Mesurament

PID_00255061

Sergio Escorial Martín

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Sergio Escorial Martín

Llicenciat en Psicologia per la Universitat Autònoma de Madrid. Doctor en Psicologia per la mateixa universitat, dins del programa de Doctorat de Ciència de la Conducta, amb especialització en metodologia. Ha publicat nombrosos articles en revistes indexades en el JCR (un 86% en revistes del Q1-Q2) i ha estat guardonat, en dues ocasions, amb el Premi TEA Edicions. En l'actualitat, exerceix com a professor contractat doctor en el Departament de Metodologia de les Ciències del Comportament de la Universitat Complutense de Madrid (UCM), càrrec amb el qual ha obtingut el certificat d'Excel·lència Docent. La seva línia de recerca actual és el desenvolupament i la validació d'instruments psicomètrics.

Cap part d'aquesta publicació, incloent-hi el disseny general i la coberta, no pot ser copiada, reproduïda, emmagatzemada o transmesa de cap manera ni per cap mitjà, tant si és elèctric com químic, mecànic, òptic, de gravació, de fotocòpia o per altres mètodes, sense l'autorització prèvia per escrit dels titulars del copyright.

Índex

Introducció..... 5

1. El procés de construcció d'un instrument..... 7
1.1. Disseny del test 7
1.2. Especificacions del test 8
1.3. Elaboració dels ítems 8
1.4. Revisió dels ítems 9
1.5. Estudi pilot 9
1.6. Estudi de camp 10
1.7. Construcció del manual 11

2. Control de qualitat dels ingredients: l'anàlisi d'ítems..... 12

3. La fiabilitat com a criteri de qualitat global de la mesura.... 15
3.1. Fiabilitat com a estabilitat temporal 16
3.2. Fiabilitat com a consistència interna 18

4. Les evidències de validesa com a criteri de qualitat de la mesura..... 20
4.1. Evidències relacionades amb el contingut 21
4.2. Evidències relacionades amb l'estructura interna de la prova 22
4.3. Evidències basades en la relació amb altres variables 23

5. Un cas per a futurs logopedes..... 25

Activitats..... 27

Bibliografia..... 28

Introducció

> «De vegades no entenc el comportament humà. Només intento fer el meu treball de la forma més eficient.»
>
> (C-3PO a *Star Wars, episodi V, L'Imperi Contrataca*)

L'objectiu final de qualsevol disciplina científica és arribar a establir principis generals que permetin descriure, predir i explicar els fenòmens de la seva àrea d'interès. Per a arribar a aconseguir aquest objectiu totes les disciplines han de recollir un conjunt de dades o observacions destinades a fonamentar les seves teories i, per aquesta raó elemental, el **mesurament** és una part absolutament fonamental del procés. Seguint a Martínez Arias, Hernández-Lloreda i Hernández Lloreda (2006) podem entendre el mesurament com un procés mitjançant el qual relacionem un sistema empíric que conté les modalitats presents en una variable o atribut d'interès amb un sistema formal (normalment representat per nombres). És a dir, el mesurament consisteix en l'atribució de nombres a atributs dels subjectes, de tal manera que els nombres reflecteixin els diferents graus (o modalitats) de l'atribut que està sent avaluat. Per exemple, quan mesurem un atribut com la intel·ligència, el que farem fer és seguir un procediment (idealment sistemàtic i estandarditzat) que ens porti a obtenir un nombre, el qual ens indicarà el grau en què aquest atribut (la intel·ligència) està present en el subjecte avaluat. Hi ha força consens sobre aquest concepte en la literatura científica (p. ex. Abad i altres, 2011; Lord i Novick, 1968; Nunnally i Bernstein, 1994).

Baixant a un terreny més pràctic, l'activitat professional del logopeda requereix en molts moments la utilització i construcció de tests o proves estandarditzades dirigides a avaluar determinats constructes que no resulten ser susceptibles a un procés de mesurament directe. Resulta usual, per exemple, en l'àmbit dels trastorns del llenguatge, l'aplicació de tests de fluïdesa verbal, comprensió escrita, o vocabulari, entre d'altres. En el terreny de la pràctica clínica, un logopeda necessita aplicar determinades proves per a diagnosticar de manera més efectiva i eficient i, fins i tot, per a valorar l'eficàcia de la intervenció desenvolupada amb un client. Per aquesta raó, cada vegada és major el nombre de tests disponibles al mercat per a la seva utilització. N'hi ha prou amb mirar els catàlegs d'empreses especialitzades per a adonar-nos de la gran extensió d'atributs que ja podem mesurar mitjançant tests. Així doncs, resulta evident que el logopeda necessita conèixer les possibilitats de cadascun d'aquests tests: la informació que aporta, com s'interpreten les puntuacions que proporciona, en quin grau ens podem fiar d'aquestes puntuacions, per a quin tipus de persones resulta apropiada la seva aplicació i la resta d'informació que ajuda el professional a fer un ús adequat d'aquests instruments. El manual d'aquests tests sol incloure dades empíriques sobre tots aquests aspectes, que determinaran en gran part les garanties que ens ofereix la prova que aplicarem. D'altra

banda, és molt possible que en algunes circumstàncies el logopeda es vegi en la necessitat de construir un test concret perquè no hi ha al mercat una prova que s'adapti als seus interessos. Això pot ocórrer, per exemple, quan es treballa en un tipus de trastorn molt específic que amb prou feines ha estat abordat amb anterioritat, quan es vol avaluar un aspecte concret que no s'ha desenvolupat prèviament o quan volem treballar amb un conjunt de subjectes que presenten unes característiques molt especials. En resum, sembla més que raonable, per tant, que un logopeda adquireixi les destreses necessàries per a valorar la informació relacionada amb el procés de mesurament que inclouen els tests comercialitzats i, a més, que conegui els mètodes i tècniques fonamentals per a dissenyar una prova concreta amb finalitats específiques. Tractarem d'ajudar-vos en això en les següents pàgines.

Per a fer-ho, començarem descrivint, de manera molt sintètica, el procés natural que se segueix en la construcció d'un test. A continuació, dedicarem un apartat a l'anàlisi dels ítems que componen el test. Per a il·lustrar la importància d'aquest aspecte, emprarem una metàfora culinària: solament podem arribar a obtenir un bon plat (en el nostre cas el test) si comptem amb ingredients de primera qualitat (els ítems que componen la prova). Els dos apartats següents, destinats a l'estudi de la fiabilitat i de la validesa del test respectivament, resulten absolutament fonamentals, atès que es refereixen a la comprovació empírica de les garanties mètriques que la prova manifesta com a instrument de mesurament. Bàsicament, aquestes garanties es refereixen a la seva precisió (fiabilitat) i a la comprovació pràctica del contingut autèntic que estem avaluant i la seva utilitat (validesa).

Finalment, què busquem incidint que els professionals, en aquest cas el de la logopèdia, manegin aquests conceptes relacionats amb el mesurament? Que siguin professionals més eficients i rigorosos. Com li succeeix al bo de C3PO –segons la cita amb què obríem aquest mòdul–, en moltes ocasions el nostre treball ens portarà a enfrontar-nos a comportaments que no acabem d'entendre, però això no ens eximeix de la responsabilitat de fer aquest treball de la manera més eficient. Conèixer els fonaments del procés de mesurament, i també les propietats que hem d'exigir als instruments amb què ens ajudarem per a la presa de decisions, ens aproparà sens dubte a aquest objectiu de l'eficiència i el rigor professional. Gaudiu del viatge..... i que la força us acompanyi!

