Anàlisi dels ítems

Albert Bonillo

PID_00198604

Universitat Oberta de Catalunya

Índex

Introducció

L'objectiu d'aquest mòdul és introduir l'estudiant en el tema de l'anàlisi d'ítems. Considerem un error que aquest tema no sigui present en tots els plans docents de l'assignatura de psicometria, ja que és prou important perquè valgui la pena estudiar-lo. És cert que, tradicionalment, l'estudi de la psicometria s'ha focalitzat més cap a l'estudi de les propietats dels instruments de mesura destinats a aspectes d'opinió i constructes psicològics, que cap als que mesuren coneixements o habilitats. No obstant això, el psicòleg de carrer treballa tant o més amb aquests últims que amb els primers.

Volem que l'estudiant sàpiga, des del principi, que aquest mòdul no el farà expert en cap dels aspectes que aborda. Li proporcionarà, ho desitgem i ho esperem, una bona introducció a cadascun dels temes, però és fàcil que per plaer –o necessitat professional, o totes dues coses– vulgui aprofundir en alguns dels aspectes estudiats. Recomanarem textos que sí els estudien en profunditat.

El mòdul s'inicia precisament distingint entre instruments en funció de l'objectiu. En segon lloc, i ja centrats en proves d'execució màxima, mostrarem quins són els aspectes que s'han de tenir en compte en la construcció dels seus ítems. En el tercer apartat veurem com s'analitzen les propietats psicomètriques de la prova i dels ítems a partir de la teoria clàssica de test (TCT). Veurem els conceptes de *dificultat* i *discriminació*, i aprendrem a valorar si un ítem és correcte o potser necessita una revisió. En el quart apartat, veurem una introducció a la teoria de resposta a l'ítem (TRI), que és una alternativa d'anàlisi a la TCT. Veurem la TRI de manera més succinta que la TCT. El model de TRI resol problemes teòrics de la TCT, però els càlculs d'aquesta última són més senzills i fàcilment aplicables que els de primera.

Deixarem per al final les conclusions que resumeixin tot el que hem presentat.

1. Tipus de proves

És tradicional que, quan des de l'àmbit de la psicologia parlem d'una prova –o d'un instrument de mesura–, pensem immediatament en una enquesta d'opinió, un test de personalitat o similars. Des del punt de vista del tipus de prova, aquestes que hem esmentat no són diferents del qüestionari de satisfacció sobre el servei que trobem a la sortida de molts hotels. Volen mesurar, en una persona, el valor determinat d'un constructe l'existència del qual es pressuposa.

Un cas diferent és una prova que vulgui ordenar els millors candidats a un lloc de treball. En aquest context, en què se suposa que hi ha un criteri –ser un bon treballador per al lloc ofert–, la mesura del constructe pot passar a un segon pla. L'objectiu de l'instrument és que cadascun dels ítems optimitzi la classificació correcta de les persones. Vegem, doncs, quines característiques tenen les proves en funció del que cerquen.

1.1. Proves d'execució típica enfront de proves d'execució màxima

Així doncs, i si classifiquem les proves per l'objectiu, en distingirem dos tipus bàsics. Anomenem *proves d'execució típica* –o *d'execució de trets*– les que mesuren aspectes no escalables o, dit d'una altra manera, les que plantegen preguntes que no tenen respostes ni correctes ni errònies, sinó que tracten d'aspectes d'opinió, de preferència o similars. Per contra, anomenem *proves d'execució màxima* les que avaluen constructes que sí són escalables, i que són aquelles en les quals té sentit parlar de respostes correctes o errònies. Un examen, un test d'intel·ligència o qualsevol instrument que mesuri l'aptitud es classificaria dins d'aquest epígraf.

Encara que tots els conceptes que hem vist fins ara en mòduls anteriors –*fiabilitat*, *validesa* i *transformació de les puntuacions obtingudes*– són aplicables a tots dos tipus d'instruments, les estratègies per a estudiar-los solen variar lleugerament i se solen estudiar aplicant-los a les proves d'execució típica. És cert que, per exemple, un test d'intel·ligència ha de ser fiable, però administrar-lo dues vegades en poques setmanes no pot tenir gaire sentit, ja que en l'interval els participants podrien haver obtingut la resposta correcta i contaminar així els resultats. No obstant això, sí té sentit repetir un test de personalitat amb pocs dies de diferència, i així comprovar si la mesura de l'instrument és tan estable com se suposa que és el constructe mesurat. En definitiva, les característiques que es volen estudiar depenen, és clar, de l'objectiu de l'instrument.

Sovint, el psicòleg professional no utilitza instruments estandarditzats, sinó que ha de crear ell mateix l'instrument. Si l'estudiant treballés en el departament d'RH d'una multinacional i aquesta li demanés una prova per a ocupar un lloc molt específic, què faria? Després de comprovar en el mercat que aquesta prova no existeix, hauria de crear-la. I hauria de fer-ho tenint en compte què es vol fer amb aquesta prova: seleccionar el millor treballador per a aquest lloc. Bé, i llavors què? Suposem que aquest lloc requereix certs coneixements. El psicòleg hauria de construir una prova que, a partir d'un nombre mínim d'ítems, pugui seleccionar el millor candidat.

Aprenguem, doncs, què s'ha de tenir en compte quan no hi ha més remei que crear una prova.

2. Directives en la construcció d'ítems

L'estudiant ja sap que l'objectiu principal d'aquest mòdul és mostrar com es mesura la qualitat d'un test de rendiment. Ara bé, no hem d'eludir l'explicació sobre què cal fer per a construir correctament una prova. Creiem que el treball d'un psicòleg no ha de ser únicament valorar si la prova està més o menys ben feta, sinó que també té molt a aportar en la seva construcció. Canviant totalment d'àmbit: no seria estrany que un arquitecte valorés edificis si no aprengués primer a construir-los?

Hi ha diversos treballs que exposen de manera molt exhaustiva quines són les directives que cal seguir per a construir correctament una prova d'execució màxima. Un dels primers i més coneguts és el de Haladyna, Downing i Rodríguez (2002). Es tracta, ni més ni menys, que de trenta-un criteris que s'han de seguir, classificats per apartats. Aquests criteris es refereixen al contingut de la pregunta (per exemple, cada ítem ha de mesurar un únic coneixement), al format, a l'estil (recomana ítems curts), a l'enunciat (evitar les negacions) i a les opcions de resposta (recomana evitar l'opció "Totes les anteriors són correctes/incorrectes").

Personalment, preferim els criteris de Moreno, Martínez i Muñiz (2004). Són menys (dotze), són molt més clars i més fàcils d'aplicar. Com podeu veure en la taula1, ara els aspectes que s'han de valorar són tres: l'elecció del contingut, l'expressió i les opcions de resposta.

