39

Quaderns de Docència Universitària

Bones pràctiques en l'ús de proves d'alternativa múltiple

Jordi Renom Pinsach
Eduardo Doval Diéguez

Universitat de Barcelona

Quaderns de docència universitària 39

Títol: *Bones pràctiques en l'ús de proves d'alternativa múltiple*

Primera edició: octubre de 2019

Recepció de l'original: 11/03/2019
Acceptació: 09/05/2019

Ediciones OCTAEDRO
Bailèn, 5, pral. - 08010 Barcelona
Tel.: 93 246 40 02 - Fax: 93 231 18 68
www.octaedro.com - octaedro@octaedro.com

Universitat de Barcelona
Institut de Desenvolupament Professional/ICE
Campus Mundet - 08035 Barcelona
Tel.: 93 403 51 75 - Fax: 93 402 10 61

ISBN: 978-84-17667-83-2

Disseny i producció: Serveis Gràfics Octaedro

ÍNDEX

INTRODUCCIÓ

Des de fa dècades els test «objectius» i exàmens «tipus test» han estat objecte d'un debat controvertit. Semblava que els nous models educatius i les noves tecnologies farien que aquestes maneres d'avaluar quedarien obsoletes enfront altres formes d'avaluació més autèntiques. Tanmateix, avui dia segueixen sent molt presents en l'avaluació universitària, en els processos d'acreditació i en la formació en general, i molt habituals en les plataformes formatives digitals.

Las proves «tipus test» o d'elecció múltiple tenen, com qualsevol altra classe de prova, avantatges i inconvenients tot i que en aquest cas els usuaris tendeixen a ponderar només les primeres. Si ens fixem en el volum d'avaluacions on s'apliquen, i les repercussions dels seus resultats per els examinats sorprèn la manca de normativa de qualitat d'aquest tipus d'exàmens. Existeixen regulacions per evitar el frau per part dels examinats i també directrius internacionals de bones pràctiques sobre la creació i administració d'exàmens, tot i que aquestes són poc conegudes i, en qualsevol cas, al tractar-se de recomanacions, la seva aplicació queda a criteri de les persones que vagin a fer ús de la prova. Tampoc és habitual que els i les docents sotmetin les seves proves a control, efectuant o sol·licitant una anàlisi de qualitat. De fet, d'acord amb el tòpic predominant s'entén que els exàmens «tipus test» són fàcils de crear i que de per sí aporten un valor d'objectivitat a l'avaluació, fent que aquest control sembli innecessari. Tanmateix l'experiència acumulada, mostra que el principal valor d'aquests proves es redueix a l'autonomia (que no facilitat) de creació i, sobre de tot, a la facilitat de correcció. Quant a la suposada objectivitat dels seus resultats no queda, ni de lluny, garantida.

Igual que en altres procediments d'avaluació, els test «objectius» comporten unes regles de joc que cal conèixer, cas contrari l'usuari assumeix una responsabilitat que pot semblar que no l'afecta directament però que sempre acaba desafavorint als examinats.

Mantenint un enfoc crític, aquesta obra va dirigida a les persones encarregades de realitzar avaluacions amb proves d'aquesta modalitat, ja

que són les que han de prendre consciència de les prestacions i limitacions dels instruments.

Per aquest motiu no es tracta d'un manual especialitzat ni exhaustiu sobre mesura en educació, la bibliografia recull algunes de les principals referències d'aquest camp i pot ajudar a aprofundir en aspectes concrets. Tampoc es pretén debatre els avantatges ni els inconvenients ni l'estat de la qüestió d'aquesta modalitat d'avaluació. A partir de l'experiència dels autors, es proposa una metodologia de base per a docents de diferent perfil interessats en aquests instruments d'avaluació que els permeti abordar de manera autònoma i amb garanties com fer *test del test* d'un examen.

1. OBJECTIU DEL TEST

En el sentit més convencional un test és un instrument format per una sèrie d'ítems (preguntes, tasques, problemes...) dissenyats per a avaluar coneixements o aptituds en l'àmbit psicològic i educatiu. En el cas dels test «tancats» d'alternativa múltiple (AM) la persona que s'examina escull la resposta correcta entre diverses opcions que cada pregunta ofereix. Els ítems s'encerten, fallen o ometen i, generalment, la suma d'encerts (puntuació total) proporciona una suposada mesura del nivell de l'examinat.

La finalitat d'aquests exàmens sol anar lligada a l'avaluació dels aprenentatges, bàsicament coneixements i competències en tota mena de continguts i àmbits. També són habituals en processos de selecció i d'acreditació (oposicions, idiomes, conducció,...).

En termes pedagògics els exàmens s'han relacionat més amb l'avaluació sumativa i menys amb la formativa. Això, en part, es deu a la forma en que s'han emprat dins l'estratègia d'avaluació.

Quant a repercussions, els exàmens varien en funció dels seus efectes. Hi ha proves de simple control i seguiment, com per exemple la modalitat denominada *quiz* (exàmens) freqüents en moltes plataformes virtuals d'aprenentatge o LMS (*learning managment systems*), que serveixen per que els estudiants realitzin exercicis de prova. A l'altre extrem hi ha els exàmens amb efectes vinculants per el futur de l'examinat, quan aprovar o suspendre'l té conseqüències clares (per exemple, repetir l'assignatura, tornar a pagar la matrícula, superar o no una oposició...). Els exàmens també poden diferir en la precisió del seus resultats. N'hi ha de tipus *screening* o exploratoris que informen, tot i que amb poca precisió, del nivell que té cada persona avaluada. Un altre cas són els de certificació que ofereixen puntuacions més precises per tal de minimitzar el risc d'error en les decisions tipus apte/no apte, a partir d'una nota de tall molt concreta.

Des d'un punt de vista semàntic la principal diferència entre test i examen rau en el caire estandarditzat del primer. Aconseguir una prova

estandarditzada requereix d'estudis previs realitzats amb una o més mostres pilot de persones, per tal d'avaluar la qualitat dels ítems, del conjunt de la prova i per concretar-ne una versió definitiva de la prova amb garanties psicomètriques. Generalment, en l'àmbit de l'avaluació, aquest tipus d'anàlisi i de comprovacions prèvies només són possibles en avaluacions fetes a gran escala, com en els estudis PISA. Per contra, en un examen no sol haver-hi assajos ni anàlisi previs de les seves qualitats mètriques amb mostres pilot d'estudiants. Habitualment un examen es crea amb una finalitat concreta, per a una ocasió determinada i els resultats s'interpreten a partir d'un criteri predefinit per a les persones que realitzen l'avaluació. Un examen (el mateix examen) no es dissenya per a diferents ocasions i grups. Altrament, la seva originalitat i confidencialitat sovint són els valors que garanteixen l'equitat de l'avaluació.

Malgrat aquestes diferències, test i exàmens també comparteixen algunes qualitats i elements metodològics importants, encara que poques vegades quedin garantides en els exàmens. Per exemple, després d'aplicar un examen no és habitual analitzar les seves qualitats psicomètriques. Generalment, l'examen simplement es crea, s'aplica i permet avaluar a les persones que es presenten a l'avaluació, assumint que la prova té garanties (no comprovades) per fer-ho. Les qualitats sempre es poden avaluar però malauradament molts cops no es realitza cap comprovació per verificar-les ja que sovint els responsables de l'examen desconeixen la seva existència i importància.

1.1. Altres qüestions clau

Un cop definits els objectius d'avaluació de la prova cal concretar-ne la forma i estructura. Això afecta a quatre preguntes clau: quin format d'ítem s'utilitzarà?, quantes preguntes ha de tenir la prova?, quines taxonomies de coneixement s'aplicaran? i cóm es corregirà i puntuarà?

Quin format d'ítem s'utilitzarà? Te a veure amb la manera com caldrà respondre als ítems. Hi ha exàmens de resposta oberta, on l'examinat elabora la resposta, i tancats del tipus verdader/fals (VF) o AM amb diverses variants. En aquesta obra el format de referència serà el d'AM

convencional amb una sola resposta correcta que puntua. Molts dels aspectes tractats seran també vàlids per el format VF i per altres variants d'AM.

Quantes preguntes ha de tenir la prova? Aquí no hi ha una pauta estricta. Segons el format un examen pot tenir diferent nombre L (longitud) d'ítems. En el cas d'AM també cal decidir el nombre *n* d'alternatives de resposta (es recomanable que tots els ítems mantinguin aquest valor). Existeix una relació matemàtica, descrita a la teoria clàssica dels test (TCT, una disciplina psicomètrica), entre *L i n* i el coeficient de fiabilitat de les puntuacions del test. En general la fiabilitat creix quant més gran és *L* i també *n*, augmentant per tant, també la precisió de la mesura (ja que l'error de mesura de les puntuacions es redueix). Per això moltes proves oficials són llargues, de vegades excessivament (és un recurs tècnic per aconseguir major precisió dels resultats i menys risc de reclamacions, impugnacions, etc.). En l'àmbit aplicat és possible estimar el punt on s'optimitza la relació entre l'esforç que suposa crear i gestionar moltes preguntes i aconseguir una major fiabilitat/precisió.

Quant al valor idoni de *n* des de fa dècades ha estat objecte d'estudi i discussió. La tendència històrica, des d'un enfoc tècnic, ha anat a la baixa estabilitzant-se avui dia entrequatre i sis alternatives (per a maximitzar la fiabilitat) o tres des d'un punt de vista pràctic (per aconseguir bones opcions de resposta). Paradoxalment, sovint el nombre d'alternatives de la prova es decideix de manera purament circumstancial, per exemple, en funció del model de full de resposta disponible, o de les prestacions de la lectora òptica, o per costum («en aquesta assignatura l'examen sempre s'ha fet així»).

A l'hora de decidir *L* i *n* també cal considerar l'estratègia com s'abordarà l'efecte de la conjectura a les respostes (apartats 7.2, 7.3 i annex) quan es sospita que les condicions de l'examen l'afavoreixen. Aquí la recomanació és simple; augmentar *n,* tenint en compte que a menys *n* es precisarà més *L* i a l'inrevés. Un exemple habitual són els exàmens de VF que solen tenir més llargada que els d'AM ja que amb només dues alternatives necessiten més ítems para compensar el possible efecte de la conjectura.

Quines taxonomies de coneixement s'aplicaran? La pregunta no es refereix al contingut o matèria que avaluen els ítems, sinó a l'enfoc com s'han plantejat. Molts exàmens AM identifiquen el domini d'una assignatura amb la capacitat de recordar o evocar dades, conceptes, principis, etc. Tanmateix aquesta limitació no és atribuïble a les preguntes AM sinó a la manera com s'han plantejat ja que es poden dissenyar amb enfocs diferents als purament memorístics (comprendre, analitzar, sintetitzar...) com es veurà a l'apartat 2. El més recomanable és que taxonomia adoptada a l'examen sigui equivalent amb l'enfoc emprat en el procés d'ensenyament-aprenentatge. La combinació entre la taxonomia i el contingut avaluat (els ítems també guarden relació amb els diferents aspectes tractats en el procés d'ensenyament-aprenentatge) determina la representativitat o validesa de contingut de l'examen.

Com es corregirà i puntuarà? La majoria d'ítems AM es puntuen com 1 (encert) i 0 (error) i es corregeixen, manual o mecànicament, a partir d'una plantilla o clau. Només es considera encert quan l'examinat tria l'alternativa correcta. Tot i això hi ha variants com quan cada alternativa de resposta s'associa a un tipus de coneixement incorrecte, parcial o total. Per exemple, per un ítem de 5 opcions de resposta, una totalment incorrecta, tres parcialment correctes en diferent grau i una altra totalment correcta, es podria assignar un pes de 0; 0,25; 0,50; 0,75, i 1 respectivament. Una altra opció és que la màxima puntuació (encert) correspongui a l'elecció d'una combinació d'alternatives (això no és molt recomanable). Aquest cas planteja dues qüestions a considerar; en primer lloc com es tractarà l'efecte de la conjectura i, en segon, la viabilitat d'una correcció mecanitzada doncs moltes lectores solen acceptar només una alternativa com certa.

1.2. Reptes i qualitats generals d'un examen

Las proves d'avaluació assumeixen tres reptes metodològics importants ja que proporcionen mesures indirectes, probabilístiques i d'interpretació relativa.

Las mesures són indirectes per que el resultat d'una prova reflecteix parcialment allò que pretén mesurar. Això connecta amb la suposa-

da representativitat dels ítems. Després d'un curs de primers auxilis, quines preguntes reflecteixen la capacitat dels alumnes per tractar una ferida? L'examen teòric per obtenir el permís de conducció de vehicles o embarcacions informa del nivell de preparació del futur conductor? Tècnicament aquestes preguntes tracten de la validesa de les inferències que es realitzen a partir de les puntuacions de l'examen (per exemple, si la puntuació assolida és alta, és que l'alumne deu sap tractar una ferida, o si l'alumne te una puntuació baixa, es que no té prou coneixements para pilotar una embarcació) i són habituals les crítiques a exàmens memorístics o sensibles a l'entrenament previ i a la conjectura (que no reflecteixen la suposada preparació de l'examinat).

Un examen escrit (obert) d'anglès reflecteix realment el nivell de competència lingüística de l'examinat? què expressa realment el resultat dels exàmens MIR (metge intern resident) o PIR (psicòleg intern resident?

En el cas concret d'AM aquests dubtes van més enllà del propi format. Molts estudiants prefereixen exàmens «*tipus test*» per la creença que donen més oportunitats d'encert que les preguntes obertes. Per definició, si els ítems són representatius del nivell de l'examinat, el format de les preguntes no hauria d'intervenir en la puntuació però en realitat això no és així. Sovint l'entrenament previ amb models d'examen, la conjectura i altres factors intervenen en la puntuació.

En segon lloc, el problema de la mesura probabilística es refereix a que tota puntuació d'un test incorpora sempre un component d'imprecisió. Psicomètricament, les puntuacions d'un test mai són valors exactes (per això la imprecisió està lligada intrínsecament amb la fiabilitat o, més aviat, amb la falta d'ella) sinó estimacions de la verdadera mesura. En un examen on s'aprova amb un 5 una puntuació de 4,95 implica suspens? En molts casos és així malgrat això comporti assumir que l'instrument té suficient precisió com per discriminar una diferència de 0,05 punts. L'examen té realment aquesta sensibilitat? A la pràctica és possible estimar els nivells de fiabilitat i de precisió d'una prova. El més complicat és que el resultat sol desconcertar als autors del test ja que els obliga a prendre consciència de la gran imprecisió que solen tenir les avaluacions realitzades.

El darrer aspecte, la relativitat de les mesures, afecta especialment als test (per exemple a les proves d'avaluació psicològiques) i no tant als exàmens. En psicologia la puntuació d'un test sol interpretar-se, seguint un enfoc normatiu, en funció de les puntuacions d'un grup de persones que s'agafa com referència (per això es denomina grup normatiu). Per tant la interpretació d'una determinada puntuació dependrà (serà relativa a) dels trets de l'esmentat grup normatiu. En canvi en educació s'acostuma a interpretar les puntuacions de l'examen seguint uns criteris establerts per l'autor/a, el/la docent, el programa, etc. (enfoc directe o criterial).

D'aquests tres problemes se'n deriva una altre de més concret; l'esbiaix de les mesures. En la Teoria dels Test hi ha biaix quan alguna persona o un determinat col·lectiu queden sistemàticament perjudicats al respondre la prova. Un cas quotidià són els exàmens amb diferents models (versions) per a diferents grups d'alumnes. Si, per exemple, aquestes versions no són equivalents en dificultat, i un grup rep un model amb preguntes més difícils que la resta, aquest grup es veurà clarament desafavorit en els resultats que s'obtinguin. En altres casos el resultat d'un examen ve condicionat per la capacitat de comprensió lectora dels alumnes. Per exemple, per respondre i encertar problemes de càlcul mental cal entendre les preguntes i això perjudica als alumnes que, tot i tenint una bona capacitat matemàtica, no disposen de suficient nivell de comprensió lectora.

Els test d'AM solen penalitzar els errors (els errors resten) per tal que l'examinat sigui prudent i eviti respondre a base de conjectures provant d'endevinar la resposta. Això té efectes col·laterals ja que involucra aspectes emocionals i personals (impulsivitat, fatiga, ansietat, autoimatge, acceptació de risc, por al fracàs, etc.) que intervenen en afrontar les preguntes. En aquests casos el biaix rau en que alguns perfils de personalitat tendeixen a puntuar pitjor per qüestions independents a la seva capacitat.

