Análisis de ítems en las pruebas objetivas

• Universidad Pontificia Comillas • Madrid •
Facultad de Ciencias Humanas y Sociales

Índice

1. El contexto: las pruebas objetivas

Al estudiar el *análisis de ítems* de las pruebas objetivas, hay que tener en cuenta *todo el contexto* de estas pruebas (ventajas, inconvenientes, tipos de preguntas, etc.) aunque aquí no tratamos del resto de los temas que se pueden y deben tratar a propósito de las pruebas objetivas.

Es obvio por otra parte que las pruebas objetivas no son el único método de evaluación, ni el mejor necesariamente.

No sobra recordar que el término *objetivo* tiene connotaciones equívocas: en las pruebas objetivas la *corrección* sí es objetiva (una respuesta o está bien o está mal), pero tanto la *formulación* de la pregunta *(qué* y *cómo* se pregunta) como dónde se pone el *mínimum* para el *apto* son ya decisiones subjetivas del profesor.

Las *pruebas objetivas (tipo-test)* pueden ser muy cómodas para el profesor, sobre todo con clases numerosas

1º Fundamentalmente porque se evita la tediosa tarea de corregir, que es la dificultad sentida de manera más inmediata con las preguntas abiertas (las pruebas objetivas se pueden corregir con *lectura óptica*, y aunque se corrija sin estas ayudas, la tarea es mecánica, simple, e incluso delegable).

2º Con las pruebas objetivas es más sencillo establecer criterios de calificación y también pueden justificarse mejor estos criterios, al menos aparentemente, a partir de un determinado número de respuestas correctas.

Una consecuencia de estas ventajas para el profesor es la proliferación de pruebas objetivas que con mucha frecuencia son de *mala calidad*, sobre todo por dos razones:

1º No es tan fácil redactar *buenas preguntas* objetivas, sobre todo si se quiere comprobar algo más que pura memorización y estimular un estudio inteligente.

2º No es frecuente *planificar* estas pruebas aunque sea de manera muy elemental, por lo tanto puede haber más preguntas de lo que es fácil preguntar y no más preguntas de lo más importante (el examen puede quedar desequilibrado, el *apto* puede depender de preguntas triviales, etc.); el disponer de un *banco de preguntas* tampoco es precisamente la solución a esta frecuente falta de planificación de las pruebas objetivas.

Es obvio por otra parte que la *calidad de las preguntas* (objetivas o de otro tipo) no es un tema irrelevante si pensamos que *el qué y cómo estudia el alumno* (y consecuentemente cómo se forma o cómo se deforma) *depende del tipo de prueba y de pregunta (o ejercicio) esperado*.

Una manera de mejorar la calidad de estas preguntas objetivas es precisamente analizarlas, aunque para preparar buenas preguntas objetivas habría que abordar también los dos puntos antes indicados:

1º *Cómo redactar buenas preguntas* (que no sean casi exclusivamente de memoria, que comprueben los objetivos propuestos, que estimulen un estudio inteligente, etc.);

2º Cómo *planificar las pruebas objetivas* para que el conjunto de la prueba esté equilibrado en función de la importancia de los diversos temas y objetivos.

Aquí no tratamos sobre la redacción de las preguntas y la planificación de la prueba, aun así los análisis que vamos a proponer nos puedan dar buenas pistas para mejorar las preguntas en ediciones sucesivas.

Las pruebas objetivas se prestan a hacer una serie de análisis de interés que pueden referirse:

a) A toda la prueba
b) A cada pregunta en particular.

Aquí tratamos sobre todo del análisis de cada ítem o pregunta.

2. Utilidad del análisis de ítems

Por qué puede ser de interés el analizar las pruebas objetivas:

1º Para ir mejorando su *calidad.* Estos análisis aportan información no ya sobre los alumnos, sino sobre *cada una de las preguntas*.

El hacer una buena prueba objetiva, incluso una mala prueba objetiva, supone un tiempo y un esfuerzo que hay que hacer rentables. Si acumulamos experiencia sin hacer nunca ningún análisis, podemos estar haciendo permanentemente pruebas objetivas de calidad muy mediocre *(diez años de experiencia no es lo mismo que un año de experiencia repetido diez veces;* no por hacer muchas veces lo mismo lo haremos necesariamente mejor).

La información que nos dan estos análisis nos permite ir mejorando las pruebas sucesivas que vayamos haciendo, aprovechando nuestro propio trabajo. Estos análisis facilitan por lo tanto la *autoevaluación* del profesor y el ir mejorando sus tareas como profesor.

2º Algunos de estos análisis aportan *información útil para comentarla con los mismos alumnos,* y darles un *feedback* matizado sobre su aprendizaje. Esta información, que puede ser muy específica, puede ayudar a caer en la cuenta de errores generalizados, a entender puntos difíciles, a condicionar un estudio posterior de más calidad, etc.

3º También nos aportan datos que pueden influir indirectamente en nuestros *criterios de calificación;* al menos disponemos de una información más completa y fácil de entender (por ejemplo podemos descubrir preguntas ambiguas, o con dos respuestas correctas, o con la clave de corrección equivocada, o preguntas con un nivel de dificultad mayor del pretendido, etc.).