Taula 1. Noves directrius per a la construcció d'ítems d'elecció múltiple

A. Elecció del contingut que es vol avaluar
1. Ha de ser una mostra representativa del contingut recollit en una taula d'especificació, evitant ítems trivials. 2. La representativitat haurà de marcar el senzill o complex, concret o abstracte, memorístic o de raonament que hagi de ser l'ítem, com també la manera d'expressar-lo.
B. Expressió del contingut en l'ítem
3. El central s'ha d'expressar en l'enunciat. Cada opció és un complement que ha de concordar gramaticalment amb l'enunciat. 4. La sintaxi o estructura gramatical ha de ser correcta. S'han d'evitar ítems massa escarits o profusos, ambigus o confusos, cuidant a més les expressions negatives. 5. La semàntica ha d'estar ajustada al contingut i a les persones avaluades.
C. Construcció de les opcions

Font: pres de Moreno, Martínez i Muñiz (2004)

6. L'opció correcta ha de ser solament una, acompanyada per distractors plausibles.
7. L'opció correcta ha d'estar repartida entre les diferents ubicacions.
8. Les opcions han de ser preferiblement tres.
9. Les opcions s'han de presentar usualment en vertical.
10. El conjunt d'opcions de cada ítem ha d'aparèixer estructurat.
11. Les opcions han de ser autònomes entre elles, sense superposar-se ni referir-se les unes a les altres. Per això, s'han d'evitar les opcions "Totes les anteriors" i "Cap de les anteriors".
12. Cap opció no ha de destacar de la resta ni en contingut ni en aparença.

Font: pres de Moreno, Martínez i Muñiz (2004)

En el contingut, s'han de preguntar coses fonamentals. Sembla obvi, però, quants exàmens recordem en què se'ns preguntaven algunes qüestions que van aparèixer poc (o gens) a classe? Una prova hauria de contenir només (però tots) els conceptes fonamentals de l'assignatura que valora. La creença que en preguntar qüestions menors, en el fons, obliguem l'alumne a estudiar tota la matèria és absurda i afavoreix l'atzar. Respecte a l'atzar, recordem els alumnes que solament estudiaven mig programa –o menys– i confiaven en la sort el dia de l'examen.

Sobre l'expressió, les tres qüestions apuntades són òbvies, però de nou no sempre es compleixen.

Un exemple paradigmàtic de Moreno, Martínez i Muñiz (2004) mostra que és millor redactar aquest ítem:

En física es denomina *sublimació* un canvi de matèria:

1. Sòlida a gasosa
2. Líquida a sòlida
3. Gasosa a líquida

que no aquest:

En física, sublimació:

1. Representa un canvi de matèria sòlida a matèria gasosa.
2. Es refereix a un canvi de matèria líquida a matèria sòlida.
3. Consisteix en un canvi de matèria gasosa a matèria líquida.

Sobre les opcions de resposta, destacarem la recomanació que les opcions siguin independents entre elles, la qual cosa comporta automàticament no usar els cèlebres "Totes/Cap de les anteriors". És obvi que per a rebutjar una opció com "Totes les anteriors són correctes" només necessitem saber que una de les altres opcions no ho és. Així, d'un cop de ploma, podem eliminar dues opcions de les possibilitats i l'elecció es facilita molt. Si el test té tres opcions, ja coneixem la resposta, i si en té quatre, fins i tot ens podem arriscar a contestar a l'atzar entre les dues restants.

Per saber-ne més

Si l'estudiant vol aprofundir en aquest tema, recomanem acudir al text original de Moreno, Martínez i Muñiz (2004), en què, en un to molt didàctic i amb exemples molt accessibles, trobarà una explicació molt exhaustiva de cadascun dels criteris.

Exemple de prova d'execució màxima

A partir de les directrius que hem mostrat, i per il·lustrar amb un exemple concret i proper els conceptes que es presentaran en aquest mòdul, hem construït l'examen següent. Conté deu preguntes sobre aquest mateix mòdul i l'opció correcta està destacada en negreta.

1. La dificultat (ID) és un índex que indica la probabilitat...

A. d'encertar-ho.
B. de fallar-ho.
C. de contestar-ho.

2. El valor de discriminació d'un ítem (ID) ha de ser...

A. negatiu.
B. diferent de 0.
C. positiu.

3. Un distractor hauria de tenir discriminació...

A. positiva.
B. negativa.
C. propera a 0.

4. Un test de personalitat és una prova...

A. d'execució màxima.
B. d'execució típica.
C. de rendiment.

5. La fórmula per a calcular *IDc* és...

A. $\frac{A-\frac{E}{1-K}}{N}$.
B. $\frac{A-\frac{K}{E-1}}{N}$.
C. $\frac{\mathbf{A-\frac{E}{K-1}}}{\mathbf{N}}$.

6. El model de TRI calcula, a partir del coneixement,...

A. la puntuació total esperada.
B. la probabilitat d'encertar un ítem.
C. la discriminació del test.

7. Els paràmetres *a*, *b* i *c* de la TRI indiquen, respectivament,...

A. discriminació, dificultat, pseudoendivinament.
B. dificultat, discriminació, pseudoendivinament.
C. pseudoendivinament, discriminació, dificultat.

8. Si, per nivell de dificultat, solament poguéssim tenir ítems d'un tipus, aquests haurien de ser, generalment...

A. fàcils.
B. difícils.
C. mitjans.

9. Un ítem que pregunti sobre un aspecte del temari difícil hauria de ser...

A. fàcil.
B. mitjà.
C. difícil.

10. L'avaluació del biaix pretén...

A. fer més justes les proves.
B. avaluar la dificultat dels ítems.

Vegeu també

En l'últim apartat d'aquest mòdul es detalla un exemple de respostes (fictícies) d'un grup de vint alumnes a aquesta prova, al costat dels càlculs de la majoria d'índexs als quals farem referència en aquest text.

C. augmentar la fiabilitat de la prova.

3. Teoria clàssica

Hi ha dues maneres principals d'apropar-se a l'anàlisi d'ítems. Distingirem, doncs, entre la teoria clàssica dels tests (TCT) i la teoria de resposta a l'ítem (TRI). La primera l'estudiarem en aquest apartat, i la segona en el següent. Quins supòsits té la TCT? Encara que s'estudia en profunditat en l'apartat de fiabilitat, es resumeixen en l'equació

$$X = V + E$$

Aquesta implica que la puntuació que una persona obté en contestar un instrument de mesura (X) conté l'anomenat *nivell vertader* d'aquesta persona (V) i una part d'error. L'objectiu de la TCT és, doncs, mesurar i minimitzar aquest error, la qual cosa implica analitzar la fiabilitat de la mesura. Com no podia ser d'una altra manera, i sempre en aquests supòsits, tots els indicadors de qualitat dels ítems depenen de la mostra de persones que els han contestat.