En els exàmens d'AM també es produeix un altre fenomen lligat a l'experiència i l'entrenament. Després de resoldre molts models de proves similars a l'examen real l'examinat adquireix una habilitat para identificar estils i patrons en les preguntes que l'ajuden a detectar l'alter-

nativa correcta. De fet en moltes certificacions i acreditacions oficials s'aconsella als candidats que entrenin a base de respondre més i més models de proves. El biaix aquí desafavoreix els que no han entrenat i responen «només» en base a la seva capacitat o coneixements relatius a allò que s'avalua.

Eliminar possibles biaixos ha estat un dels arguments tradicionals a favor de l'ús de test d'AM ja que l'examinat no ha de crear ni elaborar la resposta, només triar entre varies opcions excloent així qualsevol subjectivitat en la correcció. En els test d'AM no intervenen la bona o mala lletra de l'examinat, tampoc la seva capacitat d'expressió. Tot plegat, de temps ençà, s'ha anat identificant els test d'AM com «objectius» ja que la plantilla de correcció aparentment elimina subjectivitats i errors de puntuació (factors humans). Per aquesta raó en molts àmbits educatius s'ha associat als exàmens d'AM un valor de rigorositat que en realitat no tenen assegurat. En aquesta associació es confon l'objectivitat del mètode de correcció amb l'objectivitat (bon funcionament) de l'instrument. Fent auditories (controls de qualitat) de proves oficials i d'exàmens acadèmics hem pogut detectar moltes disfuncions greus en aquestes proves. Les causes solen associar-se a ítems que no diferencien entre examinats amb més o menys nivell (no discriminen) o a plantilles de correcció que presenten dubtes d'adequació. En molts exàmens, contra el que s'espera, existeixen preguntes amb més d'una alternativa que funciona millor que la que suposadament és la correcta (els examinats de més nivell tendeixen a triar sistemàticament una alternativa que no és la que consta a la plantilla). La situació més greu es produeix quan l'omissió és la resposta que des d'un punt de vista psicomètric funciona millor que la «correcta» (els examinats amb més nivell tendeixen sistemàticament a deixar en blanc la resposta).

Totes aquestes constatacions són habituals i porten a atribuir el valor d'objectivitat del test en molts altres elements a del seu format o l'existència d'una plantilla de correcció.

1.3. Disseny de l'examen

Quan ja s'ha decidit el format dels ítems (AM) i el contingut a avaluar, el procés complet de creació d'un examen passa por sis fases qualitatives i, opcionalment, per cinc de quantitatives.

El primer bloc de fases afecta a totes les proves i tracta del procés de creació i d'administració de l'examen. La informació la genera o recull bàsicament l'autor de la prova. Quant al segon bloc (anàlisi dels resultats) és recomanable però no sempre factible ja que precisa de software adient d'anàlisi.

Bloc 1
1. Establir la taula d'especificació d'objectius (TEO)
2. Regles de generació d'ítems (RGI)
3. Crear un banc d'ítems (BI) i establir un vector descriptor d'ítem (VDI)
4. Establir un vector descriptor de persones (VDP)
5. Maquetar i editar la/les formes de la prova i decidir el procés de aplicació
6. Sistema de puntuació i publicació de resultats (sense garanties)

Bloc 2
7. Recollir la matriu de dades en brut
8. Auditoria psicomètrica de la prova i proposta de reajustaments de l'examen
9. Sistema de puntuació i publicació de resultats (amb garanties)
10. Anàlisi comparatiu de les dades de l'auditoria amb VDI i VDP
11. Efectes i reajustaments en les fases 1, 2, 3 i 4

Aquesta seqüència de treball és orientativa i admet variants. Per exemple, l'entorn d'aplicació canvia si és tracta d'un examen convencional en paper o bé online. En el primer cas es poden establir diverses versions, models o formes amb les mateixes preguntes encara que presentades en diferent ordre. Aquests models també es poden construir amb preguntes diferents sempre que s'assumeixi que són equivalents en representativitat i dificultat.

En les proves online també es pot aleatoritzar l'ordre intern de les preguntes i de les alternatives de cada pregunta, en aquest cas un ordre diferent per a cada examinat. Així, diferents examinats poden respondre a un conjunt de preguntes, iguals o diferents, presentades en diferent ordre i trobant les seves alternatives en posicions també diferents.

Quan s'apliquen com un *quiz* també es pot associar missatges a cada alternativa de resposta de manera que si són escollides proporcionen un *feedback* explicatiu de les respostes correctes, i especialment, de les incorrectes. En aquests casos l'avaluador assumeix un principi d'Independència Local (IL), és a dir, que la resposta a una pregunta només es deu a la capacitat de l'examinat i no a l'efecte d'altres respostes a altres preguntes (aprenentatge progressiu, per eliminació o «descart», conjectura, efecte d'halo,...). La IL és una condició difícil d'acceptar i més sense una anàlisi psicomètrica que la garanteixi.

Altres avantatges de les proves online són el registre del temps dedicat a cada resposta i de la seqüència de rectificacions. En els següents apartats desenvoluparem ambdós blocs amb més detall.

2. LA TAULA D'ESPECIFICACIÓ D'OBJECTIUS

El primer pas estratègic a l'hora de crear un examen consisteix en definir una matriu o taula que combini les àrees de contingut a avaluar (files) i alguna classificació del tipus de coneixement que es pretén avaluar (columnes). La combinació d'ambdós aspectes produeix una taula d'especificació d'objectius (TEO). Els ítems de la prova presentaran les característiques identificades a les cel·les de la TEO.

Els continguts i els objectius del curs a avaluar (files) tenen a veure amb les parts, unitats didàctiques o formatives, mòduls, temes i subtemes, etc., en que s'estructura la matèria. Una mateixa matèria pot ser abordada des de diferents nivells de complexitat, depenent dels objectius formatius. Aquests són els representats a les columnes de la taula. No és el mateix preguntar sobre qué és la «humitat relativa» que plantejar quina utilitat pràctica té o quina relació manté amb altres indicadors com la temperatura o la pressió atmosfèrica. En el primer cas a l'examinat només li cal recordar la definició mentre que en els altres ha d'haver entès el concepte i ser capaç de relacionar-lo amb altres.

Un dels sistemes més reconeguts per estructurar els nivells de complexitat és la taxonomia de Bloom, que contempla 6 nivells o enfocs, des del més bàsic fins al més complex:

- **Coneixement**: capacitat de recordar coses (termes, principis, normes, mètodes, teories, etc.) prèviament apreses. Es basa en la memòria i no assumeix que necessàriament es comprengui allò que es recorda. En un examen són preguntes que demanen definir, distingir, il·lustrar, identificar, recordar, reconèixer, etc. Són útils per valorar vocabulari, terminologia, definicions, noms, dates, persones, llocs, propietats i trets, fenòmens, formes, usos, costums, regles, símbols, estils, accions, processos, classificacions, categories, mètodes, tècniques, tractaments, principis, fonaments, lleis, elements, teories, models, fórmules, etc.
- **Comprensió**: capacitat de captar el significat o sentit directe de la informació presentada (de manera verbal, gràfica, simbòlica, etc.). És un nivell superior a l'anterior, implica una interiorització del que

s'ha aprés. Les preguntes requereixen una interpretació personal dels temes plantejats. Davant un concepte conegut l'examinat l'ha de descriure emprant paraules diferents o distingir-ne aspectes essencials o derivar-ne conseqüències directes i evidents. En un test són preguntes que demanen explicar, interpretar, diferenciar, distingir, demostrar, inferir, concloure, predir, etc.

- **Aplicació**: capacitat per utilitzar els coneixements en la resolució de problemes. Les preguntes plantegen a l'examinat la solució de situacions i problemes nous emprant principis, regles, mètodes, teories, etc. prèviament apresos. Els models de problemes han de ser similars però no iguals als tractats durant l'aprenentatge. D'aquesta manera l'examinat aplica, desenvolupa, organitza, etc. els seus coneixements en circumstàncies diferents de les emprades com exemples.
- **Anàlisi**: capacitat para desglossar la informació rebuda identificant-ne els elements que la composen i també les seves interrelacions i estructura. En aquests ítems l'examinat ha de reconèixer, detectar, distingir, identificar, classificar, discriminar, contrastar, comparar, etc. les diferents parts que constitueixen l'objecte de la pregunta.
- **Síntesi**: capacitat de reconèixer els elements i parts que formen un tot. Es tracta de gestionar fragments, parts o components, organitzar-los i combinar-los per tal que formin una estructura nova que inicialment no es mostrava com a tal. Amb la informació disponible l'examinat ha de generar un producte original (idea sobre un tema, pla d'acció, model explicatiu...). En aquests ítems, és preferible que l'examinat pugui escriure, produir, transmetre, modificar, documentar, planificar, dissenyar, modificar, especificar, etc.
- **Avaluació**: capacitat de fer judicis quantitatius i/o qualitatius sobre el valor o mèrit d'idees, mètodes, instruments, resultats, projectes, programes, etc. Demana un pensament crític (no una opinió), a partir d'uns criteris preestablerts. En aquests ítems no n'hi ha prou en conèixer i comprendre una determinada matèria o contingut sinó que cal emetre judicis de valor en termes lògics o ajustats a normes i regles. L'examinat ha de jutjar, argumentar, validar i decidir.

La taula 1 mostra un exemple de TEO per a un examen d'una assignatura d'educació industrial en un curs sobre reparació de sistemes de suspensió i direcció d'automòbils.

Taula 1. Exemple d'una taula d'especificació d'objetius

		Con.	Com	Apl.	Tot.	%
Xassis i suspensió	Elàstics i amortidos de xoc	7	3		10	8.3
	Alineació	9	6	2	17	14.2
Principis de funcionament	Mecanismes de direcció	2	1		3	2.5
	Principis estabilitzadors	3	5	1	9	7.5
Servei i reparació	Diagnòstic d'errors		7	4	11	9.2
	Us de les eines	7	1	3	11	9.2
	Tècniques d'alineació	1	1		2	1.7
	Dirección i balanç	1		1	2	1.7
Tipus de sistemes de frens	Campanes i patins	2	2		4	3.4
	A disc	1	2		3	2.5
	Hidràulics	4	1		5	4.2
Principis de funcionament	Pressions mecàniques-hidràuliques	2	4	8	14	11.7
	Coeficients de fricció	3	3		6	5
Diagnòstic i manteniment	Indicacions d'error i	4	5	8	17	14.2
	Reparació de campanes, línies i cilindres	2	1	1	4	3.4
	Us d'equips i eines	2			2	1.7
	Total	50	42	28	120	
	%	41.7	35	23.3		100

La TEO determina l'estructura de futur examen. A l'exemple es va decidir que la prova constés de 120 ítems distribuïts en 6 continguts (files principals) diferents que es subdivideixen en altres de més concretes (files secundàries). Tot i l'orientació aplicada de l'assignatura, en aquesta TEO els autors van optar principalment per les categories de coneixement i comprensió (76,7% dels ítems, com es pot veure en la part inferior ombrejada de la taula) de la taxonomia de Bloom i només un 23,3% d'aplicació. A la pràctica això és habitual ja que amb ítems AM les categories que millor funcionen són coneixement i comprensió sent progressivament més difícil crear ítems d'anàlisi, síntesi i avaluació, més adients amb preguntes de resposta oberta.

Encara que la taxonomia de Bloom és la més coneguda, n'existeixen d'altres derivades, també consolidades, com la FIO (*the framework for intructional objectives taxonomy*) o LOGIQ (*logical operations for generate intended questions*). FIO és una proposta pràctica general que relaciona categories intel·lectuals amb altres de tipus psicomotriu i de valors i actituds en l'àmbit afectiu. En el disseny d'exàmens les categories in-

tel·lectuals contemplades són 10: interpretar, classificar, inferir, comparar, generalitzar, sintetitzar, analitzar, hipnotitzar, avaluar i predir. La taxonomia LOGIQ està molt més orientada a la creació de preguntes i inclou 6 categories: repetició, resum, explicació, predicció, aplicació i avaluació.

No existeix una regla fixa sobre la quantitat necessària de files ni de columnes per elaborar una TEO. El nombre de files ve determinat per les parts de la matèria avaluada que es consideren rellevants i el de columnes pels enfocs o tipus de coneixement que desitja assolir. Sigui com sigui, el principi general és que la TEO resultant hauria de reflectir bé el procés d'ensenyament-aprenentatge que s'ha seguit. Les cel·les de la taula (combinacions de continguts i nivells de complexitat), han de proporcionar elements d'ajut en la creació de les preguntes. La importància o «pes» que cada cel·la de la taula tingui dins l'avaluació, i per tant la quantitat d'ítems necessaris per poder representar l'esmentat pes, vindrà determinada de nou, per 1) el programa de l'assignatura (extensió, estructura...), 2) la importància assignada a cadascuna de les parts de la matèria (files), que pot ser proporcional a la quantitat de temps invertit en la docència de cada part, i 3) els tipus de nivells de complexitat exercitats en el procés d'ensenyament-aprenentatge (columnes).

Fent un resum, aquestes són algunes directrius i recomanacions d'ús de les TEO.

- L'estructura de la TEO ha de reflectir el contingut de la matèria i el programa del curs a avaluar. Això és fonamental per a validar l'examen.
- La TEO ha de ser exhaustiva quant al desglossament dels continguts de la matèria a avaluar.
- També hauria de representar correctament la complexitat dels nivells de coneixement (no és recomanable sobrecarregar la prova amb ítems de coneixement).
- Per la seva naturalesa el format AM no sempre s'ajusta bé a la taxonomia de Bloom.
- En ocasions l'estructura de la TEO acaba evidenciant l'existència de dos o més blocs de continguts clarament diferenciats. Quan passa això és recomanable considerar si realment es tracta d'una sola pro-

va o bé caldria fer-ne més d'una. Aquest punt és important quan es preveu fer una anàlisi o auditoria.

- Convé revisar el nombre i dificultat prevista dels ítems de cada cella de la taula i detectar possibles biaixos indesitjats (p.e. sovint els ítems més difícils són d'una part de la matèria i/o categoria concreta de la taxonomia). La dificultat de la prova té que mantenir una certa relació amb el nivell dels alumnes que seran avaluats (una mateixa prova pot ser fàcil per a un grup d'alumnes i difícil per un altre). Una cop decidit el nivell de dificultat global, és convenient que no tots els ítems tinguin un nivell de dificultat similar, sinó que hi hagi preguntes més fàcils i més difícils de manera que la majoria dels examinats en trobin d'un nivell similar al seu.
- La dificultat dels ítems prevista per a qui dissenya la prova s'hauria de distribuir entre les diverses cel·les.
- En relació amb al valor o «pes» dels ítems d'una cel·la és preferible afegir o reduir-ne la quantitat que no pas ponderar les respostes correctes.

3. REGLES DE GENERACIÓ D'ÍTEMS (RGI)

Una TEO és un bon recurs para iniciar el disseny d'un examen però a l'hora de crear el material que el forma calen pautes més concretes. Les RGI són precisament el conjunt de directrius que guien el procés de creació de les preguntes.

Des d'un enfoc estàndard els ítems AM estan constituïts per un enunciat i diverses opcions/alternatives, una d'elles correcta i la resta incorrectes (també anomenades distractors). La tasca de l'examinat és seleccionar l'opció correcta de cada ítem. L'enunciat es pot expressar directament, com una afirmació, o de manera interrogativa. Les alternatives poden fer-se verbalment (mitjançant frases), numèricament (p.e. amb fórmules matemàtiques) i/o gràficament (p.e. amb imatges).

Aparentment els ítems AM són fàcils de crear, l'experiència però mostra que construir un bon ítem AM no és una feina senzilla. Les següents recomanacions són directrius de bones pràctiques para crear ítems. La llista, que no és exhaustiva, combina pautes generals i propostes fruit de constatacions de problemes habituals en processos de revisió i auditoria.

3.1. Recomanacions generales

- El nombre d'alternatives s'ha de mantenir fix en tot el test. Aquesta condició s important quan s'apliquen penalitzacions dels errors a partir de fórmules establertes (veure annex).
- El nombre d'alternatives pot oscil·lar de 3 a 10, sent habitual 3, 4 o 5. A l'apartat 7.4 s'aprofundeix en aquest punt.
- Es preferible distribuir les alternatives verticalment i no en horitzontal.
- Cal revisar si hi ha errors ortogràfics, gramaticals, abreviatures no utilitzades prèviament, etc. Els docents solen parar més atenció al contingut i a la forma de les alternatives correctes. Per això sovint les errònies tenen més faltes ortogràfiques i d'expressió (aquest detall pot oferir pistes no desitjades).