4º Por otra parte todos estos análisis son *fácilmente programables,* y si utilizamos una hoja de respuestas de *lectura óptica* y un programa adecuado de ordenador (programa que es fácil preparar) casi sin darnos cuenta podemos acumular una información muy útil, incluso para trabajos de investigación.

4º El beneficio que podemos obtener de estos análisis compensa el tiempo o esfuerzo extra que pueden suponer; beneficio en términos de:

- Mejorar la *calidad de las preguntas*
- Dar a los alumnos *una información más específica* sobre sus aciertos y errores, con la consiguiente *mejora de la calidad de la enseñanza y del aprendizaje* de los alumnos.

3. Análisis estadísticos convencionales

No vamos a tratar aquí con amplitud sobre los análisis de carácter más estadístico (o *psicométrico)* que cabe hacer en estas pruebas, pero sí recordamos en primer lugar los *análisis estadísticos* más convencionales; unos se refieren a toda la prueba (*fiabilidad, error típico*) y otros a cada ítem (*correlación ítem-total*)[1].

3.1. Análisis referidos a toda la prueba

Además de los datos descriptivos básicos (como son la media aritmética y la desviación típica), podemos calcular el *coeficiente de fiabilidad* y el *error típico*.

3.1.1. *El coeficiente de fiabilidad*

El coeficiente de fiabilidad es una estimación de la correlación esperada con una prueba semejante y por lo tanto este coeficiente de fiabilidad indica *en qué medida en exámenes semejantes los alumnos hubieran quedado ordenados de manera parecida.*

Para interpretar estos coeficientes de fiabilidad en *exámenes convencionales* o pruebas de rendimiento hay que tener en cuenta tres factores que inciden en la magnitud de este coeficiente:

1. La *homogeneidad de los ítems:* en la medida en que los ítems *midan lo mismo* la fiabilidad será mayor; con preguntas muy distintas y poco relacionadas entre sí la fiabilidad será más baja.
2. Las *diferencias entre los examinados* (homogeneidad de la muestra); si los sujetos tienen resultados muy parecidos la fiabilidad bajará (no se puede *clasificar, ordenar* bien a los muy semejantes).
3. El *número de ítems* porque a mayor número de ítems los alumnos quedan mejor diferenciados.

Fundamentalmente la fiabilidad depende de las diferencias entre los sujetos por lo que se puede cuestionar la fiabilidad de un test o de una prueba objetiva como indicador *necesario* de su calidad: si todos saben todo o casi todo (o casi nada), la fiabilidad tiende a bajar y esto no quiere decir que el test sea malo o que se trate de un mal resultado.

Un coeficiente de fiabilidad alto *(consistencia interna)* es claramente deseable cuando las diferencias entre los sujetos son *legítimas y esperadas;* y esto es lo que suele suceder en los *tests psicológicos*, y también en exámenes finales, sobre todo si son más bien largos, y con más razón en clases numerosas y donde es razonable esperar diferencias en rendimiento. Una fiabilidad alta nos dice que el examen *deja a cada uno en su sitio;* en exámenes parecidos (con otras preguntas del mismo estilo) los alumnos quedarían *ordenados* de manera semejante.

No hay que olvidar que una fiabilidad alta no es sinónimo *sin más* de calidad porque puede faltar lo que es más importante, la *validez:* preguntas que se pueden responder

[1] Más información sobre el *coeficiente de fiabilidad* y el *error típico* en Morales, Pedro *La fiabilidad de los tests y escalas.* Madrid: Universidad Pontificia Comillas http://www.upco.es/personal/peter/estadisticabasica/Fiabilidad.pdf; publicado en el capítulo 6 de Morales (2008). En este documento el apartado 11 está dedicado a las pruebas escolares. Otros temas relacionados con las pruebas objetivas, como la *adivinación* y *diversas alternativas para corregir estas pruebas*, los tratamos en Morales, Pedro, *Las pruebas objetivas: normas, modalidades y cuestiones discutidas* http://www.upcomillas.es/personal/peter/otrosdocumentos/PruebasObjetivas.pdf (última revisión, 17, Diciembre, 2006).

correctamente de memoria cuando lo que queremos es comprobar comprensión o interpretación, etc.

3.1.2. *El error típico de las puntuaciones individuales*[2]

El *error típico* también se refiere a toda la prueba y en la práctica puede ser más útil que el coeficiente fiabilidad. Suele denominarse *error típico de la medición* y se aplica a cada resultado individual. El *error típico* se deriva del coeficiente de fiabilidad y viene a indicar el *margen probable de oscilación* de las puntuaciones de cada sujeto de unas ocasiones a otras en exámenes hipotéticamente semejantes; podríamos denominarlo informalmente el *coeficiente de buena-mala suerte*.

El error típico puede servir para *relativizar* los resultados individuales (por ejemplo, para sumar el margen probable de error o de suerte en casos límite). Equivale a una *desviación típica* y se interpreta de manera semejante en relación con la distribución normal (más o menos cada alumno hubiera quedado en el 95% de las veces entre la puntuación de hecho obtenida *más menos* 1.96 errores típicos).

