6667$$

Según la clasificación de Ebel. dado que el índice de discriminación basado en p es mayor que 0'39 podemos decir que el ítem es muy discriminativo, que "funciona muy bien" en el propósito de distinguir entre los más capaces y los menos capaces.

Fórmula de corrección de la respuesta al azar

Suponiendo que se dan las condiciones adecuadas para aplicar la fórmula de corrección, una persona, en una prueba objetiva de 20 ítems de tipo V/F, ha acertado 15 items y ha fallado 3. ¿Qué puntuación le corresponde en una escala de 0 a 10?

Solución:

$$A_{ap} = A_{tot} - \frac{E}{k-1} \rightarrow A_{ap} = 15 - \frac{3}{2-1} = 12$$

Puesta la calificación en la escala usual de 0 a 10 tenemos:

$$x = \frac{12 \cdot 10}{20} = 6$$

La persona ha obtenido un 6 como "calificación".