- Verificar que l'ítem tracta un aspecte rellevant del contingut de l'assignatura (definit a les files de la TEO) i un nivell de complexitat desitjat (definit a les columnes de la TEO). Cal que el docent tingui present les característiques de l'alumnat que respondrà correcta o incorrectament (fins i tot del tipus de resposta incorrecta) a cada ítem.
- Assegurar que, en cada ítem, hi hagi una sola resposta correcta (o la millor, si ho especifiquen les instruccions). En força auditories d'exàmens hi ha problemes amb la plantilla de correcció ja que es detecten alternatives que tot i sent considerades errònies es comporten psicomètricament com alternatives correctes (per exemple, quan els examinats amb millor puntuació total trien sistemàticament una alternativa errònia concreta o bé ometen la resposta, mentre que els de pitjor puntuació trien l'alternativa suposadament correcta).
- A no ser que es tracti d'una prova de comprensió verbal, lèxica, ortografia, etc. no hi s'ha de confondre la capacitat avaluada amb la de comprensió del redactat. El raonament verbal i la comprensió lectora són importants i necessàries, però no han de ser un element que condicioni les respostes a la prova.
- Tots els ítems han de ser independents (principio d'IL). No ha d'haver-hi connexions ni interdependència de continguts entre els ítems (p.e. la resposta a un determinat ítem no hauria de dependre de haver encertat un ítem previ).
- És convenient evitar termes absoluts com «tot», «res», «sempre», «mai», etc. Normalment van associats a alternatives errònies.
- És preferible evitar frases fetes o tòpics i expressions (o exemples) literals de llibres de text o apunts.
- Millor evitar enunciats, i alternatives, redactats en negatiu o amb dobles negacions. Si hi ha ítems amb elements negatius a l'enunciat cal destacar-ne el terme mitjançant majúscules (NO) o subratllant-lo (no) per tal que quedi ben visible.

3.2. Recomanacions referides a l'enunciat

- L'enunciat de l'ítem ha de plantejar una qüestió o formular un problema de manera comprensible, clara i específica, sense necessitat de llegir les alternatives. Cal evitar explicacions innecessàries.

- La informació bàsica de la pregunta ha de raure a l'enunciat, que hauria de ser més llarg que les alternatives.
- Cal evitar regular la dificultat de l'enunciat (i per tant de la pregunta) emprant un vocabulari de difícil comprensió, diferent de l'habitual o que no resulti familiar a l'examinat.
- Generalment una de les opcions de resposta és correcta i la resta incorrectes, però en determinades ocasions és passa al revés (una opció és incorrecta i la resta correctes). En aquest segon cas la situació hauria de quedar destacada (p.e. subratllant) a l'enunciat.
- L'enunciat d'un ítem no ha d'ajudar a respondre un altre (principi IL).

3.3. Recomanacions referides a les alternatives

- Una bona manera de plantejar les alternatives és preparar una breu justificació de cadascuna (raonant el perquè són o no correctes). Això és habitual en proves online que ofereixen retroalimentació en cada possible resposta. El fet de redactar una justificació per a cada opció fa reflexionar sobre la seva adequació, millorant la construcció de l'ítem. A més, proporciona arguments enfront possibles reclamacions.
- Les alternatives haurien de ser el més curtes possible.
- Les alternatives no han de començar totes amb una mateixa expressió. És un fet molt freqüent que sobrecarrega el temps de lectura. En aquests casos cal traslladar l'expressió repetida a l'enunciat.
- Si les alternatives són figures o gràfics mantindran una mida i aspecte proporcional.
- L'ordre de presentació de les alternatives ha de ser neutre, o preferentment alfabètic o numèric, sigui en ordre creixent o decreixent.
- Un ítem estarà ben formulat si les freqüències d'eleccions de les alternatives errònies es distribueixen aproximadament de manera uniforme al marge de l'opció correcta.
- No s'hauria de poder encertar un ítem sense dominar el contingut. Les persones amb baix nivell de domini tendeixen a respondre per «descart», i les alternatives que, per algun motiu, no semblen atractives són candidates a ser rebutjades. Per això, la redacció dels distractors ha de mantenir un atractiu, extensió, estil gramatical i aspecte similars entre sí i l'opció correcta.

- Es convenient formular les alternatives incorrectes de manera que corresponguin als errors més típics comesos pel examinats.
- Las alternatives han de presentar continguts diferenciats i no solapats.
- No són recomanables les alternatives que afirmen o neguen altres. Per exemple, es desaconsella emprar l'opció «cap de les anteriors». Si per alguna raó es utilitzar sempre es preferible «cap de les altres opcions», especialment en proves online on la posició de les alternatives pot variar per a cada examinat. Por altre banda s'ha estès la creença de que «totes les opcions anteriors/altres són certes» sol ser certa mentre que «totes les opcions anteriors/altres són falses» acostuma a ser falsa.
- Un cas problemàtic és «cap de les opcions és certa». Si és la correcta sol ser un argument habitual en processos de reclamació ja que incorre en una contradicció al implicar la pròpia alternativa invalidant l'ítem.
- Es millor evitar les alternatives humorístiques. Aquest ítems acostumen a mostrar un mal funcionament en una auditoria i propicien la conjectura.

3.4. Claus i patrons

Cal evitar les pistes, claus reveladores i paraules coincidents entre l'enunciat i les alternatives. Les més freqüents són les claus verbals (associació verbal entre pregunta i resposta), claus gramaticals (al concordar gramaticalment o per gènere o noms l'enunciat amb una alternativa) i claus per heterogeneïtat (per discordança entre conceptes).

Un efecte habitual al homogeneïtzar les alternatives és solapar continguts (efecte no desitjat, com ja s'ha indicat), això porta a l'examinat a inferir connexions i possibles patrons d'encert.

És important evitar que l'alternativa correcta segui la de més contingut o extensió (és un hàbit molt freqüent quan es redacten opcions).

En exàmens convencionals, en paper, sovint s'observen patrons sistemàtics en la disposició dels ítems. Molts cops són fruit de rutines i poca atenció per part dels autors que no en són conscients. Els pa-

trons més freqüents són de tres tipus: posició de l'alternativa correcta dins l'ítem, posició de l'alternativa correcta a la pàgina i transició entre alternatives correctes (es detallen tot seguit). Aquests patrons tenen efectes col·laterals negatius en l'avaluació i són molt fàcils d'evitar (n'hi ha prou amb aleatoritzar la presentació de les opcions).

Posició de l'alternativa correcta dins l'ítem: En exàmens universitaris en paper no sempre hi ha una distribució homogènia de la posició de l'alternativa correcta. La taula 2 correspon a un cas real de 55 preguntes de 4 alternatives A, B, C i D. A la columna «Total» de la dreta consten les vegades que cada lletra és correcta. S'aprecia una predominança de B i C respecte a A i D així com una diferència de 7 entre la més freqüent (17) i la que menys (10). En el cas de ser una distribució homogènia cada alternativa hauria de ser certa 13 o 14 vegades evitant així la creació d'expectatives.

Taula 2. Distribució de les alternatives de resposta (A, B, C i D) i del nombre de opcions correctes en un examen real de 55 preguntes

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

Posició de l'alternativa correcta dins la pàgina: és una altra constatació habitual. Si partim les pàgines d'un examen en tres seccions verticals (veure figura 1) és freqüent que a la part superior les respostes correctes siguin les últimes (C, D, E). A la secció inferior passa el contrari, les correctes acostumen a ser les primeres alternatives (A, B...). Quant al sector central no hi ha un patró tan definit i predominen altres hàbits (alternativa amb més extensió, «totes les anteriors són certes»...).

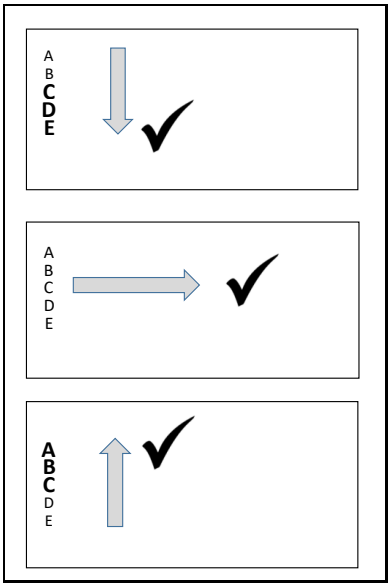


Figura 1. Distribució d'alternatives correctes en un examen

Tornant a l'exemple de la taula 2, les columnes 1 a 5 corresponen a la posició de les respostes correctes dins cadascuna de les 11 pàgines que ocupa la prova (cinc ítems per pàgina). Les cel·les informen del nombre de vegades que cada alternativa és certa destacant en gris el major valor. S'observa ja una certa tendència en les cel·les ombrejades, l'alternativa D és la més freqüent en el primer ítem de les pàgines mentre que A i B ho són en el darrer. Les posicions segona, tercera i quarta es reparteixen principalment entre les alternatives centrals. Aquest resultat torna a evidenciar una tendència anòmala que es podria evitar simplement aleatoritzant la posició de les alternatives dins les pàgines.

De transició entre alternatives correctes: en un examen no hauri-en d'haver-hi patrons sistemàtics de la seqüència d'encerts (A-D-B-A-C....). Cas contrari predisposen l'elecció de resposta i les decisions de l'examinat. La figura 2 mostra la freqüència de transició entre dos ítems consecutius d'un examen de 160 preguntes de quatre alternatives A, B, C i D.

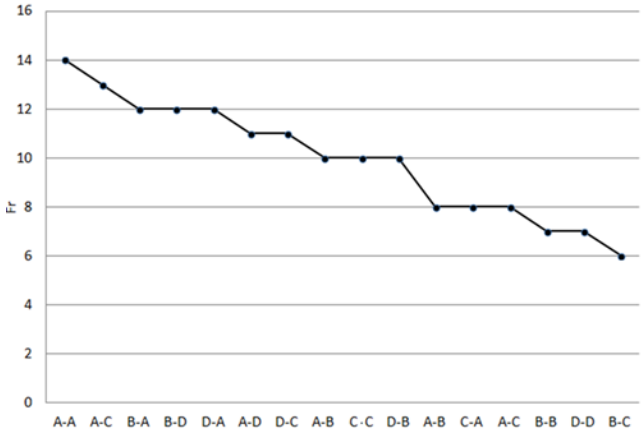


Figura 2. Freqüència de transició entre alternatives (A, B, C i D) en parells consecutius d'un examen de 160 preguntes

En ítems de quatre alternatives es poden produir 16 transicions (4^2) que en la figura s'han ordenat de major a menor freqüència d'ocurrència a l'examen. La transició més freqüent és A-A (14 vegades i la que menys B-C (sis vegades). En cas que fossin aleatòries n'hi hauria d'haver 10 de cada. A més, en aquest exemple destaca especialment la predominança del trànsit idèntic A-A per sobre dels altres tres; C-C, B-B i D-D.

Una manera gràfica de comprovar aquest patró és veure la seqüència completa de l'examen. La figura 3 correspon a una prova de 64 preguntes (eix X) de quatre alternatives 1-2-3-4 (eix Y). El traçat gruixut indica l'itinerari entre alternatives. La línia prima és un suavitzat de l'anterior para veure millor la tendència de l'opció correcta. No s'observa una pauta concreta (de vegades hi ha un salt entre l'alternativa 1 i la 4, altres entre l'alternativa 1 i la 2...) de manera que les «dents» de serra són desiguals (n'hi ha de grans i de petites). Si la seqüència no estigués suficientment aleatoritzada es podria veure una pauta concreta (per exemple, salts més habituals entre les opcions extremes, o tendències a repetir la mateixa opció de resposta en ítems consecutius).

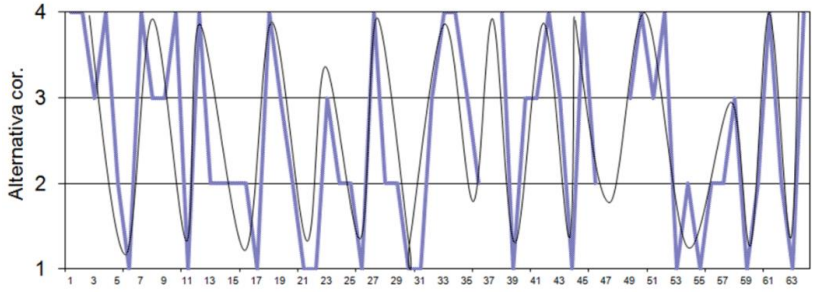


Figura 3. Patró de trànsit entre alternatives correctes d'ítems consecutius

Combinació de patrons: els patrons i claus reveladores són la base no escrita de molts recursos populars per a encertar preguntes sense saber la resposta. Des de fa anys circulen «regles», «receptes», trucs o consells entre examinats per afrontar els exàmens. A la xarxa hi ha multitud de pàgines, vídeos i foros dedicats a «com aprovar sense estudiar». Els principals arguments es deuen bàsicament al mal disseny dels test. Un examinat sense res a perdre, a qui no li preocupa que li penalitzin els errors, pot provar «sort» i aplicar algunes regles senzilles. Por exemple, davant qualsevol dubte, pot triar sempre l'alternativa amb més paraules o contingut. També pot evitar marcar l'A o la B en els primers ítems de cada pàgina i la D o la E en els últims. També ha d'evitar les alternatives amb errors ortogràfics, etc. Si a aquests aspectes hi afegim les claus reveladores particulars que faciliten alguns ítems, les possibilitats d'èxit augmenten.

Patrones d'autor: molts docents tenen hàbits de repetició a les seves classes. Reiteren certes paraules i expressions (falques, mots crossa, comodins, frases fetes, clixés...) que el caracteritzen davant dels seus alumnes encara que siguin de diferents promocions. Quelcom semblant passa amb els apunts i exàmens, ja que solen arrossegar patrons personals que es mantenen curs darrera curs, de manera no conscient per el docent però sí per a qui s'ha d'examinar i disposa de models d'apunts i d'exàmens anteriors. Existeixen pàgines web que ofereixen autèntiques biblioteques de recursos al respecte. Per exemple, hi ha docents que mantenen patrons de transició estables, altres converteixen el «cap de les anteriors» en un comodí per a totes les seves preguntes.

Els patrones personals són difícils de detectar, requereixen una revisió cronològica d'apunts i d'exemplars anteriors d'exàmens que permetin identificar reiteracions sistemàtiques.

Altres patrons: en els exàmens també hi ha patrons peculiars i sorprenents, com la llei de Benford. Segons aquesta, en el nostre entorn predominen els números (no generats aleatòriament) que comencen per 1 respecte a als que ho fan pel 2, 3 i així fins el 9. La distribució segueix un perfil molt definit (30,1 % comencen per 1; 17,6 %, per 2; 12,5 % per 3, i així successivament fins el 5 % corresponent al 9). Un estudi canadenc va revisar gran quantitat d'ítems d'exàmens confirmant que aquesta mateixa distribució es compleix amb els números emprats pels docents en les respostes de preguntes d'AM que impliquen quantitats i càlculs. Segons aquesta llei, quan es dubte hi ha un 50 % de possibilitats d'encertar si es tria un resultat que comença per 1, 2 o 3, alhora que baixen al 15 % si comencen per 7, 8 o 9.

4. EXEMPLES D'ERRORS HABITUALS EN ELS ÍTEMS

Tot seguit veurem alguns exemples reals de problemes freqüents en la redacció d'ítems d'AM. Evitar aquests errors no és només una qüestió formal o d'estil. L'experiència pràctica en auditories psicomètriques mostra com els ítems amb anomalies acostumen a tenir un mal funcionament avaluador i no resulten ser bones «peces» de l'examen.

En els exemples la resposta especificada com correcta por l'autor/a es destaca en **negreta**. S'ha respectat aquesta assignació i també el redactat original. Uns quants exemples incompleixen més d'una directriu. Tots pertanyen a exàmens diferents, per tant no es poden analitzar en conjunt intentant identificar patrons globals ja exposats.

En cada cas es presenta primer l'ítem i després una breu descripció dels problemes associats.