Vegem, ara, les propietats principals que s'han de mesurar d'un ítem.

3.1. Dificultat

L'**índex de dificultat** d'un ítem (ID) és la proporció de persones que el contesten correctament. És a dir,

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

Lectura de la fórmula

A: nombre de persones que encerten l'ítem.

N: nombre total de persones que el contesten.

En tractar-se d'una proporció, ja que els encertants són un subconjunt dels que contesten, és obvi que els seus valors fluctuen entre 0 i 1, i sovint s'expressen com un percentatge. Paradoxalment, els valors propers a 1 indiquen una dificultat baixa –s'hauria d'anomenar, doncs, *índex de facilitat* i no *de dificultat*– i valors propers a 0 indiquen dificultat màxima. En la fila titulada ID del full de càlcul annex, podem veure les dificultats dels ítems de l'exemple de prova d'execució màxima.

La fórmula anterior presenta un problema: no té en compte que una part dels encerts es donen per pur atzar. Com es tracta de preguntes amb alternatives tancades, és lògic pensar que una part dels encertants no coneixia la resposta i que si l'ha encertada és solament perquè ha escollit una de les alternatives però no perquè sàpiga la resposta. El problema que se'ns planteja és quants ho han fet? La solució és molt intuïtiva. Si 100 persones que no saben japonès contestessin un examen redactat en aquesta llengua amb preguntes de 4 alternatives, quantes preguntes esperaríem que encertessin? L'esperança matemà-

tica –i el sentit comú– ens diu que 25. I si hi hagués 5 alternatives de resposta? Òbviament, 20. Per tant, el nombre d'encerts total –i per tant, la probabilitat d'encertar una pregunta– depèn en certa manera del nombre d'alternatives de resposta.

Per tant, és recomanable utilitzar la fórmula següent:

$$IDc = \frac{A - \frac{E}{K-1}}{N}$$

Lectura de la fórmula

A: nombre de persones que encerten l'ítem.

E: nombre de persones que fallen l'ítem.

K: nombre d'alternatives (o opcions) de resposta.

N: nombre total de persones que el contesten.

Immediatament, veiem que la diferència entre totes dues fórmules és que en la segona es resta de *A* un nombre que s'obté de dividir els errors (*E*) entre el nombre d'alternatives errònies (*K* – 1).

Per aprehendre aquest concepte observem els resultats de l'ítem 4. L'han encertat 8 dels 20 participants. La seva dificultat sense corregir és, per tant, del 40%. Ara bé, en corregir l'índex per l'encert a l'atzar [(12/[(3 – 1)):20] = 3], obtenim un 30%, això és, podem suposar que el 30% de les persones l'han encertat per casualitat. Val la pena observar que amb el primer càlcul obtindríem una dificultat del 40% (el podríem etiquetar com *de dificultat mitjana*), mentre que amb el segon obtindríem un *IDc* = 10% (dificultat alta).

Encara que conceptualment no té sentit, ja que continua essent una proporció, l'*IDc* pot ser inferior a 0: en aquest cas, s'assigna *IDc* = 0. Això és el que ocorre amb l'ítem 10, que té una dificultat corregida de –5%, cosa que no té sentit. Observem també que els ítems 1 i 3 són perfectament inútils, el primer per fàcil i el segon per difícil.

Un ítem que tots encerten o que tots fallen no serveix per a res més que per a perdre el temps contestant-lo. Si tothom qui respon l'encerta, és com si regaléssim a tots els alumnes una part de la puntuació. I si tots el fallen, és com si els penalitzéssim. Suposem una prova que té 10 ítems i una puntuació teòrica entre 0 i 10. En el primer cas que hem exposat, la puntuació real podria fluctuar entre 1 i 10, i en segon entre 0 i 9. És clar que això no parla bé de les propietats de la prova.

Una vegada sabem la dificultat d'un ítem, plantegem-nos el següent: com haurien de ser les dificultats de tots els ítems d'una prova? Com diu la directriu dos de Moreno, Martínez i Muñiz (2004), la dificultat d'un ítem s'ha de relacionar amb la del concepte que recull. Això és, si un contingut és fàcil, l'ítem ha de ser fàcil. Per tant, una prova que mesura continguts diversos hauria de tenir ítems de totes les dificultats, i aquestes s'haurien de correspondre amb la dificultat dels conceptes mesurats.

Una proposta, feta per nosaltres (Bonillo, 2012), és mostrar en un gràfic la dificultat (eix *Y*) dels ítems d'una prova (ordenats de menys a més dificultat en l'eix *X*) i observar si el pendent és de prop de 45º. Això és, la línia que uneix els punts d'aquestes dificultats hauria de creuar el gràfic en diagonal.

La figura següent mostra aquesta idea aplicada a ítems dels exàmens d'accés a la formació sanitària especialitzada (les cèlebres proves de metge, farmacèutic i infermer intern resident, MIR, FIR i IIR, respectivament) per a dos anys, el 2005 i el 2006. Si sumem totes dues convocatòries, s'hi van presentar més de 23.000 aspirants i els seus resultats són molt rellevants, ja que els aspirants que les superin seran els futurs metges, farmacèutics i infermers especialistes.