1. El procés de construcció d'un instrument

El primer que cal destacar aquí és que construir un instrument d'avaluació és molt més que redactar uns ítems o preparar uns materials per a tractar de mesurar un constructe o característica d'interès per a nosaltres. Es tracta d'una tasca que, per la seva importància, ha de ser realitzada acuradament seguint les directrius establertes per a aquesta finalitat (AERA, APA i NCME, 2014). Sent molt esquemàtics, podríem establir que per a la construcció d'un instrument ens embarcarem en un apassionant viatge que consta de les etapes següents:

1) Disseny del test

2) Especificacions del test

3) Elaboració dels ítems

4) Revisió dels ítems

5) Estudi pilot

6) Estudi de camp

7) Construcció del manual

1.1. Disseny del test

En aquesta primera fase, explicarem, de manera absolutament precisa, quin és l'objectiu del test que pretenem desenvolupar. Seguint la proposta de Navas (2001), això s'aconsegueix quan es dona resposta a aquestes tres simples, però molt rellevants, qüestions:

1) Què volem mesurar amb el nostre test? Això suposa definir de manera operativa les nostres variables d'interès. De vegades, per a establir aquesta definició ens hem de posicionar en un determinat nivell d'anàlisi (cognitiu, biològic, conductual, entre d'altres) el que implica des del primer moment, una presa de decisions que és important concretar per les seves implicacions pràctiques.

2) A qui volem avaluar amb la prova? Això implica definir detalladament qui és la nostra població diana i quines característiques presenta (edat, nivell educatiu, requisits del llenguatge, motivació...). No resulta difícil d'endevinar que aquesta és una qüestió decisiva que ha de ser clarament explicitada, ja que la seva resposta determinarà en gran manera les característiques que ha de tenir la prova que desenvoluparem.

3) Per què vull mesurar? En essència, una vegada que hem deixat clar què volem mesurar i a qui, ara ens enfrontem a la finalitat amb què volem realitzar el procés de mesurament, és a dir, quin ús pretenem fer de les puntuacions que obtinguem. Són moltes les funcions que pot complir un test (diagnòstic, selecció, classificació, consell, cribratge, per exemple), per la qual cosa també hem de ser molt concrets en aquest aspecte.

1.2. Especificacions del test

En aquesta fase del procés hem de prendre una sèrie de decisions que resultaran fonamentals per a l'estructura del test.

Exemples

Quin serà el contingut del test? És a dir, es tracta d'una prova informatitzada o de paper i llapis?

Quin serà el contingut dels elements que la componen (ítems)?

Quin és el temps màxim que vull que tingui l'aplicació de la prova? Aquí cal destacar que, malgrat que les proves llargues solen tenir millors índexs de fiabilitat, els tests llargs requereixen més temps d'administració.

Es tracta d'una prova d'aplicació individual o col·lectiva? En general, les proves d'aplicació individual requereixen més temps per a la seva administració i les seves instruccions poden ser una mica més minucioses. No obstant això, les instruccions de les proves col·lectives han de ser necessàriament més senzilles i directes.

En resum, respondre a aquestes i altres qüestions de naturalesa similar ens porta a tenir molt presents una sèrie de paràmetres que definiran el marc en què es construirà la prova en concret.

1.3. Elaboració dels ítems

Una vegada establert tot l'anterior, estem en condicions de començar a elaborar elements per a tractar de mesurar el nostre constructe d'interès. Abans de seguir, hem de tenir clar que, en el procés de mesurament mitjançant el desenvolupament del test, no hi ha un únic tipus d'ítems. Tenim a la nostra disposició un gran ventall de possibilitats (vegeu el quadre 1). A més, depenent de la pròpia naturalesa del constructe que volem avaluar (i de les especificacions del test) serà necessari elaborar ítems que tinguin una resposta correcta (ítems de rendiment òptim) o ítems que les seves alternatives de resposta reflecteixin diverses maneres habituals de comportar-se, però sense que cap d'aquestes sigui millor o pitjor en si mateixa (ítems típics de rendiment).

Quadre 1. Principals tipus d'elements (ítems)

1) Ítems d'opció múltiple: els subjectes avaluats han de triar entre poques alternatives de resposta (usualment de 2 a 5).
2) Ítems d'ordenació: es proporciona a l'avaluat un conjunt de materials que ha d'ordenar de manera correcta (p. ex. dibuixos o cubs).

3) Ítems de substitució/correcció: en aquest cas, el subjecte ha de substituir o corregir els elements per alternatives correctes (p. ex. corregir faltes d'ortografia, entonació adequada de fonemes, entre d'altres).

4) Ítems de construcció (resposta oberta): en aquest tipus d'elements, l'avaluat ha de generar la resposta completa.

5) Ítems de presentació: en aquest cas, el subjecte serà exposat a un conjunt de situacions reals o simulades per a registrar amb precisió el seu comportament en les mateixes. Exemple: *role playing* en un procés de selecció.

En aquesta fase, els logopedes o l'investigador establiran una taula d'especificacions que permeti maximitzar la validesa del contingut de la prova. Una vegada establertes aquestes especificacions s'elaboraran els ítems, que es formulen de manera lògica perquè mesurin el constructe, variable o tret que interessa avaluar amb el test. Malgrat que la creació de bons ítems és en part un art, hi ha una guia de recomanacions que es pot seguir (Downing i Haladyna, 2006; Haladyna, 2004; Haladyna i altres, 2002; Moreno, Martínez i Muñiz, 2004; Roid i Haladyna, 1982).

1.4. Revisió dels ítems

Abans de donar per finalitzada la fase de construcció, que únicament permet arribar a muntar i assemblar la que serà considerada com una versió preliminar de la prova, convé fer una revisió en profunditat dels ítems que la componen. Els especialistes recomanen que aquesta revisió s'estableixi a tres nivells:

1) Experts en el constructe que es pretén mesurar (diferents dels dissenyadors de la prova), en aquest cas professionals de la logopèdia i del llenguatge per a verificar que s'hagin definit bé els constructes d'interès i els aspectes implicats en els mateixos.

2) Subjectes de la població de referència per a evitar fallades en la redacció i assegurar-nos de la seva adequació inicial.

3) Experts en mesurament per a evitar fallades metodològiques i altres problemes freqüents en l'elaboració inicial de les proves.

Els ítems que passin aquest primer control de qualitat s'organitzaran en el que serà la versió preliminar de la prova, que serà acuradament examinada en la fase següent del procés.

1.5. Estudi pilot

En aquesta fase del procés, es deixa enrere la perspectiva racional o el punt de vista dels experts i s'aplica el test a una mostra d'una grandària relativament petita i no representativa per a obtenir informació empírica sobre la qualitat dels ítems. Aquesta informació se sol agrupar en tres grans apartats:

1) Obtenció d'índexs i paràmetres psicomètrics dels ítems (com els que s'exposaran en l'apartat següent).

2) Anàlisi del funcionament dels distractors (opcions incorrectes) o de les opcions de respostes.

3) Estudio del funcionament diferencial dels ítems.