Ecosistema és...

A. La unitat fonamental per els estudis ecològics.
B. Un conjunt d'organismes vivents en una comunitat, juntament amb el seu entorn.
C. Tot alló relacionat amb les interaccions organisme-medi natural.
D. Totes les opcions anteriors són certes.
E. Cap de les opcions anteriors és certa.

L'enunciat és curt. Dues alternatives neguen o afirmen altres. Utilitza el terme «tot». Aquest ítem no es podria administrar online aleatoritzant les alternatives ja que D i E condicionen la resposta.

Marca la resposta correcta:

A. Com menys velocitat, més cap a l'esquerra anirà el vent (*veering*).
B. Com més velocitat, més cap a la l'esquerra anirà el vent (*baking*).
C. Com menys velocitat, més cap a l'esquerra anirà el vent (*baking*).
D. Com menys velocitat, més cap a la dreta anirà el vent (*veering*).

L'enunciat hauria de ser més llarg que les alternatives. L'expressió «Com» es repeteix i hauria de passar a l'enunciat.

Quan no és incorrecte afirmar que una investigació és ex-post-facto?

A. Quan l'investigador tracta amb dades qualitatives.
B. Quan l'investigador manipula la variable depenent.
C. Quan l'investigador no manipula la variable independent.
D. Quan l'investigador no té hipòtesi.

Doble negació en l'enunciat al combinar «no» amb «incorrecta», a més d'altres negacions en les alternatives C i D. Es repeteix «quan l'investigador» en les alternatives i s'haurà de passar a l'enunciat: «Quan l'investigador...».

La trombosi venosa es manifesta amb els següents signes excepte:

A. Dolor
B. Augment de volum
C. Calor local
D. Fred local

Si en preguntes anteriors s'ha anat demanant l'opció correcta ara pot ser una sorpresa tenir que marcar la incorrecta (seria recomanable destacar «excepte» a l'enunciat).

En un accident, quina de les següents afirmacions no és certa?

A. Es recomana fer el reconeixement en el mateix lloc on es troba.
B. No s'aconsella moure'l ni traslladar-lo fins que no s'hagi fet la primera valoració d'emergència.
C. Traslladar-lo pot agreujar la situació o causar-li noves lesions.
D. El primer és traslladar-lo a un lloc segur.

Hi ha una negació a l'enunciat i tres en una mateixa alternativa. A l'enunciat s'hauria de destacar «no és certa» (per exemple, «NO és certa» o «no és certa»). Les alternatives tenen diferent longitud.

Quina de les següents no correspon a una lesió per congelació profunda?

A. Necrosi de la pell o òssia.
B. Cura en 4-6 setmanes.
C. Hipersensibilitat al fred.
D. No deixa seqüeles.

Presència de negacions a l'enunciat (sense destacar amb negreta o subratllat) i a l'alternativa correcta dificultant la comprensió: L'opció D significa que no correspon que no deixa seqüeles?

Segons la Organització Mundial de la Salut, el 2016 l'esperança de vida al Japó és superior a...

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

L'esperança de vida al Japó el 2016 era de 84 anys, per tant les opcions A i B són correctes. Les alternatives de resposta no són excloents.

Si sabem que en una determinada població de 1.000 persones hi ha 300 dones i 700 homes, i volem fer un mostreig estratificat-proporcional de 500 individus, quantes dones tenim que incloure?

A. 250
B. 175
C. 300
D. Cap és certa.

Aquest ítem es pot encertar sense saber la resposta correcta (150 dones). Una de les opcions (C. 300) és poc versemblant ja que repeteix la mateixa xifra que a la població. Les alternatives no estan ordenades. En cas de mantenir el «cap...» s'hauria de referir a «les altres» o a «les anteriors»

Els cops de calor:

A. No es produeixen en ambients frescos i càlids.
B. Estan afavorits pel consum de fàrmacs simpaticomimètics.
C. No es poden prevenir amb una aclimatació prèvia.
D. No influeixen factors nutricionals ni dietètics

El germen dentari comença els moviments eruptius:

A. Quan acaba la calcificació coronària i comença la radicular.
B. Quan arriba als dos terços de l'arrel calcificada.
C. Just al finalitzar la calcificació radicular.

Exognàsia és:

A. Falta d'espai per a tercers molars.
B. Mossegada creuada bilateral.
C. Excés d'amplada d'una arcada dentària o d'un maxil·lar en la seva totalitat.

L'estil consultiu en la presa de decisions dels líders és el que deixa participar als seus subordinats en la decisió:

A. Verdader.
B. Fals, només als que tenen informació.
C. Fals, només els consulta sobre la seva decisió.
D. Fals, només rep informació o dona informació sobre el problema.

En els quatre exemples anteriors l'alternativa correcta és la de més extensió. L'alternativa A del primer sembla incoherent.

Gen es...

A. La unitat del material reproductiu.
B. El fragment de cromosoma que codifica la informació genètica de l'organisme.
C. Fragment d'un organisme que es reprodueix sexualment.
D. Totes les opcions anteriors són certes.
E. Cap de les opcions anteriors és certa.

Aquest ítem té l'enunciat molt breu. Les alternatives tenen llargades diferents (i la correcta és la més llarga). A l'opció correcta, la paraula «genètica» al·ludeix a «gen» que apareix a l'enunciat. S'utilitza el tipus «Totes són...» o «Cap és...:».

Las característiques d'un grup són diverses, entre elles podem definir:

A. Conjunt de persones interdependents.
B. Grup de persones que té un o més objectius comuns i que s'organitzen.
C. Conjunt de persones amb interessos comuns.
D. Totes les anteriors.

L'alternativa correcta té més contingut i comparteix la paraula «grup» amb l'enunciat (pista o clau). S'inclou una opció del tipus «totes són...».

Biotecnologia agrària és...

A. La modificació genètica de molècules de llavors o plantes amb finalitats aplicades.
B. Conjunt de tècniques d'inseminació i pol·linització.
C. Branca de la biologia que estudia els problemes del camp.
D. Totes les opcions anteriors són certes.
E. Cap de les opcions anteriors és certa.

Aquest ítem té un enunciat molt curt. Les alternatives D i E al·ludeixen a altres. L'alternativa correcta té més contingut i una clau reveladora «llavors i plantes» que connecta amb el terme «agrari» de l'enunciat. S'inclouen opcions del tipus «Totes són...» o «Cap és...».

La escala emprada per explorar l'estat de consciència es:

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

Quin tipus mastegador afavoreix el creixement i el desenvolupament dels maxilars?

A. Masticació maseterina.
B. Movimients d'obre i tanca.
C. Moviment del temporal.

En aquests dos exemples, tot i desconeixent la resposta, hi ha una clau reveladora entre «consciència»/«coma» i «mastegador»/«masticació» en les respectives alternatives correctes.

A l'hemisferi nord, d'esquena al vent, tens a la teva esquerra les baixes pressions i a la dreta les altes pressions. Això ho explica la llei de...

A. Buys Ballot.
B. Coriolis.
C. Murphy.
D. Cap és correcta.

En aquest ítem hi ha una alternativa humorística (C) que va ser poc escollida i que va fer augmentar les possibilitats de les altres tres. Inclou una alternativa (D) que nega totes incorrent en una contradicció.

ADN o àcid desoxirribonuclèic és...

A. L'àcid que transmet la informació heretada en els éssers vius.
B. Un àcid que caracteritza les cèl·lules sexuades.
C. Una molècula simple que composa a un bacteri.
D. Totes les opcions anteriors són certes.
E. Ninguna de les opcions anteriors és certa.

Aquest ítem redueix les possibilitats a només dues alternatives A i B que comparteixen la paraula àcid (clau) amb l'enunciat. Un altre cop, a l'hora de triar entre A i B la correcta és la que presenta més contingut. S'inclouen opcions del tipus «Totes són...» o «Cap és...».

Com es simbolitza un front fred en un mapa meteorològic? Una línia...

A. vermella amb triàngles
B. groga amb cercles
C. verda amb quadrats
D. Blava amb triàngles

Aquest ítem es veu afectat por l'associació habitual entre fred i color blau emprada en molts dispositius quotidians (aixetes, refrigeració, connexions,...) i la informació meteorològica. L'alternativa correcta és la única que comença amb majúscula.

La rinorrea és:

A. La sortida de sang pel nas.
B. Sortida de líquid pel nas.
C. Sortida de sang per l'orella.
D. Sortida de líquid per l'orella.

En aquesta pregunta l'enunciat és curt, es repeteix el començament de cada alternativa («Sortida») tot i que a la correcta és més complet («La») i entra en joc l'associació entre «rino» i «nas».

No està permès dur:

A. Ornaments al cap, joies i accessoris al cabell.
B. Pantalons curts.
C. Sabatilles esportives.
C. Genolleres recobertes.
D. Protecció per nas trencat.

Pregunta sobre reglament de Basket amb dues alternatives B i C massa evidents i sense atractiu que augmenten les possibilitats de les altres. Enunciat curt i amb una negació. L'alternativa A és la més extensa.

La pilota es considera viva quan:

A. La pilota surt fora per la línia de fons.
B. La pilota surt fora per la línia de banda.
C. La pilota surt de la mà de l'àrbitre en un salt entre dos.
D. La pilota està a les mans de l'àrbitre.
E. La pilota està a disposició de l'àrbitre.

En aquesta pregunta del mateix examen es repeteix l'inici de les alternatives (la pilota). L'enunciat hauria de ser «Es considera viva la pilota quan....». De nou la correcta és la de més contingut. El 50 % de preguntes de aquesta prova tenien l'alternativa correcta amb més text.

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

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

En aquesta pregunta d'un examen de anglès administrat a gran escala només van ser triades les alternatives A i C (baix atractiu de B i D). Encara que ofereix quatre opcions de fet funciona com un ítem de VF amb només dues. La disposició de les alternatives és horitzontal.

Segons els objectius podem distingir diferents tipus de grups:

A. Grups formals i informals.
B. Grups de treball grups de cohesió.
C. Grups de relació i grups de treball.
D. Grups de primaris i grups de secundaris.

En aquesta pregunta es repeteix innecessàriament la paraula «Grups» en cada alternativa

A la deglució adulta o madura:

A. Es contrauen els músculs posturals mandibulars.
B. Es contrauen els orbiculars.
C. Es contrauen els músculs cervicals.

En aquest exemple el terme «deglució» connecta amb «mandibulars» i facilita l'elecció per conjectura. A més, l'alternativa correcta és la de més contingut. Es repeteix «Es contrauen».

I ..... no satisfaction.

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

La resposta a aquest ítem venia condicionada pel títol d'una cançó popular del grup Rolling Stones. Al marge del nivell d'anglès els encerts augmentaven a mida que els examinats tenien més edat (esbiaix).

5. BANCS D'ÍTEMS (BI) I VECTORS DESCRIPTIUS

El resultat del procés de creació d'ítems seguint una estructura de TEO no és l'examen final sinó un banc d'ítems (BI). En aplicar les RGI per a cadascuna de les cel·les de la taula s'està produint realment una col·lecció estructurada d'ítems útils per a configurar diferents models o versions de l'examen definitiu. Des d'aquest punt de vista, un model concret d'examen no és més que una mostra possible de tots els que representen correctament la configuració de la TEO. En funció de la quantitat d'ítems generats en cada cel·la de la TEO es podran extreure més o menys models amb certes garanties d'equivalència (els ítems dels diferents models provenen de la mateixa TEO, i per tant, comparteixen continguts i objectius).

Quan es tracta de proves online, hi ha plataformes que permeten classificar els ítems en diverses categories i nivells segons criteris establerts per l'autor (per exemple, importància, dificultat, freqüència d'aparició en la docència...). En aquest cas, un cop introduïts els ítems a la plataforma i definits els criteris de configuració de l'examen (per exemple, proporció d'ítems de cada matèria i nivell de complexitat, limitacions en funció de la dificultat...) n'hi haurà prou en determinar la longitud desitjada de la prova, per tal que la aplicació generi diferents models aparentment equivalents. Algunes aplicacions incorporen també dades posteriors a l'administració de l'examen quan ja s'han analitzat les respostes. Generalment es tracta d'índex psicomètrics que informen del funcionament de cada pregunta (dificultat, discriminació, conflicte entre alternatives, etc.) i que serviran per a contrastar les dades proposades per l'autor amb les de l'anàlisi.

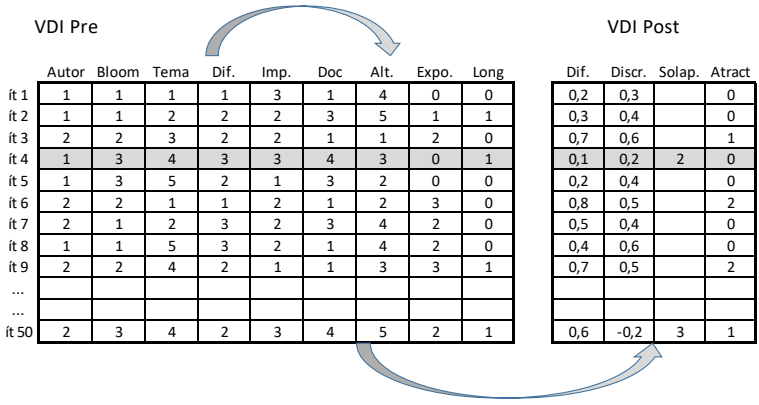
La noció de BI ha agafat força els darrers anys gràcies a propostes psicomètriques com la teoria de resposta d'ítem (TRI) i els test adaptatius informatitzats (TAI). En aquests test un algoritme decideix, en funció dels encerts i els errors de l'examinat, les preguntes que li va presentant per pantalla. Des d'aquest enfoc el centre d'atenció és el banc del que s'extreuen gradualment les preguntes. El test que finalment respon l'examinat constitueix només una breu selecció del contingut del banc ajustada al seu cas particular.

5.1. Vector descriptor d'ítem (VDI)

Tots els ítems que formen un examen estan caracteritzats per múltiples aspectes de disseny que poden aportar informació útil a diversos nivells. La unió de totes aquestes característiques codificades per a cada pregunta constitueix el vector descriptiu de l'ítem (VDI).

Ja sigui un examen generat automàticament des d'una aplicació o bé elaborat a la manera convencional, un recurs interessant per a revisar l'estructura de la prova consisteix en agrupar tots els VDI. La taula de la figura 4 mostra un exemple per a un examen de 50 preguntes amb cinc alternatives.

Cada fila de la taula correspon a una de les 50 preguntes. Les columnes contenen la informació que defineixen els VDI en aquest cas. D'aquesta forma, una determinada fila (vector) conté tota la informació rellevant per a descriure les característiques de l'ítem corresponent (per exemple, en la figura 4 s'ha ombrejat el VDI de la pregunta 4). Segons el context cada examen permetrà dimensionar VDI de manera diferent. En aquest cas s'ha considerat rellevants els nou criteris (columnes) següents:



	Autor	Bloom	Tema	Dif.	Imp.	Doc	Alt.	Expo.	Long	Dif.	Discr.	Solap.	Atract
ít 1	1	1	1	1	3	1	4	0	0	0,2	0,3		0
ít 2	1	1	2	2	2	3	5	1	1	0,3	0,4		0
ít 3	2	2	3	2	2	1	1	2	0	0,7	0,6		1
ít 4	1	3	4	3	3	4	3	0	1	0,1	0,2	2	0
ít 5	1	3	5	2	1	3	2	0	0	0,2	0,4		0
ít 6	2	2	1	1	2	1	2	3	0	0,8	0,5		2
ít 7	2	1	2	3	2	3	4	2	0	0,5	0,4		0
ít 8	1	1	5	3	2	1	4	2	0	0,4	0,6		0
ít 9	2	2	4	2	1	1	3	3	1	0,7	0,5		2
...													
...													
ít 50	2	3	4	2	3	4	5	2	1	0,6	-0,2	3	1

Figura 4. Exemple del vector descriptor d'un conjunt de 50 ítems

- Autor: aquesta prova va ser desenvolupada per dos docents, els valors 1 i 2 indiquen qui dels dos va fer cada ítem.