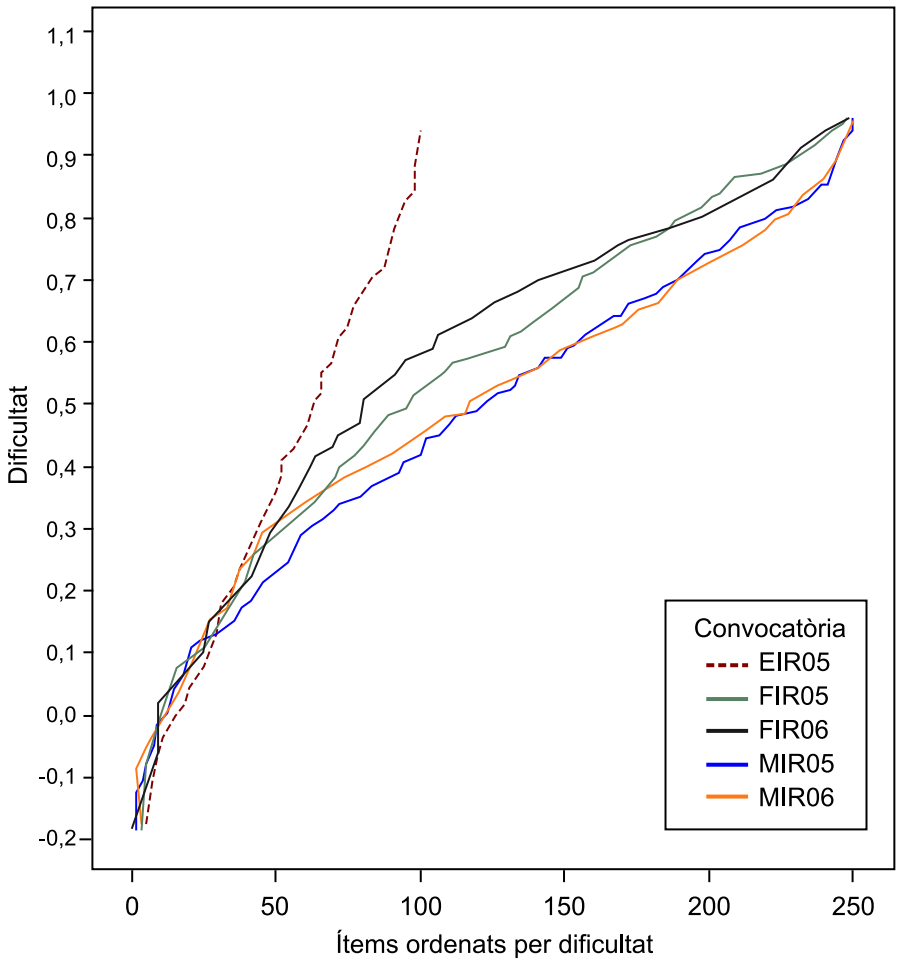

Figura 1. Ítems i dificultat de les proves d'FSE

Podem veure en el gràfic diverses qüestions que val la pena destacar. En primer lloc, veiem que hi ha dificultats negatives, i que aquestes s'expliquen perquè s'ha aplicat la correcció de l'atzar a ítems molt difícils. En segon lloc, cal tenir en compte que la prova d'infermeria consta de 100 ítems (enfront dels 250 de la resta de convocatòries), cosa que explica la diferència evident en les pendents. En tercer lloc, les corbes de les proves de farmàcia estan, en el gràfic, més altes que les de medicina, la qual cosa indica que són proves més fàcils. En quart lloc, les corbes són molt semblants intraprogrames, és a dir, la dificultat de les proves és molt semblant any rere any.

Per saber-ne més

Si l'estudiant vol conèixer amb més profunditat la proposta, que aquí solament apuntem, el remetem a l'article original: Bonillo (2012).

3.2. Discriminació

N'hi ha prou de saber si un ítem és fàcil o difícil per a decidir si és adequat o no? Intuïtivament, podríem pensar que sí, però estaríem equivocats. De fet, si haguéssim de destacar una propietat psicomètrica dels ítems sobre la resta, aquesta seria la discriminació. Si un ítem no discrimina, no és útil per al mesurament, i aquest és l'objectiu per al qual es va redactar.

Com el seu nom indica, entenem per *discriminació* la capacitat d'un ítem per a distingir entre les persones que tenen un bon rendiment en el test i els que tenen un mal rendiment.

Qui ha de contestar correctament una pregunta d'examen? No és tan important si són molts o pocs alumnes, com si els encertants són, en general, dels bons alumnes. A què ens referim quan diem "els bons"? Als alumnes que tenen una puntuació alta en la prova. És a dir, un ítem l'han d'encertar més els que han obtingut una puntuació alta en la prova que els que no l'han obtinguda. Òbviament, una pregunta no pot ser bona si solament l'encerten els pitjors alumnes: ha de passar el contrari.

L'índex de discriminació més popular és l'índex D, també anomenat *índex basat en les proporcions d'encerts*.

$$D = Pa - Pb$$

Lectura de la fórmula

P_a: proporció de persones del grup d'alt rendiment que encerta l'ítem.

P_b: proporció de persones del grup de baix rendiment que encerta l'ítem.

Les proporcions es calcularien com hem vist en la primera fórmula presentada, i, de nou, es poden expressar com a percentatges. Però, quins són els alumnes d'alt i baix rendiment? Hi ha diverses maneres de definir el punt de tall de la puntuació total en la prova per a fer aquesta classificació. D'una banda, és freqüent utilitzar la mitjana de la puntuació total en la prova, cosa que crea dos grups igual de grans. Aquesta estratègia té com a avantatge que tots els participants participen en el càlcul, però té com a inconvenient clar que els grups són poc extrems. Intuïtivament veiem que dues persones amb un rendiment molt semblant poden estar en grups diferents solament per una petita diferència.

És preferible utilitzar grups més extrems per a poder estudiar correctament aquest índex. Kelley (1939) recomana utilitzar els percentils superior i inferior del 27%. Per què el 27% i no el 25%? Encara que l'article original demostra que el 27% és lleugerament millor que el 25%, en l'exemple amb respostes fictícies que mostrem s'utilitza el 25% com a criteri per a separar el grup de rendiment alt –s'interpretaria com el que obté puntuacions superiors al 75%

dels seus homòlegs– del baix –grup que reuneix el 25% de les puntuacions més baixes. Calcular el percentil 27 no sempre és senzill, i el 25 sí, i les variacions entre l'un i l'altre són molt més petites.

Quins són els límits del D? És obvi que, teòricament, poden fluctuar entre 1 i –1. El primer valor es donaria solament quan totes les persones del grup superior encertessin i totes les de l'inferior fallessin. El valor –1 solament es podria donar en el cas contrari, i llavors hauríem de dubtar sobre si la resposta considerada correcta ho és. Cap d'aquestes dues situacions se sol donar en la realitat.

Com hem d'interpretar, doncs, aquest índex? En primer lloc, només valors positius indiquen discriminació. És clar que un ítem ha de ser més encertat entre els millors. Però, quins valors indiquen una bona discriminació? Ebel (1965) va proposar la classificació següent, que s'ha de prendre com una orientació:

Taula 2. Punts de tall dels valors (en %) de discriminació (D) i la seva interpretació

D	Interpretació de la discriminació
>40	Discriminació alta.
30-40	Acceptable.
20-30	Baixa: se suggereix revisar l'ítem.
0-20	Dolenta: s'elimina l'ítem o es reforma profundament.
<20	Inacceptable: cal eliminar l'ítem.

Un motiu per a prendre la taula anterior amb precaució és que l'índex D depèn –i molt– de la dificultat. Si un ítem és molt difícil, tindrà pocs encertants (per definició) fins i tot en el grup d'alt rendiment. Si P_a és baixa, la D solament pot ser baixa. No sembla just comparar D d'ítems de dificultats molt diferents. Una alternativa proposada a l'índex D és calcular la diferència de proporcions relativa, en lloc de l'absoluta. És a dir,

$$Dr = (Pa - Pb) \big/ P$$

Calculem la discriminació dels ítems 4 i 5, que tenen la mateixa dificultat.

