Validesa

Luis Manuel Lozano
Jaume Turbany

PID_00198602

Universitat Oberta de Catalunya

Índex

Introducció

Quan un psicòleg decideix aplicar un qüestionari és per a aconseguir un objectiu determinat. Per a això s'ha d'assegurar que el qüestionari que usarà té unes propietats psicomètriques adequades. Entre aquestes, cal destacar la propietat a la qual es fa referència en aquest mòdul: la validesa.

La validesa és un dels aspectes més importants, potser el que més, tant en l'elaboració com en l'avaluació de qüestionaris psicològics. Al cap i a la fi es tracta de comprovar que la utilització del test és correcta i que els objectius que vol assolir el psicòleg que l'utilitza són factibles.

En l'apartat "Què és la validesa?" es fa un breu recorregut històric sobre aquest concepte. Com es pot observar, és un concepte que ha estat evolucionant (i encara ho fa) fins que ha arribat a la idea que actualment està en vigor. Aquest concepte es defineix oficialment en els Standards publicats el 1999 conjuntament per l'American Educational Research Methods (AERA), l'American Psychological Association (APA) i el National Council on Measurement in Education (NCME). Aquestes entitats defensen que es poden agrupar els indicis de validesa d'un test en cinc apartats: evidència basada en la validesa de contingut, basada en el procés de resposta, basada en l'estructura interna del qüestionari, basada en la relació amb altres variables i basada en les conseqüències de l'avaluació.

En els apartats següents es veu cadascun dels indicis de validesa prèviament esmentats i es busquen estratègies per a poder obtenir aquests indicis (a excepció de l'apartat de les conseqüències de l'avaluació, en el qual només es tracta de les conseqüències que es poden esperar de l'aplicació d'un qüestionari).

En l'últim apartat, "Factors que afecten la validesa", es veuen diferents aspectes que afecten algunes de les tècniques exposades amb anterioritat per a determinar els diferents indicis de validesa.

1. Què és la validesa

1.1. Definició

Per a comprendre el concepte de *validesa* cal fer un petit estudi de l'evolució històrica que ha experimentat aquest concepte.

La utilització de qüestionaris es va veure impulsada per la Primera i la Segona Guerra Mundial. En aquells moments es va tenir la necessitat d'incorporar la població civil a l'exèrcit, destinant cada persona al lloc més adequat per a ella. Després d'emplenar els qüestionaris es comprovava al camp d'entrenament si els subjectes rendien satisfactòriament o no al lloc on havien estat destinats. Atès que en primer lloc es feia el mesurament i posteriorment s'avaluava l'èxit, es parlava de validesa predictiva. És a dir, un test té **validesa predictiva** si serveix per a predir el comportament en un constructe que serà avaluat després d'aplicar el qüestionari.

Posteriorment, es va tractar d'avaluar la relació entre les característiques de les persones que feien un treball i el seu èxit en aquest. D'aquesta manera, es tractava de saber quines característiques podrien predir l'èxit laboral i buscar-les quan es feia una selecció de personal. Atès que l'estudi s'efectuava sobre persones que ja tenien el lloc i se'n valorava l'execució, es parlava de validesa concurrent, ja que tots dos mesuraments es feien alhora. És a dir, un test té **validesa concurrent** si serveix per a predir el comportament en un constructe que és avaluat simultàniament a l'aplicació del qüestionari.

Com es pot observar, inicialment els tests s'empraven exclusivament per a predir. Així doncs, en un començament, es considerava que un test era vàlid si servia per a predir alguna variable d'interès denominada ***criteri*** (Guilford, 1946).

Per tant, es conceptualitza la validesa com a correlació entre el qüestionari i el criteri d'interès (tant si aquest és avaluat amb posterioritat com simultàniament a l'aplicació del qüestionari). Així doncs, es considera que un test és vàlid per a avaluar qualsevol aspecte amb el qual es correlacioni (Bingham, 1937; Guilford, 1946; entre d'altres).

Un dels problemes de la conceptualització de la validesa com a correlació és el fet que cal trobar una mesura del criteri adequada, és a dir, es necessiten dades del criteri que s'hagin obtingut d'una manera fiable i vàlida. Per tant, si ja es disposa d'una mesura vàlida del criteri per a què es necessita aplicar un qüestionari?