- Bloom: 1 (coneixement), 2 (comprensió) i 3 (aplicació) informen de la categoria de la taxonomia de Bloom a la que correspon cada pregunta. En aquest examen només es van utilitzar aquestes tres característiques de la taxonomia (columnes de la TEO).
- Tema: d'1 a 5 s'indica a quin dels cinc temes de l'assignatura pertany cada pregunta (files de la TEO).
- Dif.: cada autor va estimar la dificultat de les seves preguntes en tres nivells; fàcil (1), mitjana (2) i difícil (3). Aquesta estimació requereix d'un exercici quasi empàtic, doncs l'autor ha de ser capaç de preveure si la majoria dels seus estudiants contestaran correcta o incorrectament l'ítem.
- Imp.: cada autor va estimar d'1 a 3 la importància del tema tractat en la pregunta en relació amb el domini de l'assignatura.
- Doc.: indica la font principal de preparació i documentació del tema tractat en la pregunta (1) classe magistral, (2) apunts, (3) pràctiques, (4) lectures complementàries).
- Alt.: és l'alternativa correcta de l'ítem, en aquest cas d'1 a 5 (A..., E).
- Expo.: indica el número de vegades que ja s'ha utilitzat (exposat públicament) l'ítem en exàmens anteriors (0 vegades, 1 vegada, ...).
- Long.: si l'alternativa correcta és la de més contingut (longitud) apareix un 1.

Podria haver-hi més dades descriptives però aquesta informació ja serveix d'exemple per a comprovar si es produeixen relacions i efectes sistemàtics entre alguns aspectes que poden alterar el funcionament del test. La revisió d'aquesta taula és prèvia a l'ús del test (VDI pre), i és interessant per què anticipa problemes i ajuda a millorar la prova. En general, a no ser que el marc general de l'assignatura i el de l'examen indiquin el contrari, no haurien de produir-se associacions imprevistes. Té sentit que un autor solgui col·locar les alternatives correctes en una mateixa posició per a ítems amb un mateix enfoc de la taxonomia de Bloom? S'espera que els ítems considerats com a més difícils tinguin a veure amb una o una altre font d'informació? La dificultat d'una pregunta està relacionada amb el número de vegades que s'ha utilitzat amb anterioritat? Cada autor o equip responsable d'un examen hauria de constatar si es produeixen efectes sistemàtics de disseny no desitjats en el seu material. En aquest exemple hi ha poques verificacions prèvies d'aquesta mena. Una de les que es podria fer és comparar les co-

lumnes «Bloom» y «Alt.», i en aquest cas semblaria que les alternatives correctes dels ítems Bloom 1 tendeixen a ser les últimes (4 y 5), fet que indica que alguna cosa s'hauria de corregir.

El VDI adquireix realment valor quan s'incorporen noves dades objectives procedents de l'anàlisi psicomètrica (VDI post). A la part dreta de la Figura 4 es mostren quatre nous indicadors (columnes) per a cada pregunta de l'examen d'exemple. Sovint les màquines lectores d'exàmens i els sistemes d'avaluació online incorporen mòduls d'anàlisis bàsics que proporcionen dades d'aquest tipus. És suficient amb localitzar-los i incorporar-los al VDI. Les quatre que s'incorporen a la taula 4 són les següents:

- Dif: expressa la proporció d'encerts de la pregunta. Oscil.la entre 0 i 1, essent 1 un ítem que ningú no ha fallat (fàcil) i 0 un que ningú ha fallat (difícil).
- Discr: és l'índex de discriminació de la pregunta. Informa del grau en què aquesta és capaç de distingir entre els examinats més i menys preparats. Un ítem que discrimina tendeix a ser més encertat pels examinats més preparats i menys encertat pels que tenen menys preparació (permetria diferenciar o separar als que tenen més i menys capacitat en l'aspecte avaluat). Un ítem que no discrimina és encertat i fallat indistintament per examinats amb diferents nivells de capacitat. La discriminació és un indicador molt útil en el control de qualitat d'un test. Un examen amb preguntes que discriminen poc és un instrument d'avaluació defectuós. La discriminació pot calcular-se de moltes maneres. En aquest exemple s'ha obtingut a partir d'una correlació, i es consideren amb prou capacitat de correlació els ítems amb valors superiors a 0,3. Tornarem a tractar aquest índex en l'apartat 8.2.
- Solap: una opció més avançada d'anàlisi de les respostes consisteix a calcular la discriminació per a cada alternativa simulant que aquesta fos correcta. D'aquesta manera per a cada pregunta s'obtenen tantes discriminacions com alternatives. Si una o més alternatives incorrectes discriminen igual, o millor que la que és «oficialment» correcta (la considerada en la plantilla de correcció) llavors hi ha un problema de solapament. Potser per algun motiu l'alternativa considerada inicialment com a correcta no ho sigui, o que hi hagi més

d'una alternativa que funcioni com a correcta (alternatives confusores). En qualsevol cas aquest resultat constitueix una mala notícia doncs no està clar que el funcionament de l'element en el conjunt de l'examen sigui l'esperat. En auditories reals, prendre decisions sobre només uns pocs ítems que presentin solapament pot arribar a distorsionar significativament la llista de resultats de l'examen (aprovats que passen a suspesos, suspesos que passen a aprovat, ...). A la taula de l'exemple la columna «Solap» indica el nombre d'alternatives solapades en els ítems conflictius (l'ítem 4 de la prova és l'únic que presenta solapament en dues opcions de resposta).

- Atract: aquesta quarta dada indica si hi ha (1) o no hi ha (0) problemes d'homogeneïtat d'atractiu. En l'apartat sobre la RGI s'indicava que les alternatives incorrectes d'un ítem havien de tenir un atractiu similar. Per exemple, en una pregunta de cinc alternatives encertada per un 60 % d'examinats s'espera que el 40 % d'errors es distribueixi homogèniament entre les quatre alternatives incorrectes (10 % cadascuna). Aquesta condició és especialment important quan s'apliquen fórmules de penalització de la conjectura (apartat 7.2).

Aquestes noves característiques «objectives» (VDI post) poden comparar-se amb les seves equivalents «subjectives» o amb altres característiques previstes (VDI pre). En el cas de la figura 4, no sembla, en general, que hi hagi relació entre la dificultat que estimaven els autors (Dif. a VDI pre) i la dificultat final obtinguda a través de l'anàlisi de les respostes als ítems (Dif. a VDI post). Aquest és un resultat important ja que quan passa indica un desajustament entre les expectatives de disseny dels docents i la situació real dels examinats. Encara que el principal objectiu dels exàmens és obtenir evidència sobre el nivell dels estudiants, aquesta comparació entre les dificultats estimades i reals també ens proporciona informació sobre el nivell de «coneixement» que el docent que ha preparat la prova té dels seus estudiants. Es doncs, un efecte col·lateral, però positiu, de les avaluacions, i mol sovint proporciona resultats sorprenents per al docent. Ben gestionat, aquest tipus d'informació pot redundar en una millora de la docència.

Continuant amb la comparació, a primera vista sembla que els ítems que millor discriminen (Disc. A VDI post) corresponen a temes tractats principalment durant les classes (Tema en VDI pre). També s'observa

que els ítems més difícils amb menor proporció d'encerts (Dif en VDI post), són els que mai s'havien emprat abans en altres exàmens (Expo a VDI pre) i han estat elaborats principalment per l'autor 1 (Autor a VDI pre). Els problemes d'atractiu d'opcions (Atract en VDI post) estan més associats a l'autor 2 (Autor a VDI pre).

Cal tenir en compte que durant l'aplicació de l'examen poden aparèixer noves característiques rellevants a considerar en un VDI. Per exemple, sovint els examinats plantegen dubtes i fan consultes sobre una determinada pregunta. A la taula VDI es podria incloure una columna que reflecteixi el grau en què els ítems han estat objecte de consultes (quin tipus de dubte, quantes vegades ha estat consultat, ...). Més endavant, es pot utilitzar aquesta nova dada per contrastar-la amb indicadors de problemes (baixa discriminació, solapament...).

Les comparacions es podrien estendre però amb el que hem vist fins aquí ja és possible comprovar la utilitat d'aquest recurs. Proporciona elements de reflexió. Cada docent ha d'identificar relacions no esperades entre les característiques previstes per als ítems i les reals i, en cas de detectar-ne, ha de valorar la conveniència d'efectuar esmenes en els resultats de l'examen (per exemple, no considerar en el càlcul de la nota final una pregunta si té un comportament clarament no desitjat). Encara que mai constitueix una garantia total, una bona planificació de la prova (definició de la TEO i definició del VDI) minimitza el nombre de problemes posteriors. Sigui com sigui, els problemes detectats en un examen s'haurien de considerar en el futur. Només d'aquesta manera la tasca (repetitiva) d'elaborar exàmens, pròpia de qualsevol docent, podrà millorar.

5.2. Vector descriptor de la persona (VDP)

De la mateixa manera que convé acumular evidència sobre els ítems a l'VDI, també és recomanable fer-ho amb els alumnes en un vector que descrigui les característiques de la persona que considerem que poden tenir una influència en el resultat de la prova, o al menys, que pugui proporcionar-nos alguna llum sobre el mateix. Cada columna de l'VPD definirà un tipus d'informació rellevant referida a les persones. Aques-

ta informació dependrà de moltes circumstàncies (del tipus de curs, de la informació que es pugui recollir sobre els estudiants...).

Part de la informació sol procedir de la pròpia fitxa de l'estudiant, com per exemple, el grup de matrícula, però de vegades és possible que l'estudiant assisteixi a un grup diferent al que està matriculat. Aquesta és una informació més interessant que la «oficial». Informació com el grup d'assistència a les classes de teoria i de pràctiques (és un alumne d'un grup massiu de diürn o d'un minoritari de nocturn?), Si sol venir o no a classe (es pot registrar, per exemple, la assistència). La informació que descriu als alumnes sol ser difícil de recollir. No cal que una variable del VDP estigui completa per a tots els alumnes. Per exemple, de vegades, en una tutoria un alumne comenta que, a més de venir a classes, rep classes particulars. O que la matèria que impartim li «costa» particularment. Potser que només tinguem aquest tipus d'informació per a aquest alumne, però seria interessant disposar-ne incloent-la en el VDP perquè és possible que ens ajudi a entendre millor el seu resultat en la prova.

Una altra informació rellevant, associada a l'alumne però molt contextual, és la que fa referència a les condicions d'un examen en concret. Seria de molt interès saber el lloc exacte que va ocupar a l'examen (nombre de seient, fila i columna de l'aula ...). En algunes ocasions també si va lliurar ràpid o si va trigar a acabar l'examen. Si durant l'examen tenia dubtes i en quins ítems. Som conscients que aquest tipus d'informació és molt difícil de recollir, especialment quan es tenen grups amb molts estudiants, però no volem deixar d'indicar la utilitat que podria tenir disposar d'aquest tipus d'informació per contextualitzar els resultats referits a l'alumne que es descriuen en l'apartat 8.2.

6. CONFIGURACIÓ FORMAL DE L'EXAMEN I PROCÉS D'APLICACIÓ

Imaginem un examen format per 10 preguntes. Suposem que l'imprimim en un sol full de manera que a la primera pàgina, després d'una capçalera on consta el nom de l'assignatura, la data de l'avaluació, un espai per a la identificació de l'alumne i unes instruccions globals sobre la prova, queda espai per a vuit preguntes, i per tant hem d'imprimir les altres dos al dors de la pàgina. Si al final de la primera pàgina no informéssim que les preguntes continuen a l'altra cara del full, quants alumnes haurien deixat sense contestar les dues últimes preguntes simplement per no haver-les vist?

Aquest exemple pretén alertar sobre la importància dels detalls formals en confeccionar els exemplars d'examen. L'objectiu és minimitzar els aspectes formals que poden generar errors indesitjats. Cal recordar sempre que els alumnes han de poder contestar correctament o incorrectament als ítems únicament i exclusivament pel seu nivell de competència en la matèria avaluada, i no per altres factors irrellevants.

El full o el quadern d'examen ha de començar amb un espai reservat a la identificació de la prova i de l'alumne. A continuació s'han d'incloure unes instruccions on quedin clar, entre altres aspectes, el nombre de preguntes, de quina manera han d'indicar-les respostes o com es calcularà la puntuació (per exemple, si es penalitzen els errors o no, i de quina manera en cas afirmatiu). Els ítems es presentaran a continuació, i en cas que la impressió dels mateixos ocupi més d'una pàgina, hauria d'indicar-al final de les mateixes una llegenda del tipus «Continua a la pàgina següent» o simplement «Continua». De la mateixa manera, és recomanable indicar al final «Fi de la prova».

La forma en què es presenten les alternatives de resposta pot complicar de manera innecessària la comprensió de l'ítem. Vegem un exemple d'un mateix ítem amb les opcions de resposta presentades de maneres diferents:

Presentació 1
Quina de les següents capitals és d'un país asiàtic?

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

Presentació 2
Quina de les següents capitals és d'un país asiàtic?

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

Presentació 3
Quina de les següents capitals és d'un país asiàtic?

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

Presentació 4
Quina de les següents capitals és d'un país asiàtic?

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

En general es considera que l'última presentació, amb les opcions de resposta de manera vertical, és la que destaca amb més claredat les opcions, i per tant és la forma de presentació recomanada.

Si la correcció de les respostes es realitzar de forma mecànica, caldrà un full de respostes específic perquè pugui ser interpretat per una lectora òptica. En cas contrari les respostes es poden marcar en cadascuna de les opcions o es pot confeccionar un full de resposta ad hoc que s'afegirà al final de la prova. En tot cas, les instruccions de la prova han de deixar clar com i a on s'han de marcar les opcions (amb una creu, ombrejant una casella ...) i com es poden anul·lar opcions marcades inicialment com a correctes.

Un altre aspecte formal que cal cuidar és el tipus i mida de lletra. Els tipus poden classificar-se en base de diversos criteris, un de rellevant per facilitar la lectura és el de la rematada a la base. N'hi ha amb rematada (tipus Serif), com Times Roman, i sense rematada (Sans Serif)

com l'Arial. La rematada a la base de les lletres proporciona una certa continuïtat o lligam en configurar paraules, a més que distingeixen perfectament majúscules o minúscules (per exemple, l, L), cosa que no sempre passa amb els tipus Sans Serif. Així doncs, els tipus Serif solen facilitar la lectura i per això solen ser els elegits per les editorials per a la publicació de llibres impresos. La presentació dels ítems (enunciats i opcions de resposta). Tot i això, quan el que es vol és cridar l'atenció del lector (com en els avisos al final de pàgina) es recomanen els tipus Sans Serif per la seva contundència en especial quan es tracta de majúscules o negretes. En qualsevol cas es preferible evitar les tintes de colors (prop d'un de cada 10 homes és daltònic).

Pel que fa a la grandària de la lletra, el mínim aconsellat és de 12 punts (encara que això depèn també del tipus de lletra emprada). Més que la mida de lletra, el que afecta la llegibilitat del text és l'espaiat interlineal. En aquest sentit, el que es recomana és que l'espaiat sigui almenys de la mateixa mida que la lletra, encara que la lectura del text s'optimitza quan l'interlineat és entre el 20 % i el 30 % la mida de la lletra (si s'utilitza una lletra de mida 12, un interlineat òptim hauria de ser d'entre 14 i 15 punts). Tot això sol augmentar el nombre de pàgines necessàries per encabir totes les preguntes de l'examen. En aquest cas, si el format dels ítems ho permetés, els ítems podrien distribuir-se en dues columnes per pàgina, amb una separació evident entre ambdues. Les instruccions que es donin als alumnes en el moment de començar l'examen no es poden improvisar. Amb anterioritat s'han de preveure tots aquells aspectes que tinguin una conseqüència rellevant (temps d'administració, resposta a dubtes...). La persona encarregada d'administrar la prova ha d'exposar les instruccions de la manera més clara i neutra possible. Si la mateixa prova s'administra simultàniament en sales diferents, les persones encarregades d'informar s'han d'assegurar que ofereixen una informació comparable en contingut, forma i temps. També han de vetllar per que hi hagi un mínim de condicions de confort (temperatura, il·luminació...).

Sovint passa que, un cop començada la prova, cal interrompre-la per proporcionar dades addicionals sobre ella (per exemple, un aclariment sobre algun ítem o sobre algun detall d'un exercici). No es convenient abusar d'aquest recurs perquè acostuma a desconcentrar als alumnes

i a més, no se sol garantir que la informació que s'afegeix arribi a tots ells (molts estan tan preocupats per contestar les preguntes que no s'adonen del que s'està dient). Tanmateix, en cas de necessitat, convé concentrar totes les informacions de manera que el nombre d'interrupcions sigui mínim. Si les interrupcions tenen a veure amb dubtes dels examinats, cal recordar la necessitat d'incloure aquestes consultes a la taula VDI per contextualitzar millor la posterior anàlisi dels ítems.