3.2. Análisis de cada pregunta y de cada alternativa

Estos análisis (denominado convencionalmente *análisis de ítems*) son los que más nos interesan en este momento.

3.2.1. *La correlación ítem-total*

Se trata ahora de un dato de cada ítem e indica en qué medida un ítem *discrimina* (diferencia a los que saben más de los que saben menos); este tipo de información lo podemos obtener también con los índices que vamos a ver a continuación.

Aunque esta correlación suele denominarse *correlación ítem-total*, en realidad se trata de la *correlación de cada pregunta con la suma de todas las demás;* es decir, del total menos el ítem que estamos analizando (con más propiedad suele denominarse también *correlación ítem-total menos el ítem*).

Lo que expresa esta correlación (como cualquier correlación) es en qué medida el responder correctamente a un ítem está relacionado con puntuar alto en todo el test. Esta información es semejante a la que nos da el *índice de discriminación* que vamos a ver aquí:

a) Una correlación próxima a *cero* quiere decir que el responder bien a esa pregunta no tiene que ver con estar bien en el conjunto del examen.

b) Una correlación *negativa*, sobre todo si es de cierta magnitud, quiere decir que el responder bien a esa pregunta está relacionado con estar más bien mal en el conjunto de la prueba (en principio se trata de una mala pregunta, o quizás hay un error en la clave de corrección).

c) Una correlación *positiva* quiere decir que el responder bien a esa pregunta está relacionada con un buen resultado en el conjunto de la prueba. Los ítems con mayores correlaciones positivas son los más discriminantes, los que mejor diferencian a los mejores y peores alumnos.

[2] Explicado con más amplitud en Morales (2008), p.214.

3.2.2. *La correlación de cada alternativa con el total*

Cuando hablamos de la correlación ítem-total, nos referimos a la correlación entre escoger la *respuesta correcta* en cada ítem y puntuar más o menos alto en el total de la prueba; *también cabe calcular la correlación entre escoger 'cada una' de las alternativas falsas y el total del test.* Lo que podemos esperar es que el escoger una alternativa falsa correlacione negativamente con el total (las alternativas falsas las escogen los que en conjunto están peor). No es frecuente hacer este análisis pero puede ser muy informativo.

Aquí no tratamos con amplitud estos análisis estadísticos, pero sí es oportuno recordar que son relativamente sencillos, se pueden programar con toda facilidad y nuestra tarea se reducirá a interpretar y valorar los resultados. Ahora nos vamos a limitar a los análisis más frecuentes y sencillos que suelen hacerse con cada pregunta o ítem.

4. Análisis de las diversas alternativas: tabulación de las respuestas

Este análisis, que se limita a una mera *tabulación* de las respuestas:

a) Es muy sencillo y también se puede programar
b) Aporta una información de interés que se interpreta con mucha facilidad y de manera intuitiva, sin necesidad de análisis estadísticos.

Posiblemente es el análisis en principio más útil para el profesor. Se trata de *organizar las respuestas* de manera que permitan una reflexión rápida sobre las preguntas y sobre los alumnos.

El proceso es el siguiente:

1º *Se ordenan los sujetos de más a menos según su puntuación total* en la prueba (según el número de respuestas correctas, no por las notas que se les asignen) y se seleccionan el 25 % con puntuación total más alta (grupo *superior)* y el 25 % con puntuación total más baja (grupo *inferior).* También se escogen a veces el 27% o el 33% con totales más altos y más bajos, pero el 25% es un porcentaje cómodo y suficiente.

2º *Se tabulan las respuestas de estos dos grupos en cada ítem,* de manera que se pueda ver cuántos de cada grupo, superior e inferior, han escogido cada opción.

Esta tabulación de las respuestas se presta ya a muchas observaciones de interés para el profesor que ha redactado los ítems. Lo veremos mejor con un ejemplo ficticio (tabla 1)[3].

[3] Un ejemplo comentado puede verse en (página 108) Case, Susan M. and Swanson, David B. (2001). *Constructing Written Test Questions For the Basic and Clinical Sciences,* 3rd Edition. Philadelphia: National Board of Examiners (181 páginas). http://www.nbme.org/PDF/2001iwg.pdf (un excelente manual sobre pruebas objetivas en medicina).

	alternativas (la *respuesta correcta* se indica con un *)			
preguntas	**A**	**B**	**C**	**D**
ítem nº 1	**10 *** / **0**	**0** / **2**	**0** / **0**	**0** / **8**
ítem nº 2	**5** / **1**	**5 *** / **0**	**0** / **7**	**0** / **2**
ítem nº 3	**6** / **1**	**0** / **1**	**2** / **0**	**2 *** / **8**

Tabla 1

En este ejemplo suponemos que tenemos 40 alumnos, de estos 40 alumnos hemos escogido los 10 con el total más alto y los 10 con el total más bajo (el 25% de los mejores y peores resultados).