A partir d'aquestes anàlisis s'avalua si la regla de puntuació de l'ítem és l'adequada, si hi ha ítems ambigus o ítems en les distribucions dels quals s'apreciï una diferència entre subjectes que no està relacionada amb el nivell que aquests tenen en la característica mesurada, sinó més aviat amb la seva pertinença a un grup determinat, com per exemple home o dona (esbiaixats). En suma, es tracta de buscar evidències que avalin la qualitat dels ítems que componen la prova. Quan un ítem no té un bon comportament psicomètric, els autors de la prova s'haurien de plantejar la seva eliminació. Després d'aquest procés, s'arriba a la versió definitiva de la prova.

1.6. Estudi de camp

Per a entendre la importància d'aquest apartat, hem d'indicar abans que, des del punt de vista de la interpretació d'una puntuació, hi ha **tests referits a un criteri**, en què la puntuació d'un subjecte és interpretada sobre la base d'un criteri fixat amb anterioritat.

Exemple

Normalment un cinc és la qualificació mínima per a aprovar una assignatura, amb independència que aquesta matèria l'aprovin molts o pocs alumnes.

D'altra banda, hi ha els **tests referits a normes**, en què la puntuació d'un subjecte determinat s'interpreta en relació amb els altres subjectes de la població a què pertany.

Exemple

Una puntuació de 36 en la prova de Matrius progressives de Raven (escala superior) [un test que serveix per a avaluar la capacitat de raonament abstracte d'un subjecte amb un nivell educatiu superior] es correspon amb una puntuació molt elevada, perquè més del 95% de la població espanyola amb estudis superiors estaria per sota de la mateixa.

En aquest últim tipus de test resulta absolutament fonamental disposar d'un grup de referència que sigui realment representatiu de la població d'interès. En aquesta fase d'estudi de camp, s'accedeix a aquest grup. Aquí, la prova serà aplicada a una mostra representativa[1] de la població d'interès (és a dir, el grup normatiu). Si en aquesta població d'interès hi ha subpoblacions o subgrups amb alguna característica que té un cert impacte sobre la variable d'interès, s'assegurarà la seva presència en el grup normatiu en percentatges similars als de la població (p. ex. grups en funció del sexe, edat, nivell d'estudis, classes social, entre d'altres). Una vegada aplicat el test als subjectes d'aquest grup normatiu es procedirà a obtenir les dades estadístiques per a contrastar les evidències mètriques del test en el seu conjunt: fiabilitats, errors de mesura, evidències de validesa per a l'ús pretès de la prova, elaboració de normes d'interpretació (és a dir, barems).

(1)El tema de la representativitat i el mostreig és molt recurrent en manuals excel·lents d'estadística i disseny. Només cal assenyalar que les mostres obtingudes mitjançant tècniques de mostreig probabilístiques que inclouen l'atzar i que, a més, tenen grandàries relativament grans, se solen considerar representatives. No obstant això, cal tenir molt present que en proves clíniques no tenim, en moltes ocasions, la possibilitat d'aplicar procediments de mostreig probabilístics i que, moltes poblacions clíniques són tan reduïdes que és realment complicat arribar a obtenir grups normatius d'una certa grandària.

1.7. Construcció del manual

L'última fase del procés, però no per això menys important, és l'elaboració d'un manual del test. Es tracta d'un document de caràcter tècnic que permet a altres professionals de l'àmbit de la logopèdia aplicar l'instrument amb coneixement de les seves propietats i garanties (o absència d'aquestes). En general, amb independència del que s'estigui avaluant amb el test en qüestió o de la casa editorial que comercialitzi la prova, en cas de publicar-se, tot manual s'hauria d'adaptar a una estructura tipus en què caldria destacar els elements següents:

1) Fonamentació teòrica del constructe que pretén mesurar el test i dels models en què ens basa tant per a la seva operativització com per a la seva interpretació.

2) Finalitat a què es pretén destinar el test.

3) Determinació de les poblacions a què va dirigit i descripció detallada del grup normatiu empleat.

4) Instruccions detallades per a la seva administració correcta.

5) Fonamentació estadística, en què es detallen totes les evidències relatives a la qualitat mètrica dels reactius que componen la prova (ítems), els coeficients de fiabilitat i en què s'integren les evidències principals de validesa considerades conjuntament.

6) Finalment, les normes d'interpretació de les puntuacions en el test (idealment amb algun exemple de casos o perfil tipus) i els barems de la prova.

2. Control de qualitat dels ingredients: l'anàlisi d'ítems

En aquesta secció exposarem, de manera molt sintètica, en què consisteix l'anàlisi dels ítems que componen el test. Podeu trobar una aproximació molt més elaborada en els manuals de mesurament de, per exemple Muñiz, Martínez, Moreno, Fidalgo, i García Cueto (2005). Ja vam comentar en la introducció que la importància d'aquest tipus d'evidències resideix en el fet que, si volem elaborar un plat de qualitat (el test), hem d'assegurar que els ingredients que emprarem en la seva elaboració (els ítems) també tinguin aquesta qualitat.

Els ítems o elements que componen un test s'han formulat de manera lògica perquè mesurin (i a més ho facin bé) el constructe, variable o tret que interessa avaluar amb l'instrument. Ara bé, el grau en què cada ítem és un «bon mesurador» del tret d'interès és alguna cosa que es pot comprovar estadísticament de manera senzilla si obtenim alguns indicadors bàsics per a cada ítem. A continuació, explicarem molt breument els que possiblement són els dos índexs més emprats per a descriure les propietats mètriques d'un ítem:

1) L'índex de dificultat (D_j)

Aquest primer indicador serveix per a quantificar el grau de dificultat de cada ítem en proves de rendiment òptim (és a dir, proves de rendiment com ara les preguntes tipus test d'un examen, de capacitat cognitiva general o d'aptitud verbal).

Es tracta d'un índex molt fàcil de calcular ja que es defineix com el quocient entre el nombre de subjectes que han encertat un ítem en particular, que denominarem A_j i el nombre total de subjectes que l'han intentat resoldre, que denominarem N_j. És important aclarir que en l'N_j s'inclouen solament els subjectes que han contestat a aquest ítem (encerts o errors) però queden exclosos els subjectes que no ho han intentat (omissions).

Exemple

A continuació, es presenten les respostes d'una mostra de 6 persones en un test format per 6 ítems dicotòmics, és a dir, amb dues opcions de resposta (1 indica encert i 0, error):

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Subjecte 1	1	1	1	1		1
Subjecte 2	1	1	0		0	
Subjecte 3	1	1	1			1

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Subjecte 4	1	1	1	1		
Subjecte 5	1	0	0	0	0	
Subjecte 6	1	1	1	0	0	0
A_j	6	5	4	2	0	2
N_j	6	6	6	4	3	3
D_j	1	0,83	0,67	0,50	0	0,67

Sobre la base d'aquest exemple simple podem destacar alguns aspectes importants de la interpretació de D_j:

a) El valor mínim que pot assumir D_j és 0 (cap subjecte encerta l'ítem) i el valor màxim, 1 (tots els subjectes que ho han intentat l'han encertat).

b) A mesura que D_j s'apropa a 0, ens indica que l'ítem ha resultat molt difícil; mentre que si s'apropa a 1, mostra que ha resultat molt fàcil.

c) D_j està relacionat amb la variància dels ítems. Si D_j és 0 o 1, la variància és igual a zero; i a mesura que D_j s'apropa a 0,5, la variància de l'ítem augmenta.