Ítem 4: $D = 4/5 - 0/0 = 80\%$ i $Dr = \frac{0,8 - 0}{0,8} = 100\%$

Ítem 5: $D = 2/5 - 0/0 = 40\%$ i $Dr = \frac{0,4 - 0}{0,4} = 100\%$

Per a l'ítem 4, el primer valor (D = 80%) s'hauria d'interpretar de la manera següent: els millors encerten l'ítem un 80% més que els pitjors. O interpretar-lo com una diferència de probabilitats: és un 80% més probable encertar l'ítem quan es té un rendiment alt (que si es té baix). El segon valor (Dr = 100%) s'interpretaria com que els bons tenen un 100% més d'encerts (respecte als dolents). En aquest cas, tots dos valors parlen bé de la capacitat discriminativa de l'ítem. Per a l'ítem 5, les dades són semblants (D = 40% i Dr = 100%). Així doncs, tots dos ítems serien més que acceptables.

Suposem ara un ítem molt difícil, que solament l'encerti 1 de 20 participants, i suposem que aquest pertany al grup d'alt rendiment. És clar que $D = 1/6 - 0/6 = 0,17 = 17\%$. Segons

els criteris d'Ebel l'hauríem d'eliminar, però *Dr* = 100%, i per tant té discriminació relativa màxima. Llavors, què hauríem de fer? No hi ha una resposta absoluta a aquesta pregunta.

Si, per les característiques de la prova, és acceptable tenir un ítem tan difícil, aquest s'hauria de mantenir, ja que discrimina tant com pot discriminar. Si, per contra, no és convenient que tan poques persones l'encertin, s'hauria de reformar, però la decisió –en aquest cas– depèn per complet de la dificultat i no de la discriminació.

En resum, cal tenir clar que la discriminació depèn de la dificultat i no s'ha d'interpretar *per se*. En termes estadístics, la discriminació depèn de la variància de la dificultat.

Hi ha altres índexs alternatius als que s'han presentat per a mesurar la discriminació. Un dels més utilitzats és la correlació ítem-test. Habitualment, s'utilitza l'índex de correlació biserial-puntual, ja que permet quantificar la relació entre una variable binària –encertar l'ítem o no– i una variable d'escala –la puntuació total de la persona en la prova, idealment sense tenir en compte l'ítem analitzat. La fórmula i la lògica d'aquesta prova es troba explicada àmpliament tant a Muñiz (2003)[1] com a Martínez Arias (1992)[2], però és molt senzill veure que una alta correlació –propera a 1– indica una gran discriminació de l'ítem, que valors propers a –1 indiquen el contrari (que els bons fallen l'ítem i els dolents l'encerten) i que valors propers a 0 indiquen que encertar aquest ítem no té res a veure amb el coneixement que mesura el conjunt de la prova.

(1) Muñiz (2003, p. 220).

(2) Martínez Arias (1992, p. 556).

3.3. Discriminació dels distractors

Un aspecte clau per al bon funcionament d'un ítem és que els seus distractors[3] realment ho siguin. Una alternativa que ningú –o gairebé ningú– no tria no confon les persones que responen i, per tant, no és útil. I al revés: una alternativa incorrecta que els millors trien molt potser no és tan incorrecta com va pensar qui la va redactar.

(3) *Distractors* és el nom que es dóna a les alternatives de resposta incorrectes.

Com s'estudia el comportament dels distractors? L'índex habitual torna a ser l'índex D que ja hem vist, però en lloc de calcular-lo a partir dels qui encerten i els qui fallen, es fa a partir dels qui trien cadascuna de les alternatives de respostes.

Observem atentament els resultats d'administrar l'ítem 5. La discriminació de l'opció de resposta B és més alta que la de la resposta correcta, la C. Això s'ha d'interpretar com que els millors trien més un distractor –com és la C– que la resposta correcta –com és la B. Això implica que hem de revisar si, potser, la pauta conté un error. En el nostre cas no és així, i podem atribuir a l'atzar que l'opció C hagi estat tan escollida. Ara bé, potser hauríem de recalcar als alumnes què significa cada element de la fórmula (ja que és el que es pregunta en l'ítem 5) i enfortir així l'aprenentatge.

Quines propietats matemàtiques ha de tenir un distractor? Òbviament, tenir discriminació negativa, és a dir, ser més escollit entre els pitjors que entre els millors. A més, seria òptim que tots els distractors tinguessin una discriminació

semblant, ja que indicaria que les seves capacitats d'atracció són semblants. Aconseguir això és especialment difícil, i aquesta dificultat creix exponencialment amb les alternatives de resposta. En resum: és molt més difícil redactar tres distractors que siguin efectius, que dos. Per això, la majoria d'estudis recomanen usar com a molt tres alternatives de resposta.

No hem de creure que les propietats que fins ara hem presentat són independents entre elles: res més allunyat de la realitat. Si un ítem té una opció de resposta inversemblant (per exemple, Maradona com a autor d'*El Quixot*), l'ítem serà més fàcil i necessàriament discriminarà pitjor.

Com ja hem fet quan hem estudiat la dificultat, ara ens plantegem si hi ha alguna manera d'estudiar el conjunt de les discriminacions dels ítems d'una prova.

De nou, la proposta és nostra (Bonillo, 2012) i consisteix a mostrar en un gràfic, anomenat *diagrama de caixa*, la discriminació de l'opció correcta i de cadascun dels distractors, ordenats de més grans a més petits.

La figura següent mostra això aplicat, de nou, als exàmens d'accés a la formació sanitària especialitzada.