La majoria dels docents prenen algun tipus de precaució relativa al frau en les respostes. La possibilitat que algun alumne vegi i copií les respostes d'un altre és la més freqüent. Perquè es pugui produir una còpia, hi ha d'haver una font d'informació (FI voluntària o involuntària) i una font receptora (FR sempre voluntària), i per tant una prevenció bàsica consisteix a separar aquestes dues fonts. És sabut que la proximitat lateral entre l'informant i el receptor afavoreix la còpia, però altres disposicions, com la coneguda en V que es mostra a la figura 5, també. A la figura es veu que la informació procedent de la font d'informació situada al centre de la primera fila pot transmetre de manera obliqua fins arribar a l'última fila. Una manera de contrarestar aquest possible efecte és confeccionar proves equivalents (models) o si més no, permutacions de les preguntes i de les opcions de respostes, i repartir-los de manera que no coincideixin els models ni lateral ni diagonalment.

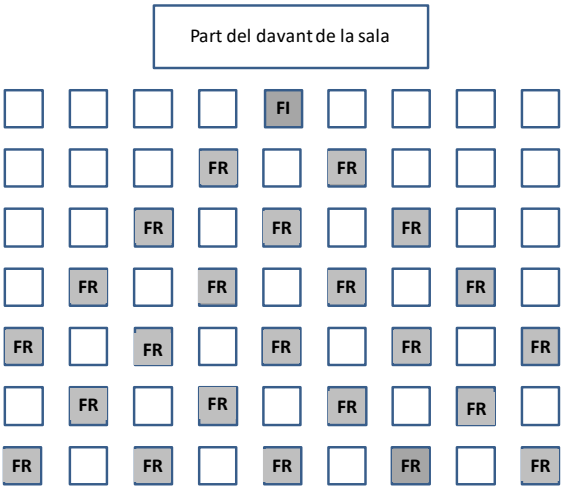


Figura 5. Disposició en forma de V que afavoreix la còpia

7. SISTEMA DE PUNTUACIÓ

Una decisió molt important a l'hora de crear un test és la que fa referència a com es puntuarà als examinats. La tendència predominant és que totes les preguntes aportin la mateixa puntuació (respostes incorrectes = 0, resposta correcta = 1) i la suma del conjunt sigui la puntuació final. Tot i això, hi ha altres opcions que bàsicament consisteixen en ponderar i penalitzar o corregir la conjectura.

7.1. Ponderació de les respostes

Consisteix a donar més valor a uns encerts que d'altres. Això pot fer-se sota el criteri del docent / autor que com a expert en la matèria decideix el pes de cada pregunta. En aquests casos convé comparar les dades VDI subjectives prèvies (importància, dificultat, ...) amb les posteriors més objectives (apartat 5.1) i confirmar que realment concorden. Sense aquesta constatació serà dubtós que la ponderació reflecteixi el valor/funcionament real de la pregunta.

Un altre tipus de ponderació consisteix en donar un valor diferent a l'alternativa escollida o a una combinació d'alternatives escollides (encara que no és molt recomanable, de vegades s'admet més d'una opció com a correcta). Aquí també interessa comprovar que l'anàlisi dels ítems (indicadors) confirma la ponderació. Existeixen ponderacions que combinen l'encert amb el nivell de certesa que manifesta l'examinat en cada resposta. En aquests exàmens els ítems aporten dues dades, la resposta i el percentatge o grau de seguretat de l'alumne en donar-la (en una escala graduada arbitrària). Un encert amb una seguretat en la resposta del 70 % pesarà menys que un altre encert amb una seguretat del 90%. Aquesta forma de respondre, i de ponderar, depèn en gran mesura de factors personals de l'examinats (autoconfiança, assertivitat...) que poden interferir en les respostes i per tant en les notes finals.

En altres casos la ponderació afecta directament la puntuació global i no als ítems. Es tracta de ponderacions basades en la coherència del

patró de resposta de l'examinat (veure més endavant els patrons atípics de resposta).

7.2. Penalitzar els errors

Consisteix, a criteri de l'autor, en restar els errors, o un fracció d'ells, dels encerts. En aquest enfocament no hi ha cap base matemàtica o fórmula que justifiqui la penalització. Bàsicament és una manera dràstica de treure incentiu a la tendència a respondre quan no està clara la resposta.

Els examinats que desconeixen la resposta a una pregunta, tendeixen a activar recursos alternatius per a contestar-la (intentar endevinar-la, descartar opcions poc plausibles, etc.). La presència de conjectura en un examen el pot invalidar ja que distorsiona el càlcul de les puntuacions (no hi hauria garanties que la nota d'un alumne sigui un fidel reflex del seu nivell de coneixements de la matèria avaluada).

Tant la penalització com la correcció de la conjectura, que veurem més endavant, comparteixen un efecte dissuasiu que involucra altres factors personals. Un examen sol comportar pressió vistes les seves conseqüències, generalment vinculants per al futur dels examinats (superar l'assignatura, passar o no el curs, tenir o no de recuperar la matèria...). Tot això provoca emocions, pensaments negatius, etc. que condicionen d'alguna manera la manera de respondre (certesa, acceptació de risc, auto-concepte...) i determinen part del resultat de l'examinat (una part que no és la que es desitja que quedi reflectida a la nota, ja que la nota hauria de poder interpretar únicament i exclusivament pel nivell de coneixements avaluats).

Les estratègies anteriors no són les úniques per a contrarestar la conjectura. Una alternativa consisteix a augmentar el nombre d'encerts (puntuació de tall) per a la presa de decisions finals (apte/no apte, barem de qualificacions...) de manera que el pes de la correcció ja no resideix en l'examinat sinó en l'exigència de la prova. En aquests casos se sol augmentar el barem de puntuació (pujar el «llistó») per compensar el possible efecte de la conjectura.

Altres opcions es basen en mantenir la puntuació de tall però augmentant la dificultat de les preguntes o simplement augmentar el nombre d'alternatives. Finalment, també hi ha estratègies analíticament més complexes basades en models psicomètrics que tracten l'efecte de la conjectura de manera diferenciada per a cada ítem. Aquesta és la via més recent i sofisticada ja que en calcular la puntuació d'un examinat es té en compte l'efecte diferenciat de la conjectura en cadascun dels ítems que ha respost. D'aquesta manera, diferents examinats amb diferents patrons de resposta (encerts i errors en diferents ítems) tindran unes puntuacions distintives en funció del seu patró.

Llevat de casos a gran escala aquest tractament de la conjectura encara és inusual en exàmens acadèmics ja que requereix grandàries de mostra importants.

Per fer els càlculs s'utilitzen mètodes analítics i programari especialitzat de la ja esmentada teoria de resposta d'ítem, especialment el model d'anàlisi de tres paràmetres.

7.3. Corregir la puntuació

Des de fa anys el recurs més popular és la «correcció» (reducció) de la puntuació de l'examinat a partir d'una suposada justificació matemàtica (veure fórmula en l'annex). Igual que ocorre amb la penalització, aquesta opció comporta uns riscos i uns efectes col·laterals que molts avaluadors assumeixen sense saber-ho.

La majoria de fórmules de correcció permeten estimar la quantitat d'encerts per conjectura i restar-la del total d'encerts obtinguts pel examinat. Amb això es «corregeix» la puntuació obtinguda permetent tornar a calcular, per a cada examinat, els seus encerts «reals» lliures d'efectes estranys. Moltes aplicacions de correcció de test i exàmens incorporades a lectores òptiques i plataformes online solen oferir el càlcul de l'annex 1 entre els seus outputs. Aparentment la fórmula soluciona de manera senzilla un problema que semblaria molt més difícil d'abordar. Tot i això, és interessant recapitular de nou les condicions d'aplicació en què es basa:

- Condició 1: tots els ítems de l'examen han de tenir el mateix nombre k d'alternatives.
- Condició 2: només hi ha encerts i errors, no omissions (no es permeten "respostes en blanc").
- Condició 3: tots els errors dels examinats es deuen al fet que intenten encertar conjecturant i no ho aconsegueixen.
- Condició 4: tots els examinats, quan s'enfronten a preguntes que no dominen, tendeixen a respondre conjecturant i les fallades són deguts a no haver-ho aconseguit.
- Condició 5: s'assumeix que la possibilitat d'escollir l'alternativa correcta és equiprobable entre les alternatives. Dit d'una altra manera, totes tenen el mateix atractiu (ja descrit en l'apartat 3).

A efectes aplicats aquesta llista planteja un primer dubte general, es compleixen les condicions en els nostres exàmens? Del que se'n deriven sis:

- Totes les preguntes tenen igual nombre d'alternatives? Hi ha exàmens en què varien. Si és així no seria lícit aplicar aquesta correcció.
- Acceptem que en una fracció de les preguntes en què els examinats provin sort la tindran i encerteran? Això implica que els examinats comparteixen un mateix estil de resposta.
- Encara que tots els ítems el mateix nombre d'alternatives (condició 1), aquestes tenen un atractiu similar en cada pregunta (condició 5)?
- Es a dir, pot assumir-se que el nombre d'encerts per conjectura és proporcional al nombre d'intents contestats per conjectura (veure annex: C = C1/k)? En la majoria de casos això no es verifica. A la pràctica, en demanar als autors d'un examen dades descriptives de les preguntes (VDI pre) molts admeten desequilibris en l'interès que susciten les diferents alternatives entre els seus alumnes. Al comparar després això amb els resultats de l'anàlisi dels ítems (VDI post) se sol confirmar l'existència d'un problema de disseny.
- Es acceptable que quan els examinats no tenen prou capacitat per respondre les preguntes tots van a provar sort (condicions 3 i 4)?
- En l'examen hem insistit perquè no hi hagi preguntes sense contestar (condició 2)? En la majoria d'exàmens hi ha omissions i més si els errors penalitzen.

En aquest últim punt hi ha una paradoxa interessant ja que d'una banda aquesta correcció potencia la conjectura (respondre tot, no ometre res) mentre que després penalitza els errors. Sigui com sigui, els examinats han d'escollir una alternativa, no importa com la triïn. Dit d'una altra manera, es promou la conjectura per després penalitzar-la assumint que ningú falla per desconeixement sinó per «mala sort» en escollir. Enfront de tot això, en la pràctica real s'observen escenaris molt diferents. Els examinats temen ser penalitzats i eviten respondre les preguntes molt difícils. En la majoria de situacions els estudiants responen les preguntes accessibles per al seu nivell deixant sense contestar les restants. Si fallen en alguna resposta no és sempre per «mala sort» sinó per què han escollit deliberadament una resposta incorrecta pensant que era correcta (han contestat en base del que sabien, encara que el que sabien no era correcte).

Un altre aspecte a tenir en compte és que aquest tipus de correccions no distingeixen entre examinats amb més o menys nivell (capacitat) mentre que la tendència a respondre per conjectura si que sol variar segons aquest factor. En moltes anàlisis de respostes a test reals els examinats del terç inferior de puntuacions (els que tenen menys capacitat sobre els coneixements avaluats) són els que maximitzen les respostes per conjectura. Tenen menys a perdre i més a guanyar. Per contra el comportament del terç superior és molt més prudent. Es tracta dels examinats amb més coneixement que es plantegen les possibles penalitzacions de manera més conservadora. Amb això es constata un efecte habitual; les correccions tendeixen a perjudicar especialment als examinats de menys capacitat. A la vista d'aquests factors encara es fa més difícil admetre que tots els examinats funcionin estadísticament com un sol individu modal. A més, molts exàmens i enquestes barregen ítems amb diferent nombre d'alternatives de resposta, això afecta la probabilitat d'encertar ja que el valor k varia incomplint clarament la condició 1 (per aquest motiu no és recomanable variar el nombre d'alternatives en els exàmens). Una cosa semblant passa quan k és constant estructuralment però no funcionalment. Suposem un ítem de cinc alternatives com tots els del test. Imaginem que per algun motiu el disseny de l'ítem fa que ningú, o gairebé ningú, esculli un parell de les seves alternatives. La forma com estan redactades i/o el contingut que tracten manca d'atractiu i els examinats tenen molt clar que aquestes

dues alternatives no poden ser correctes. Malgrat que aquest ítem estructuralment té cinc alternatives en realitat sols tres funcionen com a tals. És evident que aquesta situació facilita l'elecció de la resposta correcta (quan no es coneix l'opció correcta, resulta més fàcil encertar entre tres opcions que entre cinc). Quan això es repeteix en diversos ítems l'efecte de la conjectura augmenta i amb això la distorsió de les puntuacions dels examinats. En molts exàmens el valor funcional de k és menor que l'estructural. Només cal comprovar les freqüències d'elecció de les alternatives i constatar que l'atractiu (tendència a triar-la) varia molt. Hi ha proves amb ítems de 4 alternatives estructurals però que a la pràctica funcionen com Veritable-Fals ja que només dues alternatives de cada pregunta atreuen realment l'atenció quedant les altres dues descartades per endavant per la majoria examinats.

Tot l'anterior pot dur a reflexionar sobre l'aplicació d'aquest tipus de correccions ja que només són lícites sota condicions molt concretes i controlades. Paradoxalment la seva popularitat ha anat per davant del coneixement dels seus fonaments. A més, hi ha altres variants, com les que accepten l'existència d'omissions, però que també afegeixen noves condicions difícils de complir. Per la seva extensió i especificitat no les abordarem en aquesta obra.

En general totes aquestes expressions van ser proposades en la primera meitat del segle XX pensant en com corregir de manera objectiva (justificada analíticament) l'efecte de la conjectura en unes situacions experimentals molt concretes. Originalment s'empraven amb test denominats de velocitat de 5 o més alternatives de resposta i longitud superior a 20 ítems. En aquest tipus de proves la puntuació de l'examinat ve determinada principalment per la seva velocitat de resposta. En aquestes proves els ítems no solen ser molt difícils ja que l'important és comprovar quants encerts s'aconsegueixen en un temps breu. Aquest escenari no és generalitzable per a la majoria d'exàmens actuals ja sigui en paper o online. D'altra banda, si ja en el seu moment es van plantejar com una estimació (correcció) arriscada de la puntuació de l'examinat més difícil és encara avui dia, a la vista de les condicions, acceptar la seva validesa en escenaris complexos.

7.4. Correcció i número d'alternatives

Aquestes fórmules de correcció també han servit de base per justificar alguns tòpics com el nombre idoni d'alternatives que hauria de tenir un test.

Evidentment com més alternatives, i millor estiguin creades, es reduirà l'efecte de la conjectura. Tanmateix, això exigeix un sobreesforç de disseny que no és sostenible en la majoria d'avaluacions reals.

Des de fa anys han aparegut propostes defensant «números ideals» d'alternatives a l'hora de crear ítems. Una prové de la fórmula de correcció abans vista i porta a considerar el cinc com el mínim nombre d'alternatives idoni i realista. La figura 6 exposa aquesta proposta. El gràfic mostra com seria la puntuació corregida N (eix ordenades) per al cas de nou examinats que hagin obtingut respectivament 10, 20, 30, 40, 50, 60, 70, 80 i 90 encerts en total (A) en un examen de 100 preguntes que podria variar de tres a set alternatives (abscisses).

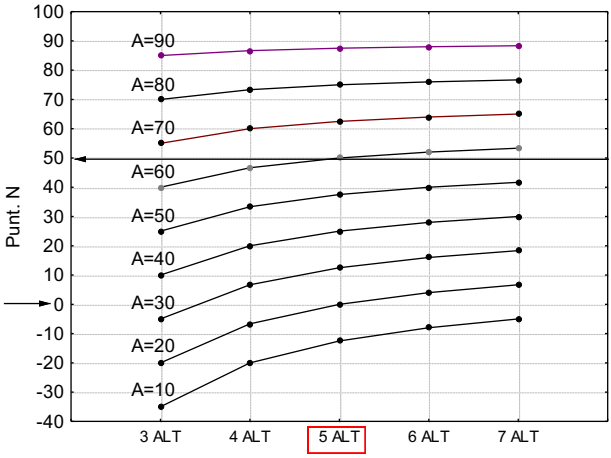


Figura 6. Relació entre la correcció per conjectura aplicada a la nota i número de alternatives de resposta dels ítems

Cada perfil representa la puntuació corregida que tindria un examinat en funció del nombre d'alternatives del suposat test. La fletxa llarga horitzontal assenyala la puntuació 50 de tall de l'examen (apte/no apte).