En resum, una bona estratègia a l'hora de dissenyar instruments de rendiment òptim seria col·locar els ítems més fàcils (amb major D_j) a l'inici, els d'una dificultat mitjana (entre 0,30 i 0,70) a la part central i els més difícils (amb menor D_j) al final. El nombre d'ítems de cada categoria de dificultat que s'han d'incloure en el test depèn dels objectius que vulgui aconseguir la persona que dissenya el qüestionari. En general, la major part dels ítems han de ser de dificultat mitjana.

2) L'índex d'homogeneïtat (H_j)

L'índex d'homogeneïtat d'un ítem (H_j), també anomenat índex de discriminació, es defineix com la correlació de Pearson entre les puntuacions dels N subjectes en l'ítem j i les puntuacions X en el total del test. Per tant, també es tracta d'un índex molt senzill de calcular.

Exemple

A continuació, es presenten els resultats d'un test format per 3 ítems amb format de resposta de categories ordenades, que es valoren entre 0 i 4.

	Ítem 1	Ítem 2	Ítem 3	X
Subjecte 1	1	2	4	7
Subjecte 2	2	0	1	3

	Ítem 1	Ítem 2	Ítem 3	X
Subjecte 3	4	3	4	11
Subjecte 4	0	0	0	0
Subjecte 5	3	2	0	5
H_j	0,76	0,93	0,83	
$H_{j(c)}$	0,51	0,85	0,52	

Aquest índex ens informa del grau en què un ítem està mesurant el mateix que la prova globalment, és a dir, el grau en què contribueix a l'homogeneïtat o consistència interna del test. Els ítems amb índexs baixos d'homogeneïtat mesuren quelcom diferent al que reflecteix la prova en el seu conjunt. Per aquesta raó, s'haurien d'eliminar els que tenen un H_j per sota de 0,2. És important tenir present que els H_j s'han de calcular per a cada domini en concret, en el cas de proves que mesurin diversos constructes diferents. A més, quan un H_j és negatiu i alt, hem de qüestionar el sistema de puntuació de les respostes en aquest ítem ja que el més probable és que ho hàgim codificat a l'inrevés.

Finalment, quan un test té un nombre petit d'ítems (com en el cas de l'exemple), resulta més apropiat obtenir l'índex d'homogeneïtat corregit ($H_{j(c)}$). Aquest índex, que s'interpreta exactament igual que l'anterior, consisteix en correlacionar les puntuacions en un ítem amb les puntuacions en el total del test després de restar d'aquest total les puntuacions de l'ítem l'índex del qual volem obtenir. Com és lògic suposar, l'$H_{j(c)}$ corregit d'un ítem sol ser inferior a la seva H_j sense corregir.

3. La fiabilitat com a criteri de qualitat global de la mesura

Imaginem un sastre que amb una cinta mètrica mesura diverses vegades un mateix tros de tela. En aquest cas, i excepte petites desviacions, sempre obtindrà el mateix mesurament, a causa que tant la cinta mètrica com el seu tros de tela romanen invariants. Ara bé, què passa quan un logopeda empra el subtest de vocabulari del WAIS-IV per a mesurar la capacitat verbal d'un subjecte? En aquest cas pot ocórrer que ni un ni l'altre romanguin invariants d'una situació a una altra. Per això és important valorar la informació proporcionada en els manuals pels especialistes en mesurament per a establir el grau d'estabilitat de l'instrument de mesurament amb què estiguem treballant i el seu grau de precisió; això és, la seva fiabilitat.

El concepte de fiabilitat es basa en el model clàssic que, al seu torn, se sustenta en una sèrie de supòsits molt simples i les seves deduccions. Per una qüestió de brevetat en l'exposició prescindirem de desenvolupar detalladament aquesta part, però remetem el lector interessat als capítols de fiabilitat presents en els manuals de mesurament (p. ex. Abad i altres, 2011; Martínez Arias i altres, 2006). El primer supòsit estableix que la puntuació observada d'una persona (i) en un test es descompon linealment en dos components: la puntuació veritable (V_i) i l'error de mesura (I_i):

$$X_i = V_i + I_i$$

En línia amb aquest primer supòsit, elegantment simple, la resta de supòsits i deduccions ens portaran a poder descompondre la variància de les puntuacions en un test en dos components, un relacionat amb els errors i un altre amb les puntuacions veritables. No obstant això, en totes les formulacions apareixen elements no observables.

$$\sigma_x^2 = \sigma_v^2 + \sigma_e^2$$

Doncs bé, sabent que la variància d'una prova σ_x^2 es descompon additivament en una variància a causa de les puntuacions veritables σ_v^2 i en variància a causa de l'error σ_e^2, la correlació entre els dos tests o mesures paral·leles ens indicarà quina proporció de variància es deu a la variabilitat en el veritable nivell del tret.

En el model clàssic, aquesta correlació entre formes paral·leles es denomina coeficient de fiabilitat ($\rho_{xx'}$). El coeficient de fiabilitat reflecteix la precisió de mesura sempre que assumim –entre altres coses– que en el grup de subjectes a què s'aplica el test hi ha una certa variabilitat en la característica que s'està mesurant (Streiner, 2003). És a dir:

> Conceptualment entendrem per fiabilitat el grau d'estabilitat, precisió o consistència que manifesta un test concret com a instrument de mesurament d'una característica determinada.

De tot el que hem dit fins a aquí, tant el model clàssic de puntuació veritable, com el plantejament de la fiabilitat com a correlació entre formes paral·leles –apuntat molt resumidament abans– s'han establert en termes paramètrics; és a dir, suposant que coneixem les dades de la població de referència. No obstant això, la situació a la pràctica és una mica diferent, ja que l'única cosa del que disposarem és de les dades que hàgim obtingut en una mostra de pacients o un grup normatiu concret. Això implica que, d'una manera directa, únicament tindrem un conjunt de puntuacions en un test en particular i a partir d'aquestes podrem obtenir els estadístics més adequats en cada cas.

En els apartats següents, indicarem molt breument en què consisteix la fiabilitat com a indicador d'estabilitat temporal i en què consisteixen els índexs de fiabilitat que fan referència a fins a quin punt les diferents parts del test mesuren una mateixa característica de manera consistent (fiabilitat com a consistència interna).

3.1. Fiabilitat com a estabilitat temporal

Una aproximació molt emprada a l'hora d'estimar la fiabilitat d'una prova és el mètode test-retest. Amb aquest mètode s'obté el coeficient de fiabilitat aplicant el mateix test dues vegades. Al coeficient obtingut així se'l denomina coeficient de **fiabilitat com a estabilitat temporal** i reflecteix el grau de concordança de les mesures preses a un mateix conjunt de subjectes en dos moments temporals diferents. Estrictament parlant no és altra cosa que el coeficient de correlació de Pearson entre les puntuacions obtingudes per un grup de subjectes en un moment temporal inicial (test) i les obtingudes pel mateix grup de subjectes en un moment temporal posterior (retest). Com més proper a ú sigui el valor de correlació, major serà la fiabilitat temporal de la mesura.