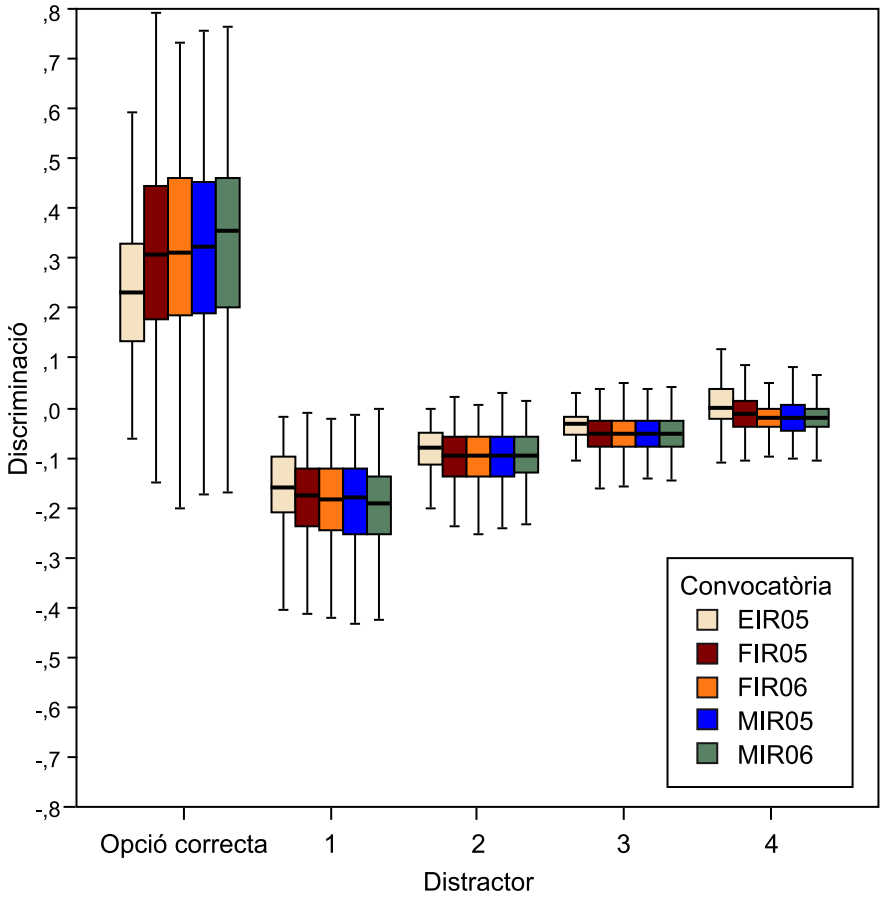

Figura 2. Discriminació dels distractors de les proves d'FSE

Cal tenir en compte que en aquesta figura apareixen les cinc convocatòries analitzades, això és [250 ítems × 2 programes × 2 anys + 100 ítems d'IIR] × 5 alternatives = 5.500 valors. Així, i per a cada ítem, el distractor 1 és el més discriminatiu, i el 4 el menys. Com sol passar en aquests gràfics, les caixes mostren la mitjana –en traç gruixut– i els quarts –en els límits de les caixes. Les patilles (*whiskers*) mostren els valors mínims i màxims no allunyats ni extrems. Els allunyats es mostren amb punts i els extrems amb asteriscos.

S'observa que les discriminacions de les alternatives correctes són semblants entre especialitats i convocatòries. Destaquen de les altres les discriminacions relatives a la prova d'IIR, que són més baixes i menys disperses. En l'anàlisi dels distractors es veu que hi ha un escalat entre aquests, però que es redueix com més alternatives es tenen en compte; és a dir, la diferència entre la tercera i la quarta alternativa és molt més petita que entre la primera i la segona. També s'observa que les alternatives tres i quatre –recordem que són ordenades per la seva discriminació i que no s'han d'identificar amb alternatives de resposta D i E, per exemple– tenen discriminacions molt baixes o gairebé nul·les. Si considerem que el límit superior de les caixes de l'última alternativa és superior a 0, podem dir que més del 25% dels ítems té una alternativa de resposta amb discriminació positiva –és a dir, més escollida pel grup amb rendiment alt. A més, l'última alternativa presenta molts valors extrems i allunyats, això és, ítems en els quals, per la seva alta discriminació positiva, seria discutible si l'opció donada com a correcta veritablement ho és, o és l'única que ho és. Conclusió que n'hem de treure: amb tres opcions seria més que suficient.

Per saber-ne més

De nou, si l'estudiant vol conèixer a fons aquesta proposta ha de consultar l'article original (Bonillo, 2012), ja que en aquest mòdul no ens podem estendre molt més del que ja hem fet.

3.4. Valoració del biaix

Un aspecte crucial quan es crea –i quan es valora– tant un ítem com un instrument de mesura és que aquests no siguin esbiaixats. De què parlem quan ens referim al biaix en un instrument de mesura? Una bàscula estarà esbiaixada si sempre infravalora el pes d'un objecte enfront d'un altre quan sabem que tots dos pesen exactament el mateix. En el context de les proves d'execució màxima, entenem que un ítem –o un test– està esbiaixat quan grups, per exemple homes i dones o rics i pobres, que tenen el mateix coneixement sobre la matèria mesurada no obtenen valors iguals, sinó que un dels grups és sistemàticament "perjudicat".

Com us podeu imaginar, els instruments esbiaixats poden tenir greus implicacions socials. Si un examen, com la selectivitat, afavorís sistemàticament un alumne amb un nivell socioeconòmic alt enfront d'un que no té aquest nivell, sempre que sabéssim que tots dos saben exactament el mateix sobre el tema, la selectivitat no seria socialment justa.

Com es pot valorar el biaix? Actualment, s'utilitza el concepte de **funcionament diferencial dels ítems** i l'índex principal que se'n deriva és el DIF[4]. Diem que un ítem té o presenta DIF[5] quan es donen diferències estadísticament

(4)La sigla DIF prové de les inicials del terme en anglès.

significatives en la puntuació d'un ítem en dos grups diferents que haurien de tenir, per lògica, el mateix nivell. Per a avaluar el DIF es poden utilitzar diferents procediments matemàtics[6], però en aquest text solament parlarem del mètode de Mantel-Haenszel (1959). Aquest és, en el marc de la TCT, el més utilitzat perquè és senzill de calcular i pels bons resultats que dóna.

La manera de calcular i la idea que hi ha darrere el DIF és molt senzilla. Suposem que volem saber si la nostra prova no està afectada pel sexe dels que responen. És obvi que no ho hauria d'estar i que si no és així l'hauríem de millorar, ja que es podria qualificar de *sexista*.

Es tracta llavors de dividir els subjectes en grups en funció de les puntuacions totals (per exemple, en cinc grups). Després, calculem una taula per grup en què observem si la variable *sexe* s'associa a encertar més. Finalment, aquest resultat s'agrega a l'estadístic de Mantel-Haenszel i es compara amb el χ^2 de referència. Si el resultat és significatiu, vol dir que hi ha trams de puntuació total en què un sexe sembla que té avantatge sobre un altre, ja que té més encerts. Llavors, podem dir que la prova no és justa. Quan creem una prova hauríem de comprovar que és independent de variables com el gènere, la raça i qualsevol altra que pugui comportar, implícitament, la discriminació de les persones.

(5)En terminologia psicomètrica col·loquial, diem que hi ha DIF.

(6)A Muñiz (2003, pp. 239-253) se'n pot veure una excel·lent revisió.

Lectura recomanada

Recomanem encaridament la consulta de l'exemple que presenta Muñiz (pp. 247-249) i que aquí descriurem però no desenvoluparem matemàticament.

4. Teoria de resposta a l'ítem

La **teoria de resposta a l'ítem**[7] parteix d'una perspectiva totalment diferent. Critica la TCT en afirmar que estudia les propietats d'un test particular en una mostra –també particular– de persones. Des de la TCT, és cert que dos tests diferents però que mesuren el mateix constructe tindran propietats diferents. I també és cert que aquestes propietats dependran de si les persones utilitzades per a calibrar el test de rendiment tenen un rendiment alt o no. La TRI permet superar aquests problemes, però a costa de complicar enormement els càlculs i ser més difícil d'utilitzar en ser més "caixanegrista".