A la part inferior hi ha el perfil de l'examinat que ha obtingut només 10 punts. S'observa que tant si el test fos de tres alternatives com de set la correcció sempre rebaixa N, que d'altra banda varia molt entre tres i set alternatives i sempre és negativa (la fletxa petita indica el valor 0 de puntuació corregida).

A l'extrem superior hi ha el perfil de l'examinat amb 90 encerts (A = 90). En aquest cas la variació de la correcció segons el nombre d'alternatives és mínima. Es manté gairebé horitzontal i sembla que el nombre d'alternatives no incideix massa en la correcció. Per als casos intermedis la lectura dels perfils és similar. En general s'observa una tendència asimptòtica a mesura que el nombre d'alternatives augmenta però que no obstant això varia en funció del total d'encerts. La variació en funció del nombre d'alternatives disminueix a mesura que la puntuació total és més gran.

Aquest tipus de gràfics ha portat a considerar el valor 5, i superiors, com idonis per contrarestar la conjectura. De fet molts models de fulls de resposta admeten un màxim d'entre quatre i sis alternatives. No obstant això, aquestes consideracions sobre el nombre òptim d'alternatives s'han de valorar des de l'escenari, poc realista, que totes les alternatives incorrectes tenen característiques equivalents (són comparables en qualitat). El que sol passar a la pràctica, però, és que la qualitat dels distractors és diferent, en part a per com ha estat el procés de creació de la pregunta. Generalment, la persona que redacta una pregunta té molt clar quina és la resposta correcta i molt probablement també una de les respostes incorrectes. No sol costar massa trobar una segona opció incorrecta, però la tercera ja és més complicat (de vegades massa costosa) i així successivament. És per això que els ítems amb tres opcions de resposta acostumen a «funcionar»millor (des d'un punt de vista mètric) que els que tenen un nombre més gran. Cal observar que aquest condicionant pràctic i freqüent no és molt compatible amb la recomanació psicomètrica d'augmentar el nombre d'opcions (per augmentar la fiabilitat, per minimitzar la conjectura...).

Dos últims aspectes relacionats amb la puntuació del test són la seva distribució i el punt de tall. En la majoria de proves psicològiques (normatives) interessa que les puntuacions es distribueixin seguint la corba normal. L'anàlisi psicomètrica de les qualitats d'un test sol assumir

aquesta forma com una condició important. Per contra, els exàmens no responen a aquest requeriment. De fet interessa que no el compleixin

La figura 7 representa la distribució esperada per a un test de norma de grup (TNG) i la d'un test referit al criteri (TRC) ja descrits en l'apartat 1.2. Si l'avaluació d'una assignatura es fa mitjançant un examen s'espera que la distribució de puntuacions tendeixi cap a la part alta de puntuació. Si el procés d'EA s'ha desenvolupat convenientment el rendiment dels estudiants ha de mostrar una desviació a la dreta.

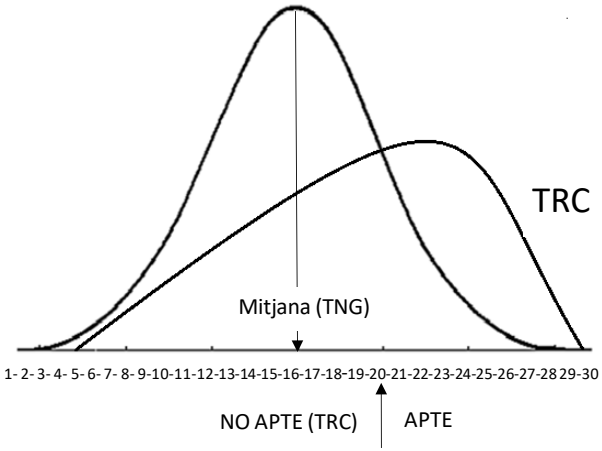


Figura 7. Distribucions de les puntuacions a dos test, un interpretat segons una norma de grup (TNG) i l'altre en base a un criteri (TRC)

L'altra qüestió és el punt de tall per a la presa de decisions importants com a apte/no apte, aprovat/suspens, etc. Hi ha la tendència a confondre la meitat d'encerts amb el 5 com a puntuació de tall més associada a l'aprovat / apte. En crear una prova cal haver previst quina serà aquesta puntuació (afecta a la dificultat dels ítems) i si ha de correspondre amb aquest 50% d'encerts o a un valor superior. De fet encertar la meitat pot semblar contradictori amb la pràctica professional i en molts exàmens i acreditacions el nivell d'exigència és molt superior (què diríem d'un metge que només encerta la meitat dels diagnòstics o d'un enginyer que rep crítiques per errors a la meitat dels dispositius que dissenya?). En molts entorns la puntuació de tall oscil·la al voltant del 70 % de domini de l'examen i aquesta dada s'ha de tenir en compte a l'hora de dissenyar la prova.

8. AUDITORIA QUANTITATIVA

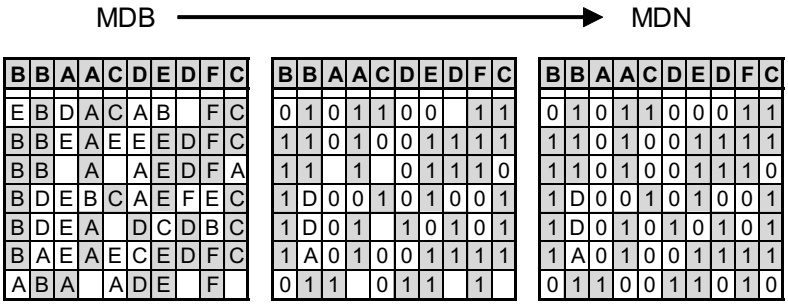
Consisteix en l'anàlisi psicomètrica del funcionament tant dels ítems com del test en el seu conjunt. Per això s'obtenen indicadors numèrics i gràfics de cadascuna de les alternatives dels ítems incloent l'omissió com una enèsima opció complementària. També es valora l'adequació de la plantilla i la coherència de les respostes. Durant el procés es comprova si cada alternativa compleix la seva funció i en quin grau aporten valor al test. Quan es detecta alguna anomalia en un ítem o a la plantilla, aquesta es contrasta amb el VDI pre i amb l'autor per tal de considerar si és o no adequat incloure la pregunta afectada en el còmput de la puntuació total dels examinats. El procés d'auditoria varia en el nivell de detall. Alguns models de lectores i LMS ofereixen anàlisis preliminars dels test que faciliten una primera valoració del material. Per anàlisis més en profunditat cal programari especialitzat o accedir a un servei d'anàlisi de test.

8.1 Dades necessàries

Moltes auditories queden limitades des d'un principi per la manca de dades adequades. Hi ha la creença que les dades a processar són les dades netes, és a dir les respostes ja corregides dels examinats (10101011...). En realitat això limita les opcions d'anàlisi ja que les realment important són les dades brutes (ABADEAD...). Les dades netes procedeixen de les brutes després d'aplicar la plantilla de correcció i això és contradictori amb el principi d'auditoria ja que, fins que l'anàlisi demostri el contrari, la plantilla es manté en quarantena i és tractada com dubtosa.

La figura 8 mostra tres versions de dades per a un mateix test de 10 ítems i sis alternatives. La taula de l'esquerra correspon a la matriu de dades brutes (MDB) amb les respostes marcades pels examinats i les omissions (la fila superior és la plantilla de correcció). La taula central correspon a la matriu de dades netes (MDN) un cop corregides les respostes amb la plantilla i respectant les omissions. La tercera taula és la mateixa que la central però convertint aquestes omissions en 0. Les

tres taules representen les tres situacions possibles, la primera és indispensable per a una auditoria completa i la tercera la menys útil.



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

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

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

Figura 8. Tres versions d'una matriu de dades per a un mateix test de 10 ítems i 6 alternatives: una de dades brutes (MDB amb respostes originals, A, B, C ...) i dos de dades netes (MDN), una codificant els encerts (1) i errors (0) i l'altra també les omissions (0)

8.2 Indicadors

Disposant de la taula MDB és possible obtenir una gran varietat d'indicadors tant globals com dels ítems encara que per la seva extensió no serà possible tractar-los tots aquí en detall. A continuació hi ha una descripció dels més comuns:

Globals

- Distribució de les puntuacions: correspon al gràfic de la figura 6.
- Coeficient alfa: és un tipus de coeficient de fiabilitat (Alpha de Cronbach) anomenat de Consistència Interna que oscil·la entre 0 i 1 sent recomanable acceptar valors superiors a 0,8.
- Error de mesura (SEM): indica la imprecisió dels resultats del test. Com a regla orientativa n'hi ha prou amb sumar i restar dues vegades el valor de SEM (± 2 SEM) a una puntuació del test (examinat) per estimar l'interval d'error en què aquesta oscil·la amb un nivell de confiança del 95 % . Com més gran sigui aquest interval menys precís és el test.
- Comparació entre omissions i puntuació: consisteix en un gràfic de dispersió que mostra la relació, i la correlació, entre les puntuacions totals dels examinats i la quantitat de respostes que omet cadascun.

D'ítem

- Dificultat d'ítem és la proporció d'examinats que seleccionen l'alternativa correcta. Oscil·la entre 0 i 1. Un ítem «fàcil» ofereix valors propers a 1 i un «difícil» s'aproxima a 0.
- Variància: és un indicador de dispersió el valor mínim és 0 i valor màxim depèn del nombre d'alternatives (per exemple, per ítems dicotòmics oscil·la entre 0 i 0,25). La variància és necessària perquè un ítem discrimini però no garanteix que ho faci.
- Discriminació d'ítem: habitualment es calcula mitjançant la correlació entre les puntuacions de l'ítem amb les puntuacions totals. Pot oscil·lar entre -1 i +1 però només és convenient acceptar ítems amb valors positius superiors a 0,3. La discriminació informa del funcionament de l'ítem i si mostra un valor negatiu pot ser que l'ítem estigui funcionant al revés del que s'esperava (el encerten els examinats menys preparats i el fallen els més preparats).
- Discriminació corregida: es calcula de la mateixa manera que l'anterior, però sense incloure les dades (respostes) de l'ítem analitzat en la puntuació total. El resultat tendeix a donar valors menors que la discriminació sense corregir i és aconsellable en proves curtes. A mesura que un test té més ítems ambdós tipus de discriminacions tendeixen a coincidir.
- Eleccions de les alternatives: és la proporció d'examinats que escull cadascuna de les alternatives. La dificultat de l'ítem coincideix amb la proporció de l'alternativa correcta.
- Omissió: és la proporció de subjectes que no han respost a l'ítem.
- Discriminació de les alternatives: és la discriminació (amb o sense correcció) que tindria l'ítem considerant que una alternativa errònia fos certa. Si l'ítem està ben construït la discriminació de l'alternativa correcta ha de produir un valor més gran que la resta.
- Discriminació de l'omissió: és la discriminació (amb o sense correcció) que tindria l'ítem considerant que l'omissió fos la resposta correcta.
- Igualtat d'atractiu: és una prova d'ajust entre la proporció d'eleccions de cada alternativa errònia pel que fa a la proporció esperada en cas que totes tinguessin un atractiu similar. Si diversos ítems mostren desajustos s'invalida la condició 1 de l'apartat 7.3. Encara que aparentment aquests ítems tinguin k alternatives en realitat funcionen amb menys i s'incrementa l'efecte de la conjectura.

- Anàlisi de la plantilla: quan una alternativa incorrecta discrimina més que la correcta pot indicar un conflicte entre alternatives o un error d'assignació a la plantilla. En qualsevol cas cal revisar el contingut de l'ítem.

Part d'aquests indicadors es poden verificar visualment a partir d'un sol gràfic de perfils de resposta d'un ítem com el de la figura 9.

El gràfic correspon a un ítem qualsevol d'un test de quatre alternatives (A, B, C i D). L'eix d'abscisses representa el rang en que oscil·la la puntuació total del test. En aquest exemple les puntuacions s'agrupen en 10 nivells o intervals que van de menys a més puntuació o capacitat (N1 a N10). L'eix d'ordenades expressa el percentatge d'examinats de cada nivell de puntuació de les abscisses que escullen una determinada alternativa de resposta. Cada perfil correspon a una alternativa i indica el percentatge d'examinats de cada nivell que l'ha escollit. També s'inclou el perfil de l'omissió (X).

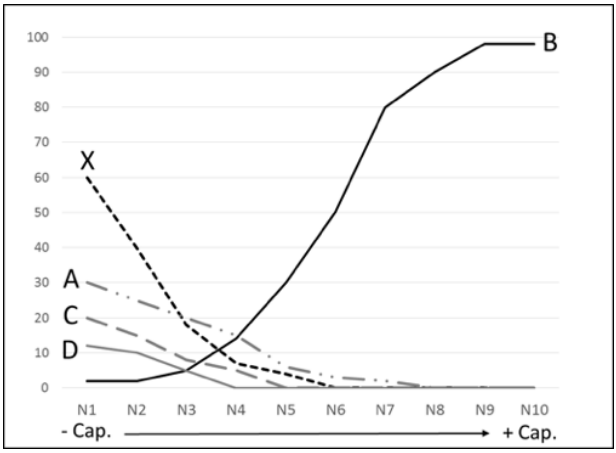


Figura 9. Corbes de resposta de les alternatives d'un ítem. L'opció correcta (B) és la creixent

El patró desitjable és que hi hagi un sol perfil creixent, el de l'alternativa correcta, i decreixents per a la resta de perfils (respostes incorrectes i omissió). Dit d'una altra manera, la tendència a escollir l'alternativa correcta ha d'augmentar gradualment cap a la dreta a mesura que els

examinats tenen més capacitat (puntuació total). Pel mateix motiu, l'elecció de les alternatives incorrectes i de l'omissió ha de decréixer.

A la figura 9 l'alternativa correcta és la B i mostra el patró esperat. El mateix succeeix amb A, C, D i X (els examinats de menor capacitat són els que més responen incorrectament o ometen la resposta).

Diverses aplicacions psicomètriques ofereixen gràfics similars al d'aquest exemple i són molt útils a l'hora de valorar l'adequació dels ítems.

Les directrius bàsiques d'interpretació per al perfil de **l'alternativa correcta** són:

- Quan més ràpidament creixi el perfil més discrimina l'ítem.
- Si ho fa suaument al llarg de tot l'eix d'abscisses pot ser que discrimini però en menor grau.
- Si el perfil està desplaçat cap a la dreta es tracta d'un ítem difícil, si ho està a l'esquerra és un ítem fàcil.
- Si el perfil és pla al llarg de l'eix d'abscisses o decreixent indicarà que l'ítem no discrimina o ho fa al revés. És convenient revisar-lo.
- Si la part esquerra del perfil ja és inicialment elevada i paral·lela a l'eix d'abscisses (asimptòtica) per després créixer a mesura que va cap a la dreta, es tracta d'un ítem afectat per respostes conjecturades (endevinació, atzar...).

Per a les **alternatives incorrectes**:

- Regla general: han de decréixer d'esquerra a dreta.
- Si mostra un perfil creixent (com el que correspondria a una correcta) i el de la correcta és decreixent o creix menys que el de l'errònia, és important revisar la plantilla de correcció per confirmar quina alternativa és realment la correcta (potser hi ha algun error a la plantilla).
- Si tots els perfils es mantenen plans i estables pot indicar que la majoria de respostes es fan per conjectura (endevinació, descart, atzar...). En tot cas, l'ítem no funciona bé i no s'hauria d'utilitzar per avaluar els examinats.

Per a **l'omissió**:

- Com a regla general ha de mostrar una tendència decreixent. No obstant això, per efectuar una interpretació adequada cal considerar si els errors penalitzen o no i, si ho fan, en quin grau.
- Si es manté estable (pla) al llarg de tot l'eix d'abscisses indica que els subjectes amb més puntuació total també tendeixen a deixar en blanc aquesta resposta. No és un resultat esperat.
- Si el perfil és creixent cal revisar l'ítem ja que contràriament al que s'esperava els subjectes amb més puntuació tendeixen a ometre la resposta. Si això coincideix amb que el perfil de l'especificada com a correcta és decreixent, o creix menys que el d'omissió, cal verificar la clau de correcció i confirmar quin és realment l'alternativa correcta (els subjectes amb major puntuació tendeixen a ometre aquesta pregunta i això pot evidenciar un error de disseny que només ells detecten). Aquest és un dels pitjors casos possibles.