En la figura 1 tenemos cómo se han distribuido las respuestas entre las cuatro opciones de cada pregunta: en el *supuesto numerador* tenemos el número de alumnos del grupo superior que ha escogido cada opción, y debajo el número de alumnos del grupo inferior que ha escogido esa misma opción; la respuesta correcta está señalada con un *asterisco.*

Esta *mera tabulación de frecuencias* se presta ya una serie de consideraciones aun sin conocer el contenido de las preguntas (como sucede en este ejemplo); por ejemplo:

- Ítem nº 1:
  La alternativa correcta (la A) la han escogido todos y solos los del grupo superior: se trata de una pregunta que discrimina muy bien; diferencia claramente a los que saben de los que no saben.

  Los del grupo inferior se han ido casi todos a la opción D: es una *buena alternativa incorrecta*, que atrae al que no sabe o no entiende; sabemos dónde o porqué fallan los que saben menos (qué confunden con qué…); un resultado de este tipo se presta a una buena explicación a la clase porque nos dice dónde fallan los que menos saben.

  Esta presentación de los datos puede tener un claro valor *diagnóstico*. La alternativa C no la ha escogido nadie, ni siquiera de los que menos saben. En otra edición convendrá modificarla, y si observamos que con frecuencia hay alguna opción que no la escoge nadie o muy pocos (y esto sucede con mucha frecuencia), podremos pensar en pasar de cuatro a tres alternativas.

- Ítem nº 2:
  Los que más saben se distribuyen entre dos alternativas, la B (correcta) y la A (incorrecta). Es posible que las dos sean correctas, o que la pregunta sea ambigua; al menos se trata de una pregunta que conviene examinar. La opción C también es un buen *distractor* que atrae a los que no conocen la respuesta correcta.

- Ítem nº 3:
  Aquí tenemos un resultado anómalo: los que aciertan son sobre todo los que menos saben. El grupo superior prefiere la opción A (incorrecta). Pregunta que podemos hacernos: ¿Estará mal la *clave* de corrección? En cualquier caso una pregunta que

favorece a los que menos saben es en principio una mala pregunta y habrá que examinarla.

Naturalmente no hay *interpretaciones automáticas*, pero esta tabulación puede decir mucho al profesor que conoce sus propias preguntas.

Esta simple tabulación de las respuestas puede ser muy informativa:

- Para *comentar los resultados en clase* (con la consiguiente reflexión, corrección de errores, etc.)
- Para ir *mejorando la redacción de los ítems* (y entonces el tiempo y la experiencia serán rentables.)

También pueden tabularse las respuestas dividiendo a la clase en tres segmentos: el superior, el medio y el inferior, pero el tener en cuenta las respuestas de los dos grupos extremos es suficiente para el fin que se pretende.

Los índices que vamos a exponer a continuación son ampliamente utilizados, sin embargo a profesores no acostumbrados a análisis numéricos pueden resultarles poco claros o simplemente incómodos; en cambio éste observar cómo se distribuyen las respuestas en cada ítem de los que más y menos saben, ofrece buenas pistas de reflexión de manera intuitiva y nada complicada (este *nivel* de análisis puede ser suficiente en seminarios y actividades de formación del profesorado).

5. Índices de dificultad y discriminación

Estos índices no se calculan con toda la muestra sino, como en el caso anterior, con el 25% con una puntuación total *más alta* en todo el test y con el 25% con una puntuación total *más baja;* también suelen hacerse a veces con otras proporciones (como el 21%, 27%, 30%) pero el 25% es suficiente[4]. El número de sujetos en ambos grupos es por lo tanto el mismo; sólo se analizan las respuestas del 50% de los alumnos. (se prescinde del 50% central). Este tipo de análisis es análogo al que se hace cuando se construye una escala de actitudes. Los símbolos utilizados son los siguientes (tabla 2)

Símbolos utilizados	
N = número de sujetos en uno de los dos grupos (los dos grupos tienen idéntico número de sujetos)	AS = número de *acertantes* en el *grupo superior* (con puntuación total más alta)
N + N = número total de sujetos analizados	AI = número de *acertantes* en el *grupo inferior* (con puntuación total más baja)

Tabla 2

[4] Pueden verse numerosos documentos sobre estos análisis (poniendo *item analysis* en *search*) en The University of Washington's Office of Educational Assessment, http://www.washington.edu/oea/

5.1. Índice de dificultad

Índice de dificultad

$$Df = \frac{AS + AI}{N + N}$$

Indica la *proporción de aciertos* (tanto por ciento si multiplicamos por 100) en la muestra de alumnos que estamos utilizando (el 50% del total, los dos 25% con puntuaciones totales extremas).

Este índice es la *media* de este 50% de sujetos analizados. También la media del ítem, obtenida con toda la muestra, nos indica el grado de dificultad (media más alta, ítem más fácil), sin embargo este *índice de dificultad* suele utilizarse rutinariamente junto con los índices de discriminación.

El término *índice de dificultad* se presta a equívocos: un índice mayor indica una pregunta más fácil (mayor proporción de aciertos), no más difícil (quizás podría denominarse con más propiedad *índice de facilidad).*

5.2. Índices de discriminación

Los *índices de discriminación* expresan en qué medida cada pregunta o ítem diferencia a los que más y menos saben. Decimos *índices* (en plural) porque hay dos ampliamente utilizados (quizás más el primero).