Un altre problema d'aquesta conceptualització és que deixava fora un gran nombre de tests educatius. En aquests no es tracta de predir la conducta, es tracta de comprovar quant s'ha après després d'un període de formació. En aquests qüestionaris la puntuació obtinguda és un indicador del que el test vol avaluar (coneixement en matemàtiques, en anglès, etc.) i no un predictor de criteris diferents del test. Des d'aquesta perspectiva, la validesa fa referència al fet que els ítems que componen el qüestionari siguin representatius d'allò que es vol avaluar. Aquest concepte es va denominar ***validesa de contingut*** (Anastasi, 1954).

D'altra banda, al llarg dels anys trenta hi ha un auge de les teories que tracten de conèixer l'estructura factorial de la intel·ligència. Amb aquestes teories es comença a conceptualitzar un test com a vàlid quan representa de manera fidedigna el constructe psicològic que vol mesurar, com també les relacions esperades entre els diferents constructes. D'aquesta manera neix la validesa de constructe (Cronbach i Meehl, 1955). Les tècniques estadístiques emprades per a poder comprovar aquesta validesa són, tradicionalment, l'anàlisi factorial exploratòria i les matrius multitret-multimètode (Campbell i Fiske, 1959), i més recentment l'anàlisi factorial confirmatòria. Per exemple, si s'empra un test que avalua la tríada cognitiva des del model cognitiu de depressió de Beck (Beck, Rush, Shawn i Emery, 1979) (pensaments sobre un mateix, pensaments sobre el món i pensaments sobre el futur), el qüestionari tindrà **validesa de constructe** si avalua les tres dimensions i aquestes tenen les relacions que s'esperen, per exemple, amb ansietat.

Fins als anys vuitanta es podia parlar de validesa predictiva, validesa concurrent, validesa de contingut i validesa de constructe d'un qüestionari. Si bé les dues primeres en els estàndards dels tests i manuals educatius i psicològics publicats per l'APA, AERA i NCME el 1966 i el 1974 s'englobaven com a **validesa de criteri**.

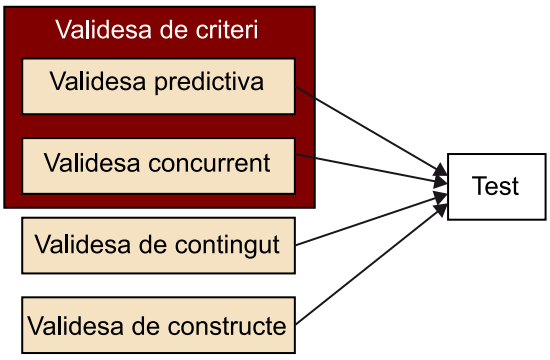

Figura 1

Posteriorment, Cronbach (1971) va puntualitzar que en un test que vol mesurar un tret de personalitat no hi ha només un criteri rellevant per a predir ni un contingut per a mostrejar (validesa predictiva i de contingut respectivament). Per contra, es disposa d'una teoria sobre el tret i sobre les seves relacions amb altres constructes i variables (validesa de constructe). Si se suposa que

la puntuació del test és una manifestació vàlida de l'atribut, es pot contrastar l'assumpció analitzant les seves relacions amb altres variables. Per tant, va començar a haver-hi una tendència que considerava la validesa com una cosa unitària, essent la validesa de constructe la científicament més admissible i estant la validesa de criteri i de contingut incloses en aquesta (Messick, 1989). Així doncs, s'imposa la concepció que la validació de constructe constitueix un marc integral per a obtenir proves de la validesa incloent-hi les procedents de la validació de criteri i de contingut. De fet, es deixa de parlar de les diferents categories de validesa i es comença a parlar de diferents evidències implicades en els tres tipus tradicionals de validesa (criteri, contingut i constructe).

Atès que tant l'estudi de l'estructura del constructe com les relacions d'aquest amb altres constructes es passa a considerar la forma principal de validesa, aquest procés es pot concebre com un cas particular de la contrastació de les teories científiques mitjançant el mètode hipoteticodeductiu (Prieto i Delgado, 2010).