Us serà senzill comprendre que, en essència, aquest mètode es deriva fàcilment del model lineal clàssic, apuntat anteriorment, segons el qual es defineix la fiabilitat com la correlació entre les puntuacions empíriques en dues formes paral·leles, ja que no hi ha un major grau de paral·lelisme entre dos tests que quan en realitat es tracta de la mateixa prova administrada dues vegades.

Atesa la seva naturalesa, aquest tipus de coeficient de fiabilitat només s'ha d'emprar quan el tret o constructe a mesurar s'assumeix estable i, per tant, els canvis de les puntuacions en el temps reflecteixen una falta de precisió en la mesura.

En aquest mètode, un element absolutament fonamental és la determinació de l'interval temporal entre les aplicacions i, per tant, s'ha d'informar del mateix (AERA, APA i NCME, 2014). Per a determinar aquest interval, tots els efectes en les respostes a causa de la doble aplicació (p. ex. efectes de l'aprenentatge, la fatiga, la maduració, el record, la motivació o el desig de congruència), haurien de ser analitzats i controlats.

Finalment, els usuaris de proves han de tenir molt present que si l'interval temporal que els investigadors van adoptar en l'aplicació d'aquest procediment és massa curt i no hi ha efectes de fatiga, se sol produir una sobreestimació de la fiabilitat perquè es recorden les respostes (aprenentatge), mentre que si l'interval és massa llarg es pot produir una infraestimació perquè es poden donar canvis reals en el tret o constructe avaluat (efectes maduratius). No hi ha una regla fixa en aquest sentit i l'interval temporal que es recomana establir per a minimitzar aquests efectes està en funció de:

1) la variable sobre la qual s'estigui treballant (no és igual una mesura de capacitat cognitiva que una d'empatia),

2) les característiques dels subjectes (no és igual un adolescent que una persona de la tercera edat), i

3) la interacció d'ambdues.

Com una recomanació general, que té les seves excepcions i adaptacions, els possibles efectes a causa de l'aprenentatge són minimitzats si l'interval temporal entre el test i el retest és d'almenys 8 setmanes (o més), mentre que els possibles efectes a causa de la maduració també són minimitzats si l'interval temporal entre el test i el retest és de menys de 6 mesos (pot ser, fins i tot, de més temps si els subjectes són adults).

3.2. Fiabilitat com a consistència interna

Imaginem que acabem de realitzar un examen tipus test de la nostra assignatura favorita del grau en Logopèdia. És esperable (i seria desitjable) que les preguntes (els ítems) que formen part de la prova representin la varietat dels continguts que formen part del programa d'aquesta assignatura en particular. Malgrat això, esperaríem que les diferents parts del test mesuressin amb la mateixa precisió (concordança), per a assegurar-nos que les notes de l'examen són adequadament interpretades. Doncs bé, aquesta concordança entre les puntuacions dels subjectes en els diferents ítems (preguntes de l'examen) és una propietat psicomètrica molt rellevant a què anomenarem **fiabilitat com a consistència interna**[2].

(2)Aquesta introducció a la fiabilitat com a consistència interna ens dirigeix cap a debats apassionants sobre les implicacions del mesurament en qualsevol procés d'avaluació. Per exemple, si aquest examen no té una adequada consistència interna, llavors, les puntuacions de dos alumnes que obtenen la mateixa nota, per exemple un 4.5, responent a diferents preguntes cadascun d'aquests (és a dir, a diferents parts de l'examen), haurien de ser interpretades de la mateixa manera? o, atès que la nota de l'examen serveix als docents per a prendre decisions rellevants per a la vida (i l'economia) dels alumnes, quin hauria de ser el nivell mínim de precisió que hauríem d'exigir als nostres exàmens per a prendre aquesta decisió?

Com vam dir al començament d'aquest apartat, a partir del model clàssic s'arriba a plantejar el concepte de fiabilitat de les puntuacions en una prova d'avaluació (test), que representa la proporció de la variància de les puntuacions observades que es deu a la variància de les puntuacions veritables. En termes generals podem considerar que aquesta propietat ens indica la replicabilitat o concordança de la mesura, però ja hem vist que la mateixa es pot operativitzar de diferents maneres. Així, hem indicat que podem estimar el coeficient de fiabilitat com una correlació entre dues formes paral·leles d'un mateix test, o bé mitjançant l'administració d'un test a un mateix grup de subjectes en dos moments temporals diferents.

No obstant això, també hi ha tot un conjunt d'índexs proposats per a estimar el coeficient de fiabilitat que es basa en una única aplicació del test. Per tant, a la pràctica aquests índexs són molt menys costosos d'obtenir pel que solen ser els que es presenten amb major freqüència. Amb aquests mètodes s'estudia la concordança entre les puntuacions dels subjectes en diferents parts del test. Un procediment clàssic és el denominat *mètode de dues meitats*. Aquí dividirem el test en dues meitats (habitualment formades per ítems pars i imparells cadascuna d'aquestes) i s'observa com correlacionen les puntuacions dels subjectes en les dues meitats. La correlació entre ambdues meitats es pot entendre com el coeficient de fiabilitat de la meitat del test i a partir d'una fórmula senzilla coneguda com Spearman-Brown, pot ser fàcilment extrapolable per a calcular la fiabilitat del test complet.

Finalment, un dels indicadors de consistència interna del test és el denominat **coeficient alfa de Cronbach** (α; Cronbach, 1951). És important destacar que el valor d'aquest índex reprodueix el coeficient de fiabilitat del test si tots els ítems són paral·lels. Malgrat que en la pràctica això resulta molt difícil, té sentit aplicar-ho per a valorar el grau de covariació global dels ítems de la prova, el grau en què estan mesurant una única dimensió. Quan α té un valor elevat (proper a 1) els ítems covarien fortament entre si; mentre que quan α presenta

valors baixos (propers a 0) els ítems que componen el test són linealment independents. Ull, és possible que α presenti valors negatius, però des del punt de vista del mesurament, seria totalment inacceptable.

En resum, el coeficient α se sol considerar una *estimació per defecte* del coeficient de fiabilitat. S'ha d'interpretar com un indicador del grau de covariació entre els ítems i és aconsellable complementar-ho amb altres tècniques estadístiques abans d'interpretar-ho com una mesura de la dimensionalitat del test. El coeficient α és probablement l'indicador de fiabilitat més utilitzat, no obstant això, la discussió sobre com s'ha d'interpretar segueix generant polèmica entre els professionals del mesurament. Amb vista als professionals aplicats, possiblement el criteri més rellevant per a la seva interpretació sigui l'ús proposat per al test que estiguem considerant. En aquest sentit, un coeficient alfa de 0,65 pot indicar una alta consistència interna si la prova té només 5 ítems. No obstant això, aquesta alta consistència interna no legitima el seu ús, perquè la precisió d'aquesta mesura serà clarament insuficient. Quan la mesura s'utilitzi amb finalitats diagnòstiques o per prendre decisions que impacten en la vida de les persones, el valor de fiabilitat (de precisió) que hem d'exigir al coeficient alfa hauria de ser superior a 0,90. És important acabar aquesta breu exposició indicant que hi ha molts altres indicadors relacionats amb la consistència interna del test, però la majoria d'aquests rarament són aplicats a la pràctica. El lector interessat a aprofundir en aquests aspectes pot trobar una guia excel·lent en els manuals d'Abad i altres (2011) o Martínez-Arias i altres (2006).