5.2.1. Índice de discriminación 1

Índice de discriminación 1:

$$Dc_1 = \frac{AS - AI}{N}$$

Es la diferencia entre dos proporciones, los acertantes del grupo superior menos los acertantes del grupo inferior: (AS/N)-(AI/N); como los denominadores son iguales (idéntico número de sujetos en cada grupo) la fórmula queda simplificada.

a) Es el índice probablemente más utilizado consiste en *la diferencia entre dos proporciones:* proporción de aciertos en el grupo superior (AS/N) menos proporción de aciertos en el grupo inferior (AI/N). Expresa por lo tanto hasta qué punto la pregunta *discrimina, establece diferencias*, contribuye a situar a un sujeto en el grupo superior o inferior. A mayor diferencia en número de acertantes entre los grupos superior e inferior, el ítem es más discriminante, contribuye más a situar a un sujeto entre los primeros o entre los últimos.

b) Equivale a una *estimación* de la *correlación ítem-total* y puede interpretarse de la misma manera; sin embargo puede ser más clara una interpretación literal *(diferencia entre dos proporciones).*

c) Los valores extremos que puede alcanzar este índice son 0 y *más/menos* 1.

Si todos responden correctamente (pregunta *muy fácil):* $Dc_1 = \frac{N - N}{N} = 0$

Si todos se equivocan (pregunta *muy difícil)*: $Dc_1 = \frac{0 - 0}{N} = 0$

Es decir, las preguntas muy fáciles o muy difíciles no discriminan, no establecen diferencias; nos dicen que todos saben o no saben una pregunta, pero no *quién sabe más y quién sabe menos.* Estas preguntas no contribuyen a la fiabilidad, pero eso no quiere decir necesariamente que sean malas preguntas (son malas para *discriminar*…).

Si *todos* y *solos* los del grupo superior aciertan la pregunta, tendremos que: $Dc_1 = \frac{N - 0}{N} = 1$

Si acertaran solamente los del grupo inferior tendríamos que $Dc_1 = \frac{0 - N}{N} = -1$

Por lo tanto 1 y -1 son los valores máximos de este índice. Las preguntas con *discriminación negativa* favorecen al grupo inferior y en principio deberían ser revisadas (posibilidades: preguntas mal formuladas, ambiguas, error en la clave de corrección, etc.)

d) Las preguntas que discriminan mucho *(diferencian bien a los que saben más de los que saben menos)* no son muy difíciles; tienden a ser de *dificultad media* (responde bien la mitad de los sujetos analizados). En este caso (discriminación máxima porque aciertan sólo y todos los del grupo superior) tendríamos que el índice de dificultad sería:

$$Df = \frac{N - 0}{N + N} = .50$$

e) Una limitación de este índice está en que el valor máximo de 1 sólo se alcanza cuando aciertan todos los del grupo superior y se equivocan todos los del grupo inferior. *Puede haber preguntas que discriminan bien pero que son difíciles* (y fallan algunos del grupo superior) *o son fáciles* (y las aciertan algunos el grupo inferior). Por estas razones algunos prefieren el otro índice de discriminación que expondremos a continuación (Dc_2), aunque se pueden programar y utilizar los dos.

Valores máximos del índice de discriminación

Puede tener su interés conocer el *valor máximo* que puede alcanzar este índice de discriminación. El valor máximo que puede tener de hecho este índice depende de la dificultad de la pregunta (fórmulas en la tabla 3).

Valores máximos del índice de discriminación (Dc_1)		
Cuando Df = .50 (*aciertan la mitad)* Dc_1 máximo = 1	Cuando Df > .50 *(aciertan más de la mitad)* Dc_1 máximo = 2 (1 - Df)	Cuando Df < .50 *(aciertan menos de la mitad)* Dc_1 máximo = 2Df

Tabla 3

El *índice máximo de dificultad cuando aciertan más de la mitad* es el que hubiera habido 1º manteniendo el mismo número de aciertos (grado de dificultad de la pregunta) pero de manera que 2º ninguno del grupo superior hubiera fallado.

Por ejemplo, en una clase de 40 sujetos tenemos que N = 10 (25% superior e inferior).

Si AS = 9 y AI = 3 tendremos que $Df = \frac{9 + 3}{10 + 10} = .60$ (60% de aciertos)

$$Dc_1 = \frac{9 - 3}{10} = .60$$

El *valor máximo de discriminación,* manteniendo los 12 aciertos, es el que hubiéramos obtenido si AS = 10 (todos los del grupo superior aciertan) y AI = 2 (los dos aciertos restantes se los dejamos al grupo inferior). En este caso el índice de discriminación hubiera sido

$$Dc_1 = \frac{10 - 2}{10} = .80 \text{ [ó 2 (1- .60) = .80]}$$

El índice máximo de discriminación cuando han acertado menos de la mitad, es el que hubiéramos obtenido si todos los acertantes pertenecieran al grupo superior. En el mismo caso anterior (una clase de 40 y N = 10), obtenemos estos resultados:

Si AS = 4 y AI = 2 tendremos que $Df = \frac{4 + 2}{10 + 10} = .30$ (30% de aciertos)

$$Dc_1 = \frac{4 - 2}{10} = .20$$

El índice máximo de discriminación en este caso (han acertado menos de la mitad) es el que hubiéramos obtenido si todos los aciertos estuvieran en el grupo superior (AS = 6) y todos los del grupo inferior se hubieran equivocado; el índice de discriminación hubiera sido:

$$Dc_1 = \frac{6 - 0}{10} = .60 \text{ [ó 2 (.30) = .60]}$$

Cuando el índice de dificultad es .50 (acierta el 50%) las dos fórmulas anteriores llevan al mismo resultado, y el índice máximo de dificultad es siempre 1.