Fixeu-vos que en aquests moments, a mitjan anys vuitanta, hi ha un canvi molt rellevant. Mentre que al començament es conceptualitza la validesa com una propietat inherent al test, es passa a concebre que el que realment es valida no és el test en si mateix, sinó les inferències que es fan a partir d'aquest. Per això, el responsable d'assegurar la validesa ja no és només el constructor del test, sinó que també ho és l'usuari que empra aquest qüestionari per a una finalitat determinada. Sovint, els problemes d'un qüestionari sobre la validesa es deuen no al disseny del qüestionari, sinó a l'ús que se'n fa.

Actualment, en l'última edició fins ara dels *Standards for educational and psychological testing* (AERA, APA, NCME, 1999), molt influenciats pel capítol escrit per Messick (1989) i el llibre de Shepard, Camilli, Linn i Bohrnstedt (1993), es defensa que la validesa fa referència al **grau en què l'evidència empírica i la teoria donen suport a la interpretació de les puntuacions dels tests relacionada amb un ús específic**. Com es pot apreciar, la validesa es concep com un concepte unitari. Per a comprovar la validesa s'ha d'atendre a cinc evidències de validesa:

- **El contingut de test**: els ítems que constitueixen el test són rellevants i representatius del constructe psicològic que es vol mesurar.

- **El procés de resposta**: el procés que segueixen les persones en contestar el test permet extreure respostes indicadores del que es vol avaluar.

- **L'estructura interna**: les relacions dels ítems entre ells són congruents amb el model teòric emprat a l'hora de definir el constructe que es vol avaluar.

- **La relació amb altres variables**: les relacions que s'estableixen entre el constructe que s'avalua i altres constructes són les esperades segons el marc teòric en el qual s'ha definit el constructe a avaluar.

- **Les conseqüències de l'aplicació del qüestionari**: les conseqüències tant positives com negatives que ha d'emprar un test són les previstes.

Com a breu resum del que s'ha exposat més amunt, es presenta la taula següent, en la qual es pot apreciar l'evolució del concepte en els diferents estàndards publicats per l'APA.

Taula 1

Edició	Validesa
1954	Constructe, concurrent, predictiva, contingut
1966	Criteri, constructe, contingut
1974	Criteri, constructe, contingut
1985	Unitària (però mantenen criteri, constructe i contingut)
1999	Unitària: 5 fonts d'evidència

1.2. Importància de la validesa

El concepte de *validesa* és central en la psicometria. Tal com s'ha comentat anteriorment, per a comprovar la validesa s'han d'acumular evidències que proporcionin una base científica per a interpretar les puntuacions d'un qüestionari de manera adequada. Per això el que realment es valida no és el qüestionari en si mateix, són les interpretacions que es fan a partir d'aquest. Per tant, no es pot defensar que un test sigui vàlid o per contra manqui de validesa. Un test pot ser adequat per a un propòsit però no per a un altre.

> Si s'aplica un qüestionari amb el qual es vol mesurar l'autoestima, les respostes es poden emprar amb diferents finalitats (conèixer el nivell d'autoestima d'una persona per a saber si és un problema que s'ha de tractar en teràpia, en selecció de personal, com a recerca sobre el propi constructe, etc.). Per a poder usar el qüestionari amb una finalitat determinada, s'han d'acumular evidències que indiquin que l'ús és correcte (evidències de validesa). En cas contrari, s'estaria fent un mal ús dels tests, eines principals en el treball psicològic, i les conclusions que se'n traguessin no serien correctes. En l'exemple anterior no se sabria si és un aspecte sobre el qual s'ha d'intervenir terapèuticament, no se sabria si la persona seleccionada realment té l'autoestima que es desitja o no se sabria si realment s'està mesurant autoestima.

Per a poder fer el treball correctament com a psicòlegs, s'ha de saber si les conclusions que es treuen a partir dels tests emprats són adequades, ja que, en cas contrari, es corre el risc de no saber exactament què s'està avaluant o si aquest mesurament realment és útil per al propòsit del psicòleg.

2. Evidència de validesa basada en el contingut

2.1. Concepte

Moltes de les inferències i assumpcions que es deriven de la interpretació de les puntuacions en un test es poden avaluar més fàcilment si s'examinen els procediments que s'han emprat per a generar les puntuacions. Per exemple, si es vol inferir alguna cosa a partir de les puntuacions en un test sobre determinada conducta o constructe psicològic, se suposa que els ítems que componen el qüestionari són tan **rellevants** (que la informació que es demana està directament relacionada amb el que es vol mesurar) com **representatius** (les qüestions que es presenten han de ser una mostra adequada de tot el que es vol mesurar) d'aquesta (Kane, 2006).