D'examinat

Imaginem dos estudiants A i B que empaten amb 12 encerts en una prova de 20 preguntes. A efectes oficials se'ls considera amb el mateix nivell de capacitat ja que la seva puntuació és una simple suma sense entrar a mirar en quins ítems han aconseguit els encerts. Tanmateix, si ordenem les 20 preguntes de l'examen de la més fàcil a la més difícil (en funció de la proporció d'encerts de tot el grup examinat) observem que A ha aconseguit els seus 12 punts encertant i fallant ítems fàcils i difícils indistintament. Dit d'una altra manera, ha encertat preguntes molt difícils però també ha fallat altres fàcils (poc esperat). Pel que fa a B ha aconseguit els 12 punts de manera més coherent, ha encertat els ítems de poca i mitjana dificultat fins que ha començat a fallar o ometre les respostes i així fins a la pregunta més difícil.

Examinat A: fàcil 01011010111010110101 difícil
Examinat B: fàcil 11111111111100000000 difícil

A efectes d'avaluació les preguntes aquí són, tots dos examinats tenen realment la mateixa capacitat? Estan empatats? Observant la sèrie de respostes sembla que B té un patró de resposta més coherent que A i que per tant la seva puntuació global hauria de reflectir d'alguna ma-

nera aquesta qualitat (quedar millor ponderada). Si l'exemple fos un examen de càlcul, A hauria encertat arrels quadrades i fallat en simples sumes la qual cosa no contribueix molt a saber quina és la seva veritable capacitat o nivell (es posa en dubte la validesa de les inferències sobre el nivell de coneixements que es realitzen a partir de les notes obtingudes). Aquest tipus de reflexions fan dubtar de la validesa de la puntuació que atorguem a l'examinat A, o millor dit, del que inferim a partir d'aquesta nota (haurà contestat a l'atzar? Si és així, la nota que ha aconseguit descriu correctament el seu nivell de capacitat?).

Hi ha índexs psicomètrics que permeten detectar patrons incoherents o atípics de resposta (PAR). Però no pensem que tots els PAR són deguts a «trampes» en l'examen per aconseguir més nota. En algunes ocasions, alumnes d'alt nivell contesten malament preguntes molt fàcils (per a ells són tan fàcils que no creuen que es puguin preguntar en l'examen i per això «li busquen els tres peus al gat»). O li dediquen tanta atenció als aspectes difícils de la matèria que deixen de banda els més senzills. En aquests casos, al contestar de forma incoherent amb el seu nivell, aquests estudiants tindran una nota més baixa del que els correspondria per la seva capacitat. L'auditoria dels patrons de resposta pot identificar aquests i altres casos. No pot, però, explicar les seves raons. És aquí quan la combinació dels resultats quantitatius amb la informació recollida en els VDP pot ser crucial (per exemple, és possible que les pautes de respostes incoherents d'un examinant puguin justificar-se per assistència a classes particulars).

Un altre tipus d'índexs es centra exclusivament en intentar detectar «proximitat» en les pautes de resposta, especialment de patrons d'errors similars (PES). Aquest tipus d'índexs pretendre detectar la probabilitat de còpia entre examinands, però com passa amb l'anàlisi dels PAR, hi ha altres explicacions alternatives als PES. Per exemple, és habitual que grups d'estudiants es reuneixin per estudiar la matèria de l'examen, per la qual cosa no seria estrany trobar patrons similars d'encerts i d'errors en diferents membres del mateix grup d'estudi. També aquí la informació de la VDP és necessària (per exemple, si s'ha enregistrat la localització fila-columna a l'aula dels dos examinands «sospitosos» de copiar el dia de la prova).

A la pràctica real encara és poc freqüent l'anàlisi de patrons de resposta. En corregir exàmens la majoria de docents no tendeix a considerar la pauta o origen dels encerts, errors i omissions ni la seva coherència. La nostra experiència indica que, quan es realitza, el professorat està més interessat pels aspectes «negatius» (còpia, respostes a l'atzar) que pels positius (detectar a alumnes en que la capacitat queda infravalorada a la prova). En aquest sentit destaquem la capacitat formativa que tenen aquests instruments analítics (per exemple, detectar pautes incorrectes d'estudi que poden ser millorades). Tot és qüestió d'interès.

ANNEX. CORRECCIÓ DE LA PUNTUACIÓ PER CONJECTURA

Una de les expressions més estesa entre usuaris de test i productes online és:

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

Per tal de conèixer la puntuació corregida *N* d'un examinat només s'ha de restar al número d'encerts *E* el número de respostes fallades *F* dividit entre el número *n* d'alternatives de resposta dels ítems menys 1. Aparentment és un càlcul simple, el problema és que sovint es desconeixen les condicions que assumeix. Un exemple hipotètic servirà de base per a conèixer l'origen d'aquesta expressió i les seves condicions d'aplicació (dins els parèntesis apareixeran citades les condicions que es descriuen més endavant).

Imaginem un examen de *K* ítems amb el mateix número *k* d'alternatives de resposta cadascú (condició 1). Un examinat qualsevol encerta un número *E* de preguntes y falla un número *F*. No existeix la possibilitat d'ometre (condició 2). Dels *E* encerts aconseguits s'assumeix que una quantitat *N* es deu al seu nivell de capacitat i altre quantitat *C* que ha triat la resposta correcta per conjectura (condició 3). Així, aquests *C* ítems que ha encertat por conjectura només són una part del total de preguntes *C1* en que, suposadament, ha intentat endevinar la resposta (condició 4). D'aquesta manera podem desglossar la puntuació de l'examinat en els següents components:

K: número d'ítems de l'examen

k: número d'alternatives de resposta de cada pregunta

E: número total d'encerts

F: número total d'ítems fallats

N: número d'ítems encertats pel propi nivell de capacitat i sense conjecturar

C1: número d'ítems en que el examinat ha intentat conjecturar

C: número d'ítems encertats por conjectura

El total de *K* ítems de l'examen és la suma de les respostes encertades i de les fallades.

$$K = A + F$$

D'això se'n deriva que el total d'encerts és la suma d'encerts sense conjecturar N i encerts gràcies a la conjectura C.

$$A = N + C$$

Essent els encerts N fruit de la capacitat de l'examinat:

$$N = A - C$$

Quant als encerts aconseguits conjecturant (C) podem assumir que són només una part d'aquells que ho han intentat (C1). Com cada ítem té la mateixa quantitat *k* d'alternatives el valor *C* serà la següent fracció de *C1* (condició 5).

$$C = C1/k$$

D'altra banda el total de respostes fallades F serà la diferència entre el número d'ítems en que l'examinat ha intentat conjecturar (C1) i la quantitat que ha aconseguit encertar conjecturant (C).

$$F = C1 - C$$

Com hem vist abans, *C* també e pot expressar com una fracció de manera que

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

Que equival a:

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

I, acceptant abans que *C=C1/k* i substituent ara *C1* queda:

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

Simplificant *k*, queda que:

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

Tornant a la expressió inicial referida als encerts deguts al propi nivell de capacitat N, podem ara prendre l'expressió:

$$N = A - C$$

I substituent *C* retrobar la fórmula presentada al principi.

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

GLOSSARI

Administrar un test: aplicar la prova a un o més individus.

Alternativa correcta: lletra o número que identifica la resposta que puntua.

Alternativa múltiple (AM): ítem format per un enunciat i unes opcions de resposta on cal seleccionar la resposta correcta o la millor possible.

Anàlisi (en Bloom): capacitat de subdividir la informació rebuda.

Anàlisi d'alternatives incorrectes: discriminació de l'ítem en el cas que cadascun dels distractors fos correcte.

Anàlisi dels ítems: procés d'examen dels indicadors dels ítems del test.

Aplicació (en Bloom): capacitat d'abordar situacions o resoldre problemes nous utilitzant principis i regles prèviament apresos.

Auditoria de test: control de la qualitat del test. És el «test del test».

Auditoria qualitativa: revisió dels aspectes formals i de disseny de la prova.

Auditoria quantitativa: obtenció d'indicadors numèrics i gràfics que informen del funcionament del test a partir de les respostes rebudes.

Banc d'ítems (BI): col·lecció o biblioteca d'ítems de format i contingut estandarditzat.

Biaix: per a grups homogenis d'examinats, un mateix ítem o test dóna puntuacions diferents en funció d'una característica aliena a l'objectiu del test (cultura, raça, nivell social, etc.).

Clau o plantilla de correcció: codi d'encerts, errors i puntuació que s'atorga a les respostes dels ítems.

Coeficient de fiabilitat: indicador basat en l'equivalència o consistència de mesures d'un mateix grup d'examinats (expressat com correlació).

Coherència de resposta: patró de resposta d'un examinat ajustat al que s'esperava.

Comprensió (en Bloom): capacitat de captar el significat o sentit directe de la informació presentada.

Conjectura: recursos que ajuden a encertar un ítem quan es desconeix la resposta.

Coneixement (en Bloom): capacitat de recordar termes, principis, normes, etc.

Consigna: instruccions verbals o escrites per respondre una prova.

Correcció de la conjectura: descompte de la puntuació del test suposadament a causa de la conjectura.

Corba d'ítem: gràfica que relaciona la capacitat d'un examinat amb la probabilitat d'encertar un ítem.

Corba d'omissió: gràfica que relaciona la capacitat d'un examinat amb la probabilitat que ometi un ítem.

Corba normal: model matemàtic que relaciona la desviació típica de les puntuacions amb la proporció de casos, o àrea de la corba.

Dificultat d'alternativa de ítem: proporció en què una alternativa és escollida.

Dificultat d'ítem: proporció d'examinats que escull l'alternativa correcta.

Discriminació: capacitat de l'element per diferenciar entre examinats de diferent capacitat.

Discriminació d'alternativa: discriminació considerant que una alternativa errònia d'un ítem fos la certa.

Discriminació negativa: els que encerten l'ítem són els examinats amb menor puntuació total i els que el fallen els de millor puntuació total.

Enunciat: premissa, nucli o part introductòria de l'ítem que planteja l'examinat la tasca a exercir.

Error estàndard de mesura (SEM): indicador de la precisió de les puntuacions d'un test.

Igualtat d'atractiu de les alternatives: distribució homogènia de les eleccions errònies entre les alternatives incorrectes.

Independència local (IL): només la capacitat de l'examinat determina la seva resposta als ítems.

Ítem: element d'un test. Una pregunta és un tipus d'ítem però un ítem no sempre és una pregunta.

Ítem obert: l'examinat elabora la resposta

Ítem tancat: l'examinat tria la resposta entre diverses opcions.

***Learning managment system* (LMS)**: sistema online de gestió de l'aprenentatge.

Matriu de dades brutes (**MDB**): lletres o números de les alternatives marcades per cada examinat (AABDADCCA...).

Matriu de dades netes (MDN): procedeixen de les dades brutes després de comparar o corregir amb la plantilla o clau de correcció (01101001...).

Longitud del test (L): nombre d'ítems del test.

Patrons atípics de resposta (PAR): pautes de resposta que no es corresponen amb el model psicomètric de la prova.

Patrons d'error similars (PES): en parelles o grups d'examinats amb un patró d'errors molt similar especialment en els ítems difícils.

Quiz: examen o prova d'assaig i autoavaluació habitual en entorns LMS.

Regles de generació d'ítems (RGI): directrius per a crear ítems adequadament.

Síntesi (en Bloom): capacitat de reunir elements/parts per formar un tot.

Solapament: una alternativa errònia fa el paper de l'alternativa especificada com a correcta.

Taula d'especificació d'objectius (TEO): estructura que relaciona els continguts a avaluar amb la forma com s'avaluaran.

Teoria clàssica dels test (TCT): teoria psicomètrica basada en el concepte que la puntuació que obté una persona en una prova és el resultat de sumar a la seva puntuació veritable una puntuació de causa de l'error de mesura. Permet quantificar aquest error i per tant estimar la fiabilitat.

Teoria de resposta a l'ítem (TRI): conjunt de principis psicomètrics i models matemàtics que relacionen la capacitat dels examinats amb la probabilitat d'obtenir determinada puntuació en els ítems.

Test adaptatiu informatitzat (TAI): test personalitzat administrat per ordinador. Es basa en algoritmes de presentació d'ítems que decideixen, per a cada persona i en cada pas de la prova, l'ítem òptim per avaluar el seu nivell de coneixements.

Test de norma de grup (TNG): interpreten la puntuació d'un individu en relació a l'execució global del grup al qual pertany.

Test referits al criteri (TRC): produeixen mesures directament interpretables independents als resultats del grup.

Validesa: qualitat del test que informa si mesura el que pretén.

Validesa de contingut: basada en la representativitat dels ítems.

Vector descriptor d'ítem (VDI): unió de descriptors codificats de les característiques dels ítems

Vector descriptor de persona (VDP): unió de descriptors codificats de característiques rellevants de les persones examinades abans (pre) i després de l'examen (post).

Verdader/fals (VF): format d'ítem en què només cal valorar si el que planteja és correcte o no.

BIBLIOGRAFIA

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

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

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

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

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

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

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

—— (2009a). *Nuevos usos de los formatos de respuesta de selección en la evaluación diagnóstica y formativa*. Comunicació presentada en el VI Congrés Internacional de Docència Universitària i Innovació. Barcelona.

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

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

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

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

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

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

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

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

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

International Test Commission (2013). *International guidelines on quality control in scoring, test analysis, and reporting of test scores*. Disponible en: <www.intestcom.org>.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

NORMES PER A LA PRESENTACIÓ D'ORIGINALS PER A LA COL·LECCIÓ

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

NORMES PER ALS COL·LABORADORS

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

EXTENSIÓ

Les propostes del Quadern no podran excedir **l'extensió de 50 pàgines (en Word)**, uns 105.000 caràcters, espais, referències, quadres, gràfiques i notes incloses.

PRESENTACIÓ D'ORIGINALS

Els textos han d'incloure, en format electrònic, un **resum** d'unes deu línies i tres paraules clau, no incloses al títol. Igualment han de contenir el **títol**, un ***abstract*** i tres ***keywords*** en anglès.

Per a les **formes de citar i referències bibliogràfiques** han de remetre's a les utilitzades en aquest *Quadern*.

AVALUACIÓ

L'acceptació d'originals es regeix pel **sistema d'avaluació externa per pars**.

Els originals són llegits, en primer lloc, pel **Consell de Redacció**, que valora l'adequació del text a les línies i objectius dels *Quaderns* i si compleix els requisits formals i els mínims de contingut científic exigit.

Els originals són sotmesos, en segon lloc, a **l'avaluació de dos experts**, especialistes en la temàtica de la qual tracta l'original i l'àmbit disciplinari corresponent. Els autors reben els comentaris i suggeriments dels avaluadors i la valoració final amb les esmenes i canvis que cal fer, si és el cas, abans de ser acceptat per a la seva publicació.

Si els canvis exigits són significatius o afecten bona part del text, el nou original és sotmès a l'avaluació de dos experts externs i d'un membre del Consell de Redacció. El procés es duu a terme com a «doble cec».

Revisors

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

L'Institut de Ciències de l'Educació (ICE) de la Universitat de Barcelona inicià fa uns anys la publicació dels **QUADERNS DE DOCÈNCIA UNIVERSITÀRIA** amb l'objectiu de posar a l'abast del professorat universitari documents i materials de treball referits a temes relacionats amb la docència superior que facilitessin la seva formació, l'intercanvi d'experiències i la difusió de «bones pràctiques» docents. Amb aquests *Quaderns* pretenem estar atents als temes nous i emergents en l'actual conjuntura universitària, per tal de donar a conèixer i difondre iniciatives innovadores en el camp de la docència universitària, que responguin a les línies següents:

- Propostes de marcs de referència rigorosos i generals que ajudin a clarificar conceptes clau.
- Estratègies docents i bones pràctiques de planificació, metodologia i avaluació de l'ensenyament-aprenentatge, desenvolupades en contextos acadèmics específics i diversos.
- Tècniques i tàctiques, de marcat caràcter didàctic, presentades en materials i propostes concretes de treball i reflexió sobre la pràctica d'equips docents disciplinaris o interdisciplinaris.

ice
IDP
Institut de Desenvolupament Professional
UNIVERSITAT DE BARCELONA