4. Les evidències de validesa com a criteri de qualitat de la mesura

En la nostra opinió, aquesta és la propietat més important que ha de tenir un instrument de mesurament. Una vegada que es té una certa perícia desenvolupant instruments, aconseguir que els mateixos posseeixin índexs de fiabilitat (com els presentats en la secció anterior) que resultin acceptables és –creieu-me- relativament senzill. Fins i tot quan el professional empra aparells i instruments amb un cert grau de sofisticació, els mateixos solen ser bastant precisos (fiables). No obstant això, la qüestió de la validesa que analitzarem aquí està relacionada amb la utilització adequada de l'instrument per a la finalitat que nosaltres pretenem.

Exemple

Si ens prenem la temperatura corporal amb un termòmetre senzill, com el que la majoria de nosaltres pot tenir a casa, en dues ocasions consecutives, sense que passi res de temps entre l'una i l'altra, proporcionarà dos mesuraments que tendiran a ser idèntics (excepte unes desviacions molt petites). No ens equivocarem si sobre la base d'aquest tipus d'evidències mantenim que el termòmetre de casa nostra és un instrument fiable per a mesurar la temperatura. Ara bé, qualsevol temperatura? el nostre termòmetre precís seria un instrument vàlid (adequat) per a mesurar la temperatura dels casquets polars? Confio que ningú no emetrà una resposta afirmativa a aquestes qüestions.

Això és precisament el que es persegueix amb la validesa: disposar de proves empíriques que estiguem realment mesurant el que volem mesurar i que, a més, aquesta mesura és adequada en el context en què nosaltres la volem utilitzar i per a la finalitat amb què volem fer-ho.

Per tant, podríem dir que la validesa és l'aspecte del mesurament vinculat amb la comprovació i l'estudi del significat de les puntuacions obtingudes pels tests. D'acord amb una orientació marcadament empírica de la logopèdia, podem centrar el seu estudi en l'examen de les variables definides en i pel test, i en les seves relacions amb les variables externes, observades o latents, amb l'objectiu de sustentar les interpretacions proposades (Elosua, 2003). Des d'una de les primeres definicions sobre la validesa aportada per Guildford (1946) quan estableix que, d'una forma molt general, un test és vàlid per a allò amb què es correlaciona, fins a una de les últimes formulada per Messick (1989) quan afirma que «la validació d'un test abasta totes les qüestions experimentals, estadístiques i filosòfiques per mitjà de les quals s'avaluen les hipòtesis i les teories científiques» (pàg. 14), hi ha més de quaranta anys de diferència, període en què es produeix una gran evolució sobre tot el relacionat amb el mesurament. És precisament la definició de Messick (1989), la que continua sent àmpliament acceptada per la comunitat científica i la que va servir de marc de referència per als estàndards de 1999, que continua vigent a la versió dels estàndards del 2014.

En aquesta última edició dels estàndards (APA, AERA i NCME, 2014) es defensa que el que es valida, per tant, no és un test sinó les inferències fetes a partir del mateix. En definitiva, s'entén la validació com un procés unitari en continu estat de revisió i –fonamental– amb un marc teòric de referència. En aquesta obra es va a subratllar la importància del concepte «ús proposat per al test», i sens dubte, el seu objectiu final (gairebé tan ambiciós com simple en la seva formulació) seria dotar els tests d'avals, tant científics com ètics. En paraules de Messick:

> «Una visió integradora de la validesa (...) ha de distingir dues facetes interconnectades del concepte unitari de validesa. Una faceta és la font de justificació (...). L'altra faceta és la funció o resultat del test.»
>
> (Messick, 1989, pàg. 20)

Per tant, per a referir-nos a la validesa d'una prova, ja no serà suficient una justificació substantiva de les puntuacions, sinó que és necessari delimitar els fonaments teòrics en un context extern en relació amb el propòsit o interpretació proposada.

En resum, la validació pot ser vista com el desenvolupament i l'adquisició d'un cúmul d'arguments científics sobre la validesa per a mantenir la interpretació proposada de les puntuacions d'un test i la seva rellevància per a l'ús que es proposa (APA, AERA i NCME, 2014). D'aquesta manera, tal com reflecteixen els últims estàndards, la validesa és l'aspecte més rellevant tant en el desenvolupament com en l'avaluació de la qualitat dels tests. No obstant això, malgrat la consideració de la validesa com un concepte unitari, la tradicional trinitat (contingut, constructe i criteri) se segueix mantenint temps després, encara que no com a tipus de validesa sinó com a estratègies de validació o evidències de validesa (Schuler i Guldin, 1991).

4.1. Evidències relacionades amb el contingut

Quan s'analitzen les evidències basades en el contingut de la prova es persegueix determinar el grau en què els seus ítems representen el domini dels continguts o de les conductes de la variable que es pretén mesurar (Wernimont i Campbell, 1968). La **validesa del contingut** també ha rebut els noms de «validesa lògica» o «validesa per definició» (Anastasi, 1954) i, en aquest cas, no se sol tractar d'un concepte estadístic, sinó que depèn dels judicis que els experts facin sobre la pertinència dels ítems per a capturar la variable d'interès.

L'anàlisi de la validesa del contingut d'un test se centra fonamentalment en dos aspectes: la rellevància dels ítems i la representativitat del test. En aquest sentit, els ítems d'un test són considerats una mostra del domini que interessa avaluar i el que es comprovarà és *si són tots els que estan* (rellevància dels ítems) i si *estan tots els que són* (representativitat del test).

Tradicionalment, la rellevància dels ítems és jutjada per un grup d'experts mitjançant un procediment estructurat que els permet aparellar aquests amb el domini que, en la seva opinió, s'està avaluant. Això requereix una definició prèvia del domini, que determinarà en gran manera les àrees de contingut que ha de cobrir la prova. Per a quantificar la rellevància de cada ítem es pot utilitzar una escala d'1 a 5, on el 5 representaria un ajust perfecte de l'ítem a la seva variable corresponent i l'1, la falta d'ajust, i utilitzar la mitjana de les puntuacions donades per diversos jutges experts per a definir la rellevància de cada ítem.

D'altra banda, la representativitat del test pot definir-se com la precisió amb què es podrien realitzar inferències sobre la puntuació o el nivell de cada subjecte en la variable d'interès a partir de la seva puntuació en el test. En la pràctica, aquesta representativitat es pot establir com el grau en què els continguts de l'àrea d'interès queden coberts pels ítems que figuren en el test. Per a quantificar la representativitat, es pot adoptar un criteri similar al plantejat per a quantificar la rellevància. Es pot utilitzar una escala, per exemple, de 0 a 10, on 0 significa que els ítems considerats en el test representen molt malament en el seu conjunt la característica que es pretén avaluar, mentre que el 10 significa que els ítems que preveu el test representen a la perfecció la característica que es pretén avaluar. A partir d'aquestes dades, es pot utilitzar la mitjana de les puntuacions donades pels jutges.

4.2. Evidències relacionades amb l'estructura interna de la prova

Ja vam comentar anteriorment que la validació d'un test implica l'obtenció de proves a favor del constructe d'interès, i també la demostració que el test és un instrument adequat per a mesurar aquest constructe. Tradicionalment, aquests tipus d'estudis s'engloben en el que es coneix com a validesa del constructe. Des del nostre enfocament, la validació del constructe d'un test es pot abordar des de dos punts de vista: un d'extern (les relacions d'un test amb altres mesures) i un altre intern (relacions entre els ítems d'un test). És aquest últim punt de vista a què ens referirem en aquest apartat. Així, la validesa de constructe interna es refereix al grau en què les relacions entre els ítems reprodueixen l'estructura hipotetitzada de dimensions o factors en un test.