No es fácil en la práctica establecer una *magnitud óptima* del índice de discriminación; una buena orientación es interpretar estos índices en *términos relativos* y examinar cuáles son más y menos discriminantes en una situación dada.

5.2.2. Índice de discriminación 2

Este índice es menos utilizado; cuando se habla de *índice de discriminación* sin más especificaciones hay que entender que se trata del índice anterior; sin embargo este segundo índice de discriminación es también informativo.

Índice de discriminación 2:

$$Dc_2 = \frac{AS}{AS + AI}$$

Este índice indica la *proporción de aciertos en el grupo superior con respecto al número total de acertantes.* Puede considerarse satisfactorio si al menos es superior a .50: esto quiere decir que más de la mitad de los acertantes pertenecen al grupo que *sabe más*.

Este índice es *independiente del grado de dificultad de la pregunta;* con el índice anterior nunca se llega al valor de 1 si falla alguno del grupo superior (preguntas más difíciles); en cambio este índice llega a 1 *si todos los acertantes, aunque sean pocos, pertenecen al grupo superior*. Este índice nos dice cuánto discrimina el ítem lo mismo si es muy fácil como si es muy difícil; de hecho se utiliza menos que el anterior pero también aporta una buena información.

Vamos a verlo con dos ejemplos:

1º Suponemos una pregunta *muy fácil;* con N = 10 en cada grupo (superior e inferior, en una clase de 40 alumnos), la aciertan los 10 del grupo superior y 9 del grupo inferior; los dos índices de discriminación serían estos:

$$Dc_1 = \frac{10 - 9}{10} = .10$$

El *primer índice* (Dc_1) nos dice que la pregunta apenas discrimina; es muy fácil;

$$Dc_2 = \frac{10}{10 + 9} = .526$$

El *segundo índice* (Dc_2) nos dice que aunque es una pregunta muy fácil, mas del 50% (casi el 53 %) de los acertantes pertenece al grupo superior; de fallar alguien esta pregunta, pertenece al grupo de los que menos saben (este índice debe alcanzar al menos el valor de .50).

2º Suponemos ahora una pregunta *muy difícil;* solamente la responden bien 2 alumnos del grupo superior y ninguno del inferior.

$$Dc_1 = \frac{2 - 0}{10} = .20$$

El *primer índice* (Dc_1) nos dice que la pregunta discrimina muy poco porque es muy difícil;

$$Dc_2 = \frac{2}{2 + 0} = 1$$

El *segundo índice* (Dc_2) nos dice que la discriminación es perfecta; aunque se trate de una pregunta muy difícil, de saberla alguien, éste pertenece al grupo superior, donde están los alumnos que más saben.

Podemos ver que este índice (menos utilizado que el anterior como ya se ha indicado) es sumamente útil, pues nos dice en qué medida una pregunta contribuye a distinguir a los que saben más de los que saben menos independientemente de la dificultad o facilidad de la pregunta. Los dos índices de discriminación se pueden programar y utilizar conjuntamente.

6. Índices de dificultad y discriminación referidos a todo el test

De manera análoga se pueden calcular los índices de dificultad y discriminación referidos a *todo el test:*

$$\textit{Índice de dificultad de todo el test} = \frac{\text{media}}{\text{número de ítems}}$$

Se trata simplemente de la *proporción de respuestas que corresponde a la media;* en un test de 40 preguntas si la media es igual a 30.5, el índice de dificultad será 30.5/40 = .76 (la media de respuestas correctas es del 76%).

Este índice es útil para comparar la dificultad de varios tests (o distintas partes del mismo test) sobre todo si tienen un número distinto de ítems.

$$\text{Índice de discriminación } \textit{de todo el test} = \frac{\text{puntuación más alta obtenida} - \text{puntuación más baja obtenida}}{\text{número de ítems}}$$

El número de ítems del denominador es la *diferencia máxima posible* (la que habría entre un sujeto que hubiera respondido bien a todos los ítems y el que no hubiera respondido

a ninguno). Por lo tanto este índice equivale a la *diferencia máxima obtenida* dividida por la *diferencia máxima posible* (o lo que es lo mismo, la *amplitud* dividida por el número de ítems). Si en un test de 40 preguntas la puntuación mayor es de 35 y la más baja es de 20, el índice de discriminación sería (35-20)/40 = .375.