(7)En anglès, *item response theory* (IRT).

La TRI es basa en el càlcul, per a cada ítem, d'una sèrie de paràmetres, que assumeixen un model matemàtic molt concret. L'objectiu últim és la mesura del tret latent[8] a partir de tres paràmetres:

(8)*Tret latent* és el nom que rep el constructe que s'ha de mesurar, per exemple el coneixement d'una assignatura.

- la discriminació de l'ítem,
- la dificultat i
- l'encert a l'atzar.

(9)En anglès, *item response function* (IRF).

Aquests tres paràmetres es poden veure en la figura següent, que resulta clau per a entendre què és la TRI[9]: es coneix com la corba característica de l'ítem (CCI).

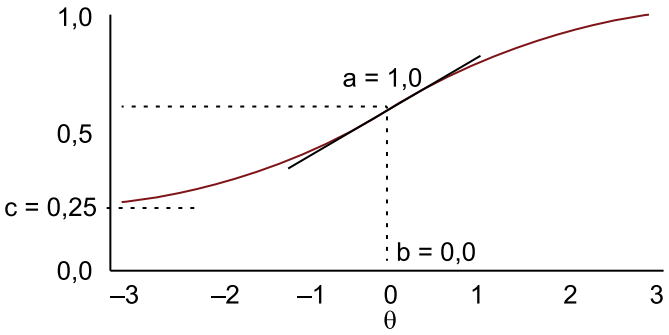

Figura 3. Exemple de corba característica d'un ítem (CCI)

La CCI mostra, en l'eix de les ordenades (l'eix *Y*), la probabilitat d'encertar l'ítem a partir de la magnitud del tret latent (eix d'abscisses o *X*). Aquesta probabilitat segueix una funció sigmoïdal (en forma de *S*, també anomenada *logística*) i el tret latent s'indica com a θ (theta). La funció logística té com a propietat que no pot ser més petit que 0 ni més gran que 1, i θ sol estar compresa entre –3 i +3.

A partir de la CCI podem mesurar els tres paràmetres clau de la TRI. El paràmetre indicat amb *a)* mesura la discriminació de l'ítem. Una corba molt plana expressaria que no és important tenir un alt coneixement del tret per a

augmentar la probabilitat d'encert. És a dir, com més gran sigui el pendent, més gran serà la discriminació. El del gràfic que es mostra és 1, valor positiu i acceptable.

El paràmetre indicat com a *b* mesura la dificultat a partir del punt de tall de l'eix *X*, que correspon a una probabilitat d'encert del 50%. S'interpretaria com el nivell de tret latent necessari per a tenir un 50% de probabilitat d'encertar l'ítem. Com més gran sigui la *b*, més difícil serà l'ítem, ja que caldrà més coneixement per a poder arribar a aquest 50% de probabilitat desitjat. El valor que es mostra en el gràfic és 0, que s'interpretaria com que és un ítem de dificultat (exactament) mitjana. El tercer paràmetre, indicat com a *c*, mesura el nivell d'atzar i també es coneix com a *índex de pseudoendevinament*. Gràficament, correspon al valor de *X* que talla l'eix *Y*. Recull, lògicament, la probabilitat d'encertar quan el coneixement de l'ítem és nul. El valor del gràfic indica que aquest és alt: del 25%.

Aquests càlculs es fan amb programari molt específic. Mostrar amb quines eines es fa i com, excedeix els objectius d'aquest mòdul.

Enllaç recomanat

La pàgina següent conté un conjunt de programes gratuïts que permeten aplicar la TRI: http://www.psychology.gatech.edu/unfolding/FreeSoftware.html. Ara bé, hem de recordar a l'estudiant que aplicar la TRI no és senzill i que requereix una formació més completa que la que necessita l'aplicació de la TCT.

Exemple d'avaluació de les propietats dels ítems a una mostra d'alumnes fictícia

En la taula següent es mostren les respostes dels 20 subjectes (per files) a les deu preguntes de la prova (columnes). En negreta es marquen les respostes correctes, i la seva pauta apareix en la primera fila.

Pauta	**A**	**C**	**B**	**B**	**C**	**B**	**A**	**C**	**C**	**A**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**
S1	**A**	**C**	C	**B**	**C**	**B**	**A**	**C**	**C**	B
S2	**A**	**C**	C	**B**	B	**B**	**A**	**C**	**C**	A
S3	**A**	**C**	C	**B**	B	**B**	**A**	**C**	**C**	**A**
S4	**A**	**C**	C	A	**C**	**B**	**A**	**C**	**C**	**A**
S5	**A**	**C**	A	**B**	B	**B**	**A**	**C**	**C**	**A**
S6	**A**	A	A	A	A	A	B	A	B	B
S7	**A**	**C**	A	**B**	**C**	A	B	B	B	B
S8	**A**	**C**	A	A	V	A	B	B	B	B
S9	**A**	A	A	**B**	**C**	A	B	B	B	C
S10	**A**	**C**	A	**B**	A	A	B	B	B	B

Pauta	**A**	**C**	**B**	**B**	**C**	**B**	**A**	**C**	**C**	**A**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**
S11	**A**	A	A	**B**	**C**	**B**	B	**C**	**C**	**A**
S12	**A**	A	A	A	A	**B**	B	**C**	**C**	**A**
S13	**A**	A	A	A	**C**	**B**	B	B	B	C
S14	**A**	A	A	A	**C**	**B**	B	B	B	B
S15	**A**	A	A	A	**C**	**B**	B	B	B	C
S16	**A**	A	A	A	B	C	B	B	**C**	B
S17	**A**	A	A	A	B	A	B	B	B	B
S18	**A**	A	A	A	A	C	B	A	A	C
S19	**A**	A	A	A	A	A	B	A	A	C
S20	**A**	A	A	A	A	A	B	A	A	C

La taula reprodueix l'anterior, però codificant la resposta com a 1 si s'ha encertat, i com a 0 si no s'ha encertat. La columna titulada "P. total" conté la puntuació total de cada persona, i és el resultat de sumar els uns de la taula. En la part inferior veiem dues files, titulades "ID" i "IDc". Com sabem, la primera correspon a l'índex de dificultat i la segona al mateix però en la versió corregida. Observeu la gran diferència entre tots dos i recordeu que el segon corregeix l'encert per atzar que el primer obvia.