En el marc de la Teoria dels tests (Muñiz, 2010), l'estudi de la **unidimensionalitat** dels constructes avaluats amb cadascuna de les escales d'un test o qüestionari té un lloc absolutament prioritari. Així, per exemple, alguns dels estadístics tradicionals que s'han presentat en aquest capítol, com són l'índex *de dificultat* (D_j) o l'índex *d'homogeneïtat* (H_j), tenen sentit únicament si s'està mesurant un sol atribut. Però, sense cap dubte, el punt més crític a l'hora de valorar la importància concedida a l'estudi de la unidimensionalitat es pot situar en l'obtenció de les puntuacions que seran assignades als subjectes en el procés d'avaluació. L'obtenció de les puntuacions globals en un test per mitjà de la suma de les puntuacions en els ítems implica, de manera immediata, l'assumpció que s'està mesurant el mateix constructe amb tots aquests, en cas

contrari aquesta suma mancaria de sentit. És a dir, quin sentit tindria sumar puntuacions de preguntes que estan mesurant coses diferents que no tenen res a veure? i aquesta puntuació que resultés, seria interpretable? La veritat és que difícilment.

Un problema que es presenta quan es recorre a la literatura específica a la recerca d'una definició acceptada d'unidimensionalitat és que molts dels autors que proposen alguna definició per a aquesta propietat ho fan en funció del tipus de prova que estan pensant utilitzar per a avaluar-la. El reflex d'aquesta diversitat es pot trobar fent un repàs de les diferents definicions d'unidimensionalitat.

Aquí, per la seva àmplia accepció i extensió, proposarem al lector l'aproximació a la unidimensionalitat basada en les teories del tret latent, que l'operativitzen com l'existència d'un únic tret (o factor) subjacent a les respostes dels subjectes en un ítem. Dins d'aquesta aproximació, Ackerman (1992) situa la dimensionalitat d'un test en la interacció que es produeix entre els subjectes i els ítems del test. Aquesta interacció pot ser unidimensional de tres formes diferents:

1) Un ítem pot necessitar la utilització de diverses destreses per a obtenir una resposta concreta, però si els subjectes varien únicament en una de les destreses o en la mateixa combinació d'aquestes, aquesta interacció pot ser modelada unidimensionalment.

2) Si els ítems només mesuren una dimensió és igual que els subjectes variïn en diverses dimensions, la interacció és de nou unidimensional.

3) El cas extrem d'un test format per un únic ítem.

Des d'aquesta perspectiva, la tècnica d'anàlisi més freqüentment utilitzada és l'anàlisi factorial, sia en el seu vessant exploratori o confirmatori. En essència, es tracta d'un procediment estadístic que permet reduir la dimensionalitat, és a dir, a partir d'un gran nombre de variables (els ítems) trobar un nombre molt menor de dimensions que els representin (factors). Ja no parlarem més aquí sobre aquest procediment que està en continu desenvolupament i evolució des de fa més de 100 anys i segueix actualment en desenvolupament mentre escrivim aquestes línies.

4.3. Evidències basades en la relació amb altres variables

Des d'aquest punt de vista, es persegueix trobar evidències de validesa per a les escales d'un instrument basant-nos en el grau de relació d'aquestes escales amb altres mesures externes. Aquestes altres mesures externes poden ser:

1) Altres escales i altres constructes (en què les correlacions reflecteixen les relacions implicades en la teoria dels constructes analitzats).

2) Altres mesures externes considerades com a criteri sobre les quals ens interessa comprovar el poder predictiu de les escales que componen el nostre test d'interès.

Per a analitzar la validesa de constructe externa emprant com a criteri les mesures dels mateixos i diferents constructes, s'han proposat diferents procediments i tècniques estadístiques. Un dels més emprats és la utilització de *matrius multitret-multimètode*. En realitat, la matriu multitret-multimètode és una matriu de correlacions entre diferents trets mesurats per mitjà de diferents mètodes, per la qual cosa és possible analitzar dos aspectes diferents sobre aquesta matriu: la **validesa convergent** i la **validesa discriminant**. El primer d'aquests aspectes es refereix al grau d'acord entre les múltiples mesures d'un mateix tret o atribut; és a dir, dos o més mesures d'un mateix constructe seran vàlides si la seva correlació és elevada. El segon d'aquests aspectes es refereix al grau en què les mesures de diferents trets són diferents; és a dir, dos o més constructes són únics si les mesures de cadascun d'aquests no es correlacionen massa.

D'altra banda, hi ha altres estudis que tracten de recaptar l'evidència empírica per al procés de validació basada en les relacions de les variables mesurades amb el nostre instrument amb unes mesures externes que es denominaran criteris. Hi ha dos tipus de **validesa referida al criteri**: la *validesa concurrent* i la *validesa predictiva*. La diferència fonamental entre tots dos tipus de validesa és el moment en què els investigadors obtenen la mesura del criteri. Es parla de validesa concurrent quan s'obtenen simultàniament les mesures de les variables predictores i del criteri. D'altra banda, es parla de validesa predictiva quan primer s'obtenen les mesures de les variables predictores i, després, les del criteri. El model de la validesa concurrent ha estat molt criticat i molts autors consideren que, des d'un punt de vista teòric, els resultats que ofereix són inaplicables al model de la validesa predictiva. No obstant això, en recerques empíriques amb prou feines s'han observat diferències entre els coeficients de validesa obtinguts en tots dos tipus de disseny, la qual cosa justifica que segueixi sent el més emprat en la pràctica, a més dels avantatges de temps i cost econòmic que presenta el procediment de la validesa concurrent.

5. Un cas per a futurs logopedes

Són moltes les recerques que han trobat resultats que suggereixen que els problemes en la lectura de paraules i una pobra competència lingüística s'associen amb dificultats en la comprensió lectora. Això no és nou. No obstant això, la Cristina, una brillant logopeda recentment graduada està interessada per aprofundir a l'apassionant món de les dificultats en la comprensió lectora. Decideix realitzar una cerca bibliogràfica. Després de començar a llegir el que se sap sobre aquest tema, aviat s'adona que el paper que poden arribar a jugar els processos cognitius sobre la comprensió lectora ha estat escassament explorat. La Cristina radia de felicitat!!! Acaba de trobar un tema per a la seva futura tesi doctoral.

El primer que decideix fer és definir el seu criteri (la comprensió lectora), i pensar en com pot arribar a avaluar-la. Recorda que en el seu moment va llegir en un excel·lent manual (com el mòdul que té ara mateix entre les mans), que no era recomanable realitzar un diagnòstic o una avaluació basant-se exclusivament en els resultats d'un únic test. Per aquesta raó, decideix avaluar la comprensió a partir de dues proves: El test d'estratègies de comprensió (TEC; Vidal-Abarca i altres, 2007), i la prova de comprensió de textos del test PROLEC-SE (Ramos i Cuetos, 2003). El coeficient de fiabilitat alpha de Cronbach del primer és de 0,79 i el del segon, de 0,77. I, a més, la correlació que se sol obtenir entre les puntuacions d'aquestes dues proves és de 0,86.