L'evidència de la validesa de contingut fa referència a la relació que hi ha entre els ítems que componen el test i el que es vol avaluar amb aquest, parant esment tant a la rellevància com a la representativitat dels ítems. Aquest tipus d'evidència es recull principalment en el moment de l'elaboració del test.

Suposem que es vol elaborar un test per a avaluar la personalitat. En aquest cas, es decideix treballar dins del marc teòric dels cinc factors de la personalitat (extraversió, obertura, responsabilitat, amabilitat i neuroticisme). Atès que es tracta d'un test que s'emprarà en una selecció de personal concreta, només interessen les dimensions de responsabilitat (a), amabilitat (b) i neuroticisme (c). En aquest exemple, el constructe és la personalitat que està composta per les cinc dimensions. Les dues primeres, per als interessos del test que es planteja, són informació irrellevant. Les altres tres són el domini que interessa avaluar. A partir d'aquest domini es construeixen ítems destinats a avaluar la responsabilitat (a'), l'amabilitat (b') i el neuroticisme (c'). Aquests ítems han de tenir relació amb el factor que volem mesurar, és a dir, els ítems que avaluen responsabilitat estan relacionats amb la definició que existeix en la comunitat científica sobre aquest factor (rellevància). Però, al seu torn, els ítems han de preguntar per la totalitat del domini que s'ha d'avaluar (representativitat).

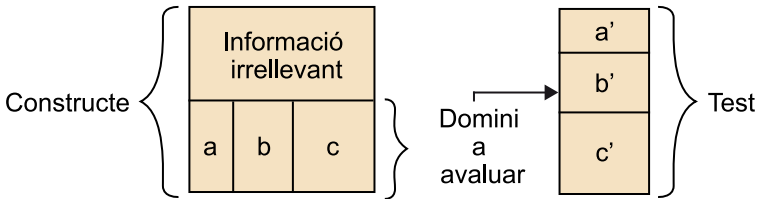

Figura 2

En les proves educatives, les evidències de validesa basada en el contingut són fonamentals. Si no es comprova que el test és consistent amb els objectius curriculars que es persegueixen (rellevància), és a dir que està lliure de material irrellevant i que el que hi ha representa adequadament el domini que es vol avaluar (representativitat), la utilitat del test es veurà seriosament afectada i, per tant, les conclusions que s'obtinguin a partir d'aquest seran errònies. En aquestes situacions, atès que el domini que es vol avaluar està clarament defi-

nit, se sol recomanar emprar els diferents mètodes estadístics de mostreig per a obtenir una mostra representativa dels continguts que han de constituir el test (Muñiz, 2003).

El problema sorgeix quan no es disposa del domini tan clarament definit.

> Per exemple, si es vol elaborar un test que avaluï la intel·ligència, el primer que s'ha de preguntar el constructor del qüestionari és què és la conducta intel·ligent. En aquest cas, atès que no hi ha un domini perfectament definit, s'han de buscar altres estratègies per a obtenir l'indicador de la validesa de contingut.

2.2. Procediment

En aquest apartat es presentarà el procediment més habitual en la valoració de l'evidència basada en el contingut, si bé hi ha altres mètodes menys emprats. En podeu trobar una recopilació a Sireci (1998).

Si es vol desenvolupar un test, el primer que s'ha de fer és definir de manera operativa el domini que es vol avaluar. Després d'elaborar-ne una o acceptar-ne una que ja existeix, s'ha d'elaborar una **taula d'especificacions**. Es tracta de fer una descripció detallada del test, determinar la proporció o el nombre d'ítems que avaluaran cada contingut o habilitat del domini que es vol avaluar i determinar el format dels ítems i de les respostes (AERA, APA, NCME, 1999) (usualment en aquest pas també es determinen les propietats psicomètriques que es vol que tingui la prova).