La información que nos da este índice puede ser cuestionable porque se puede ver afectado por unas pocas puntuaciones extremas y muy atípicas, aun así puede ser útil para comparar en discriminación tests con distinto número de ítems o el mismo test en grupos distintos. También se puede calcular excluyendo a los sujetos con puntuaciones muy extremas y atípicas (y advirtiéndolo en este caso).

7. Valoración de estos índices

1. Estos índices *describen* cómo *ha funcionado* una pregunta en una situación dada; no hay que asociar necesariamente *juicios de valor sobre la calidad de la pregunta* al valor de estos índices (por eso decimos en primer lugar que estos índices *describen* qué ha sucedido; *luego vendrá nuestra valoración).*

   Las preguntas que son muy fáciles o muy difíciles, por ejemplo, no son discriminantes y tendrán una baja correlación ítem-total) y *tienen su lugar*. Otra cosa es cuando estos índices nos *sorprenden* porque no esperábamos estos resultados (si las examinamos podemos ver quizás que la pregunta es ambigua, que alguna alternativa está mal formulada, que la clave de corrección está equivocada, que hay más de una respuesta correcta, etc.).

   Un índice bajo de discriminación (o una correlación con el total muy pequeña) pueden estar indicando que esos ítems miden *algo distinto* que la mayoría del resto de los ítems (por ejemplo un ítem que mida comprensión o capacidad de aplicar principios puede tener un índice de discriminación bajo si la mayoría de los ítems son de memoria).

3. Estos índices (sobre todo el índice de discriminación 1, el más utilizado y del que suele tratarse cuando se habla del índice de discriminación) tienen la ventaja clara de que son muy fáciles de entender, pero son poco fiables calculados en muestras pequeñas (como son los alumnos de una clase); pueden variar mucho de muestra a muestra. Con muestras pequeñas describen bien lo que ha sucedido en *esa* muestra y permiten dar un *feedback* muy específico a los alumnos, pero hay que ser muy cauteloso cuando se trata de prescindir de algunos ítems en ocasiones sucesivas; con esta finalidad hay que utilizar muestras grandes (o acumular análisis). Cuando se descartan ítems en función de análisis hechos con muestras pequeñas se corre el riesgo prescindir de buenos ítems; por otra parte ningún análisis puede sustituir un examen cuidadoso de la formulación del ítem (Burton, 2001).

   Para *extrapolar* los resultados harían falta muestras grandes (N= 400, ó unos 100 en los grupos extremos; estas muestras se pueden obtener acumulando datos); sin embargo la experiencia dice que los índices obtenidos con grupos pequeños, si se mantiene constante el tipo de muestra, dan una buena idea de lo que se puede esperar en grupos similares.

3. La correlación ítem-total aporta una información semejante al índice de discriminación y puede ser preferible porque se basa en los datos de todos los sujetos. Si se ha impuesto más (al menos en textos de evaluación) el índice de discriminación es por la facilidad de cálculo antes de que se popularizaran los programas de ordenador. Sin

embargo los índices se discriminación siguen siendo más fáciles de entender para los que no están familiarizados con la estadística.

4. Las preguntas muy discriminantes (que por definición no suelen ser ni las más difíciles ni las más fáciles) nos indican *dónde* fallan, sobre todo, los que tienen malos resultados; pueden incluso indicar *por qué* fallan cuando varias preguntas muy discriminantes tienen alguna relación entre sí.

5. La discriminación supone *diferencias* (lo mismo que la *fiabilidad* calculada con todo el test) y el que haya diferencias *no es necesariamente un buen resultado,* por ejemplo cuando las preguntas son en principio fáciles, versan sobre objetivos mínimos, etc. Sí es, en cambio, importante que las preguntas (bastantes al menos) discriminen cuando se trata de *clasificar,* de *seleccionar,* etc., pero no es éste el caso en muchos exámenes convencionales.

6. En exámenes *largos* (sobre todo en *exámenes finales),* en los que se pregunta de todo, con grupos relativamente numerosos, la *no discriminación* (lo mismo que una fiabilidad muy baja) puede indicar que no se detectan diferencias que de hecho existen (por ejemplo, puede haber alumnos que saben más de lo que pueden manifestar en un determinado examen).

   En este tipo de exámenes habrá preguntas que no discriminen porque o son fáciles, o son importantes y todos las han estudiado; casi todos las responden bien y éste será un buen resultado; otras no serán discriminantes porque son muy difíciles y ya se contaba con ello (y tampoco tiene que valorarse como un mal resultado); pero en el conjunto del examen y para poder calificar con cierto matiz, debe haber preguntas de dificultad media que discriminen bien.

7. Las preguntas *muy discriminantes* (que nunca serán las más difíciles) pueden ser útiles en exámenes de segunda convocatoria, prescindiendo de lo muy fácil y de lo muy difícil; con exámenes más cortos obtenemos la información suficiente. Claro está que puede haber *otros criterios* para seleccionar estas preguntas, como son temas u objetivos determinados, al margen de que las preguntas discriminen mucho o poco.