A continuació, i seguint un procediment similar a l'establert per al criteri, la Cristina selecciona diferents tests i mesures per a identificar cadascun dels factors cognitius que emprarà com a potencials predictors del seu criteri variable. Entre els factors cognitius que la seva revisió bibliogràfica la porta a plantejar hi ha: la capacitat de memòria, l'accés al lèxic, la velocitat de processament i la capacitat d'atenció. Per a avaluar aquests aspectes, la Cristina tria tres mesures per a cada dimensió, és a dir, un total de 12 mesures. Després d'aquest procés, i molt satisfeta amb el seu treball de moment, se'n va a parlar amb els responsables educatius de diversos instituts de la seva ciutat, Barcelona, perquè la deixin accedir a una mostra que resulti representativa i que tingui una grandària suficient de participants. És tan persistent i està tan motivada que acaba tenint dades per a totes les variables (predictores i criteri) en una mostra de 1.200 éssers humans.

Abans de plantejar l'anàlisi principal, la Cristina vol estar segura que les coses empíricament van com la teoria dicta que han d'anar. Per a això és necessari comprovar, per exemple, que cadascuna de les mesures avalua el que se suposa que mesura i no altres factors cognitius. Així, decideix realitzar una anàlisi factorial exploratòria (primer per a cada prova a nivell d'ítems) i després amb

les puntuacions de les 12 mesures cognitives empleades i, tot en ordre, és a dir, s'identifiquen exactament els seus quatre factors cognitius d'interès (els seus quatre constructes predictors!).

Finalment, la Cristina està en el moment més àlgid de la seva recerca, això és, està en disposició d'esbrinar si els quatre factors cognitius avaluats tenen la capacitat predictiva sobre el seu criteri variable (la comprensió lectora). S'asseu davant de l'SPSS, selecciona l'anàlisi adequada, introdueix les variables i marca les opcions pertinents, està emocionada i amb la pell de gallina, sospira, recorda el moment en què, recentment graduada, va decidir emprendre aquest viatge. Han passat diversos anys de treball dur i ara està a un sol clic d'obtenir les respostes a les seves preguntes. Per cert, parlant de preguntes...

Activitats

1. Et sembla adequada l'opció de la Cristina a l'hora de definir el seu criteri (comprensió lectora)? Per què?

2. Quantes evidències de validesa diferents estan implicades en la recerca que ha desenvolupat la Cristina?

3. En el cas de la validesa referida al criteri, segons el disseny descrit, es tractaria d'una validesa concurrent o d'una validesa predictiva?

4. Si la Cristina et demanés ajuda per a millorar el seu treball, quines altres evidències empíriques li proposaries recaptar o què faries d'una manera diferent? Per què?

5. Vols ser com la Cristina? Si la resposta és que sí, pots començar llegint aquest article:

Lectura recomanada

Casas, A. M.; Andrés, M. I. F.; Castellar, R. G.; Mínguez, R. T. (2011). «Factores que predicen las estrategias de comprensión de la lectura de adolescentes con trastorno por déficit de atención con hiperactividad, con dificultades de comprensión lectora y con ambos trastornos». *Revista de Logopedia, Foniatría y Audiología* (núm. 31, vol. 4, pàg. 193-202).

Bibliografia

Abad, F. J.; Olea, J.; Ponsoda, V.; García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Ackerman, T. A. (1992). «A didactic explanation of item bias, item impact and item validity from a multidimensional perspective». *Journal of Educational Measurement* (núm. 29, pàg. 67-91).

American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014). *Standards for educational and psychological testing* (5a. ed.). Washington DC: American Educational Research Association.

Anastasi, A. (1954). *Psychological Testing*. Oxford: Macmillan.

Balluerka, N.; Gorostiaga, A.; Alonso-Arbiol, I.; Haramburu, M. (2007). «La adaptación de instrumentos de medida de unas culturas a otras: una perspectiva práctica». *Psicothema* (núm. 19, pàg. 124-133).

Borsboom, D. (2006). *Measuring the mind*. Nova York: Cambridge University Press.

Downing, S. M.; Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: LEA.

Elosua, P. (2003). «Sobre la validez de los tests». *Psicothema* (núm. 15, vol. 2, pàg. 315-321).

Gómez, J.; Hidalgo, M. D.; Guilera, G. (2010). «El sesgo de los instrumentos de medición. Tests justos». *Papeles del Psicólogo* (núm. 31, vol. 1, pàg. 75-84).

Guildford, J. P. (1946). «New standards for test evaluation». *Educational and Psychological Measurement* (núm. 6, pàg. 427-439).

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3a. ed.). Mahwah, NJ: LEA.

Haladyna, T. M.; Downing, S. M.; Rodriguez, M. C. (2002). «A review of multiplechoice item-writing guidelines for classroom assessment». *Applied Measurement in Education* (núm. 15, vol. 3, pàg. 309-334).

Lord, F. M.; Novick, M. R. (1968). *Statistical theories of mental test scores*. Nova York: Addison-Wesley.

Martínez Arias, R.; Hernández-Lloreda, M. J.; Hernández-Lloreda, M. V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Messick, S. (1989). «Validity». A: R. Linn (ed.), *Educational Measurement* (3a. ed). Nova York: Macmillan (pàg. 13-104).

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

Muñiz, J. (1996). *Psicometría*. Madrid: Universitas.

Muñiz, J. (1998). «La medición de lo Psicológico». *Psicothema* (núm. 10, vol. 1, pàg. 1-21).

Muñiz, J. (2010). «Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems». *Papeles del psicólogo* (núm. 31, vol. 1, pàg. 57-66).

Muñiz, J.; Martínez, R.; Moreno, R.; Fidalgo, A. M.; García Cueto, E. (2005). *Análisis de los Ítems*. Madrid: La Muralla.

Navas, M. J. (2001). *Métodos, diseños y técnicas de investigación psicológica*. Madrid: UNED.

Nunnally, J. C.; Bernstein, I. (1994). *Psychometric Theory* (3a. ed.). Nova York: McGraw Hill.

Olea, J.; Ponsoda, V. (2001). *Tests adaptativos informatizados*. Madrid: UNED.

Ramos J. L.; Cuetos, F. (2003). *Batería de evaluación de los Procesos Lectores en los alumnos del tercer ciclo de Educación Primaria y Educación Secundaria Obligatoria, PROLEC SE*. Madrid: TEA Ediciones.

Roid, G.; Haladyna, T. M. (1982). *Technology for test-item writing*. Nova York: Academic Press.

Schuler, H.; Guldin, A. (1991). «Methodological issues in personnel selection research». *International review of industrial and organizational psychology* (núm. 6, pàg. 213-264).

Streiner, D. L. (2003). «Starting at the beginning: An introduction to coefficient alpha and internal consistency». *Journal of personality assessment* (núm. 80, vol. 1, pàg. 99-103).

Vidal-Abarca, E.; Gilabert, R.; Martínez, T.; Sellés, P.; Abad, N.; Ferrer, C. (2007). *Test de Estrategias de Comprensión (TEC)*. Madrid: Instituto Calasanz de Ciencias de la Educación.

Wernimont, P. F.; Campbell, J. P. (1968). «Signs, samples, and criteria». *Journal of Applied Psychology* (núm. 25, pàg. 372-376).