Després de crear els ítems s'ha d'acudir a un grup d'experts en la matèria que faran de jutges. Per a evitar qualsevol biaix, aquests jutges no han d'estar implicats en l'elaboració del qüestionari, sinó que han d'analitzar cadascun dels ítems valorant en quina mesura són **representatius** i **rellevants** per a avaluar el domini d'interès, prenent com a definició d'aquest l'aportada pels autors del test.

Es pot defensar que hi ha, per tant, tres aspectes ben diferenciats que s'han de tenir en compte a l'hora de comprovar les evidències de la validesa de contingut: la definició del domini, la representació dels ítems que avaluen el domini i la seva rellevància (Sireci, 1998).

És recomanable que cada jutge valori els ítems per separat per a evitar, d'aquesta manera, possibles biaixos a l'hora de respondre. Una vegada es tenen les valoracions de tots els experts, s'han de buscar els ítems en els quals hi hagi concordança i seleccionar-los per a formar part del qüestionari.

Per exemple, si 8 dels 10 jutges determinen que un ítem destinat a mesurar depressió realment avalua allò que vol avaluar, aquest ítem tindrà un índex de congruència de 0,8. Se solen considerar adequats els ítems que tenen un índex de congruència igual o superior a 0,7 (Sireci, 1998).

No cal eliminar els ítems en els quals no hi hagi acord (que no aconsegueixin un índex de congruència de 0,7). És recomanable que amb aquests ítems es faci un grup de discussió amb els experts perquè comentin les diferències tractant d'arribar a un punt d'acord per a millorar aquests ítems.

Aquest és el procediment més habitual a l'hora de valorar els indicis de validesa de contingut, si bé no està lliure de crítiques. El problema principal que es planteja en la utilització d'experts és que aquests són altament competents en el contingut que s'avalua, per la qual cosa poden passar per alt un text amb un nivell que no és adequat per a la comprensió dels subjectes que s'han d'avaluar o que es pot interpretar malament amb facilitat. És a dir, encara que l'expert ens pot proporcionar informació molt rellevant, el que realment importa és com percep el test o l'ítem la persona que hi està responent i com hi reacciona (Leighton, 2004).

2.3. Contingut esbiaixat

L'ús d'experts per a valorar tant la rellevància com la representativitat dels ítems té com a finalitat evitar que el qüestionari tingui continguts esbiaixats. Es diu que el contingut d'un test està esbiaixat si els ítems que el componen avaluen aspectes no rellevants per al domini (**biaix per falta de rellevància**) o si no representen de manera adequada tot el domini a avaluar (**biaix per falta de representativitat**). Com es pot comprovar, un test està esbiaixat si no cobreix adequadament el domini que vol mesurar o si inclou qüestions que no són necessàries per a valorar el domini correctament.

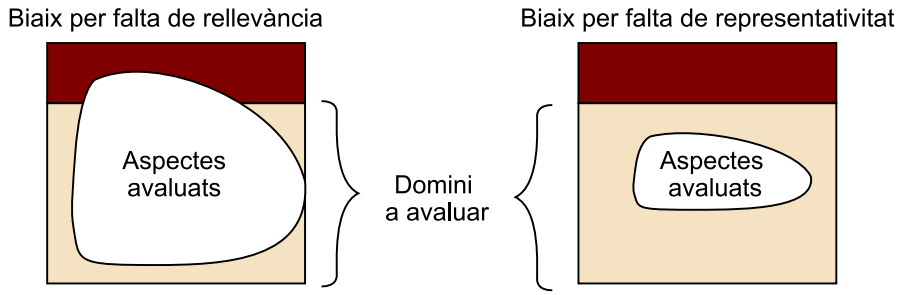

Figura 3

3. Evidència de validesa basada en el procés de resposta

3.1. Concepte

Aquest tipus d'evidència és un concepte que es va introduir com a nou en els *Standards* publicats el 1999, si bé ja havia estat esmentat per alguns especialistes en la mesura del psicològic com Messick (1989). Els *Standards* descriuen aquest tipus d'indicis com l'ajust entre el constructe avaluat i el procés de resposta que han dut a terme les persones que responen al test.