8. No hay que olvidar, cuando se calculan e interpretan estos índices, que en principio una pregunta es buena:

   - Si es clara y está correctamente formulada,
   - Si permite comprobar el objetivo deseado,
   - Si condiciona en el alumno un tipo de estudio inteligente o al menos deseable
   - Y tampoco hay que olvidar que *una mala pregunta muy analizada sigue siendo una mala pregunta*

9. Estos índices *describen* cómo han *funcionado* los ítems en una muestra y situación concretas y son útiles para *evaluar las preguntas, sugerir qué se puede revisar,* etc., pero *malas preguntas* (triviales, que no comprueban nada importante, que no responden a los objetivos, que condicionan un estudio poco inteligente, etc.) pueden tener índices que podrían considerarse como óptimos (por ejemplo pueden discriminar muy bien). Es peligroso interpretar estos índices como *indicadores automáticos* de la *calidad* de una pregunta.

10. Estos índices (y cualquier otro análisis semejante) no son prueba de *validez*, es decir, de que realmente estamos comprobando lo que deseamos comprobar (comprensión, capacidad de análisis, etc.). La validez la verificamos con un cuidadoso examen de la formulación el ítem y también viendo su relación (de cada ítem, de bloques de ítems, de toda la prueba) con otros criterios.

11. El *análisis de las diversas alternativas* expuesto en la tabla 1, comprobando cuántos eligen cada una, en toda la muestra o mejor en los dos grupos extremos, es un análisis sencillo, fácil de entender y comunicar y que da una información sumamente útil *para ir mejorando las preguntas* en ediciones sucesivas sin necesidad de calcular ningún índice.

12. Estos índices (lo mismo que otros datos descriptivos como la media, la desviación y la correlación ítem-total) son sin embargo importantes:

    - para *comunicar* (y publicar) resultados,
    - para *resumir* la información y conservarla para una reflexión posterior,
    - para hacer algún tipo de *investigación*, etc.

13. ¿Y qué sucede con las *preguntas abiertas?*

    No estamos tratando de estas preguntas (u otro tipo de ejercicios, problemas, etc.) pero sí es útil advertir que se pueden hacer *análisis semejantes* si todas las preguntas se corrigen *con la misma clave* o con el mismo sistema de corrección o calificación (en vez de tener siempre el valor de 0 ó 1, como las preguntas objetivas, podrán puntuar 0 ó 1 o también de 0 a 2, de 0 a 5, etc., según como se establezca la clave de corrección).

    En estos casos:

    - El *índice de dificultad* es la *media* de cada ítem,
    - El *índice de discriminación* es la *diferencia entre las medias* de los dos grupos con puntuación total más alta y más baja.

      También podríamos utilizar como un indicador de la discriminación la *t de Student* o preferiblemente el *tamaño del efecto* (no es éste lugar para explicar estos cálculos) en vez de la mera diferencia entre las dos medias[5].

      Si utilizamos estos dos indicadores para apreciar en qué medida una pregunta diferencia a los que más y menos saben, las preguntas o ejercicios pueden tener claves de corrección distintas (los valores de la *t de Student* y del *tamaño del efecto* son independientes de la escala métrica utilizada y son comparables entre sí).

8. Bibliografía citada y direcciones sobre el *análisis de ítems* en Internet

Case, Susan M. and Swanson, David B. (2001). *Constructing Written Test Questions For the Basic and Clinical Sciences,* 3rd Edition. Philadelphia: National Board of Examiners (181 páginas). http://www.nbme.org/PDF/2001iwg.pdf

Burton, Richard F. (2001). Do Item-discrimination Indices really Help Us to Improve Our Tests? *Assessment and Evaluation in Higher Education,* Vol. 20, nº 3, 213-220

[5] Puede verse Morales (2008) cap. 8

Kehoe, Jerard (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, 4(10). http://pareonline.net/getvn.asp?v=4&n=10 This paper has been viewed 60,488 times since 11/13/99 (05, 05, 09).

Matlock-Hetzel, Susan (1997). *Basic Concepts in Item and Test Analysis*. Texas A&M University. http://ericae.net/ft/tamu/Espy.htm

Michigan State University. *Scoring Office*. *Item Analysis* http://www.msu.edu/dept/soweb/itanhand.html#uses

Morales Vallejo, Pedro (2008). *Estadística aplicada a las Ciencias Sociales*. Madrid: Universidad Pontificia Comillas.

Morales Vallejo, Pedro *La fiabilidad de los tests y escalas.* Madrid: Universidad Pontificia Comillas http://www.upco.es/personal/peter/estadisticabasica/Fiabilidad.pdf (última revisión, 1 de Mayo de 2007).

Morales, Pedro, *Las pruebas objetivas: normas, modalidades y cuestiones discutidas* Madrid: Universidad Pontificia Comillas http://www.upcomillas.es/personal/peter/otrosdocumentos/PruebasObjetivas.pdf (última revisión, 17, Diciembre, 2006)

The University of Texas at Austin, Measurement and Evaluation Center (MEC) *Item analysis* http://www.utexas.edu/academic/mec/scan/scanitem.html

The University of Washington's Office of Educational Assessment, http://www.washington.edu/oea/ (poniendo *item analysis* en *search*; numerosos documentos sobre análisis de ítems).