	Encert										**P. total**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**	
S1	1	1	0	1	1	1	1	1	1	0	8
S2	1	1	0	1	0	1	1	1	1	1	8
S3	1	1	0	1	0	1	1	1	1	1	8
S4	1	1	0	0	1	1	1	1	1	1	8
S5	1	1	0	1	0	1	1	1	1	1	8
S6	1	0	0	0	0	0	0	0	0	0	1
S7	1	1	0	1	1	0	0	0	0	0	4
S8	1	1	0	0	0	0	0	0	0	0	2
S9	1	0	0	1	1	0	0	0	0	0	3
S10	1	1	0	1	0	0	0	0	0	0	3
S11	1	0	0	1	1	1	0	1	1	1	7
S12	1	0	0	0	0	1	0	1	1	1	5
S13	1	0	0	0	1	1	0	0	0	0	3
S14	1	0	0	0	1	1	0	0	0	0	3

	Encert										P. total
S15	1	0	0	0	1	1	0	0	0	0	3
S16	1	0	0	0	0	0	0	0	1	0	2
S17	1	0	0	0	0	0	0	0	0	0	1
S18	1	0	0	0	0	0	0	0	0	0	1
S19	1	0	0	0	0	0	0	0	0	0	1
S20	1	0	0	0	0	0	0	0	0	0	1
ID	100,0%	40,0%	0,0%	40,0%	40,0%	50,0%	25,0%	35,0%	40,0%	30,0%	
IDc	100,0%	10,0%	–50,0%	10,0%	10,0%	25,0%	–12,5%	2,5%	10,0%	–5,0%	

Els punts de tall dels quartils de la puntuació total són, respectivament, 1,75 i 8. Això és, quina puntuació mínima cal tenir per a estar per sobre del primer i del tercer quartil. Els valors compresos entre tots dos corresponen al grup anomenat *intermedi*, i que conté el 50% de les persones amb valors entorn de la mitjana. A partir d'aquests valors es calcula la columna que podem veure a la dreta de la taula, i que classifica les persones en tres grups.

	P. total	Grup rend.
S1	8	Superior
S2	8	Superior
S3	8	Superior
S4	8	Superior
S5	8	Superior
S6	1	Inferior
S7	4	Intermedi
S8	2	Intermedi
S9	3	Intermedi
S10	3	Intermedi
S11	7	Intermedi
S12	5	Intermedi
S13	3	Intermedi
S14	3	Intermedi
S15	3	Intermedi
S16	2	Intermedi
S17	1	Inferior
S18	1	Inferior

	P. total	Grup rend.
S19	1	Inferior
S20	1	Inferior

Les cel·les de la taula següent mostren els càlculs relatius a la discriminació de cadascun dels ítems. En les cel·les titulades "Encert del grup superior", es mostra el percentatge de persones del grup de rendiment superior que s'ha escollit, respectivament, les alternatives *A*, *B* i *C*. Un càlcul idèntic es fa en les cel·les titulades "Encert del grup inferior". Què fem després amb aquests valors?

Gràcies a aquest càlcul, podem obtenir de manera senzilla la discriminació de cadascuna de les opcions de resposta. En les cel·les titulades "Índex D alternatives" veiem la discriminació de l'alternativa *A*, i ídem per a les *B* i *C* amb les files inferiors. En negreta es mostra la discriminació de l'alternativa correcta. Observeu que la negreta coincideix amb la lletra que, en la primera taula d'aquest apartat, apareixia en la pauta de respostes correctes.

	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
% encert grup superior										
A	100,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%
% encert grup inferior										
A	100,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
B	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
C	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Índex D alternatives										
A	0,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%

Resum

Al llarg d'aquest text, hem exposat una multitud d'aspectes sobre els instruments de mesura. Sobre una prova hem vist què cal estudiar, com i quan s'ha de fer i hem insistit –especialment i conscientment– en el perquè cal fer-ho. La idea principal que ens agradaria haver transmès és que el que no podem fer és no fer res. És a dir, la pitjor de les situacions que podem imaginar –en aquest context, és clar– és aplicar una prova sense estudiar-ne cap de les propietats. No fer-ho convertirà l'error en norma i no en excepció.

Aquesta situació descrita és, lamentablement, la realitat. En l'àmbit universitari, que és el que ens és més proper, són pocs els professors que estudien els exàmens després d'administrar-los. Fins i tot els que ho fan, rares vegades publiquen els resultats, que permetrien als alumnes contrastar la justícia de la prova amb la qual se'ls va examinar. Per què passa això? Creiem que és més atribuïble a la falta de formació que no a la falta de transparència. El professor no és format en la manera com ha de valorar les seves proves, i tampoc no és format en la manera com ha de fer una classe. En el nostre sistema educatiu, i ens referim a totes les etapes, la cultura de treballar amb evidències no està implantada.

No podem no centrar-nos en l'àmbit educatiu, que és al qual pertanyem, però, no penseu que s'haurien de publicar les dades que permetin valorar la justícia d'unes oposicions, com són les proves d'accés de la formació sanitària especialitzada? De fet, en l'article que ja hem esmentat (Bonillo, 2012) aquesta va ser la nostra proposta principal. Així, els opositors podrien impugnar preguntes, proposar altres correccions i, en definitiva, tots podríem estar més tranquils en saber que les proves són justes i premien de debò els millors.

En aquest mòdul s'han presentat molts conceptes nous i s'ha fet de manera molt breu. Considerem que l'estudiant està preparat per a aplicar-los demà? En absolut. Ens agradaria pensar que, per a qui un dia necessiti crear o valorar una prova, aquest text és el primer d'altres. Aquests altres estan citats o se'n poden buscar alternatives.

Per a l'estudiant que no necessiti el que aquí s'exposa, que serà la gran majoria, confiem que els conceptes exposats s'entenguin. Quan això ocorre, i en el futur es necessita aplicar-los, recuperar-los de la memòria i de la biblioteca és senzill.

En qualsevol cas, i per a tots, esperem haver insuflat esperit crític i desig de treballar amb evidències i per les evidències. Al final, aquests dos elements són els que separen la psicologia d'altres coses que ens han de ser alienes.

Bibliografia

Baker, Frank (2001). *The basics of item response theory. ERIC Clearinghouse on Assessment and Evaluation*. Maryland: University of Maryland.

Bonillo, A. (2012). Pruebas de acceso a la formación sanitaria especializada para médicos y otros profesionales sanitarios en España: examinando el examen y los examinados. *Gaceta Sanitaria*, 26 (3), 231-235.

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

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood: Prentice Hall.

Haladyna, T. M., Downing, S. M., i Rodríguez, M. C. (2002). A review of multiple-choice item-writing guidelines for Classroom Assessment. *Applied Measurement in Education*, 15 (3), 309-334.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30 (1), 17-24. DOI: 10.1037/h0057123.

Mantel, N. i Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Martínez Arias, R. (1996). *Psicometría: Teoría de los Tests Psicológicos y educativos*. Madrid: Síntesis.

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

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.