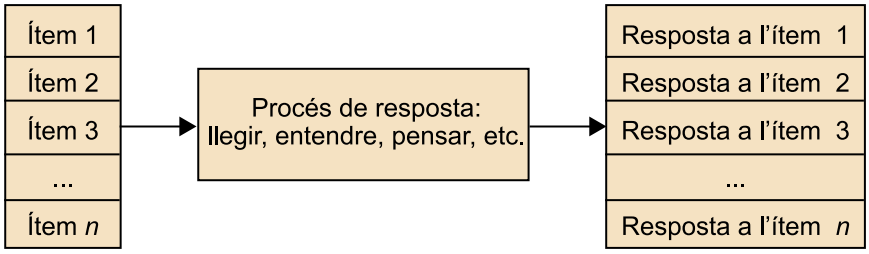

Figura 4

Per **procés de resposta** s'entenen totes les conductes que es necessiten per a poder contestar un ítem, com poden ser llegir les preguntes, comprendre-les, decidir la resposta que es vol donar i finalment respondre a l'ítem.

> Un exemple sobre aquest indici de validesa es pot trobar en un examen de matemàtiques que es faci a nens que estan aprenent a sumar. Si l'enunciat de l'ítem és "3 + 2 =", probablement, si han adquirit el coneixement necessari, poden donar una resposta correcta. Aquest ítem també pot tenir un enunciat com "Calculeu el valor resultant de dur a terme una operació additiva entre els valors 3 i 2". Evidentment, un nen que estigui aprenent a sumar pot respondre al primer enunciat, però no al segon (no té el vocabulari necessari, la seva capacitat lectora no li permetrà comprendre la pregunta, etc.). Com a resultat del primer enunciat es conclourà que ja té adquirit cert nivell del domini avaluat, però amb el segon es conclourà que no. És a dir, atès que el segon enunciat manca de validesa de resposta, portarà els avaluadors a conclusions errònies sobre el nivell d'habilitat del nen a l'hora de fer sumes.

A l'hora de respondre a un test, s'han de combinar tant les característiques de les preguntes com les de les respostes que es poden donar i les de la persona que respon. Per això hi ha diferents factors que poden afectar la resposta:

- **Factors relacionats amb els ítems**. En aquest apartat s'han de tenir en compte diversos factors.
  - Contingut dels ítems. S'ha d'assegurar que el contingut és adequat a la població que es vol avaluar. Per exemple, no es pot usar un test per a avaluar depressió infantil si es va construir per a avaluar-la en adults.

Si es fa, es poden trobar preguntes del tipus "Ha notat canvis en el seu desig sexual?", que evidentment són inadequades per a avaluar nens.

– Redacció dels ítems. El llenguatge emprat per a redactar l'ítem no ha de superar la capacitat comprensiva de les persones que respondran. Un exemple d'això es pot observar en l'exemple de la suma que hem exposat anteriorment.

– Validesa aparent de l'ítem. Quan s'avaluen coneixements és desitjable que les persones que responen al qüestionari pensin que és adequat. Si s'avalua el coneixement en psicometria mitjançant un test, s'espera que els alumnes que facin el test pensin que les preguntes són adequades per a mesurar el coneixement en psicometria. Això és el que es denomina *validesa aparent*. Per contra, en els tests de personalitat s'ha d'intentar que la persona que respon no sàpiga exactament el que s'avalua. D'aquesta manera s'intenta evitar que respongui el que més l'afavoreix o el que pensa que s'espera d'ell.

- **Factors relacionats amb la resposta als ítems**
  – El nombre d'alternatives que s'ofereixin com a resposta. Als tests d'actituds se sol respondre en un format tipus Likert. En aquestes escales es demana a les persones en quin grau estan d'acord amb l'afirmació que se'ls presenta i han de respondre en una escala en què, per exemple, 0 significa totalment en desacord i 5 totalment d'acord. El problema sorgeix quan s'empra una escala que supera la capacitat discriminativa de les persones. En estudiants universitaris, una escala de 0 a 10 és perfectament comprensible però, per contra, aquesta mateixa escala emprada en persones sense estudis pot ser excessiva. Un universitari comprèn perfectament la diferència entre un 4 i un 5, però aquesta diferència pot ser menys clara en una persona sense estudis, de manera que s'introduiria un error en l'avaluació.

  – Les instruccions a l'hora d'emplenar el qüestionari. A l'hora d'emplenar un qüestionari, les instruccions han de ser clares i comprensibles. S'han d'adaptar al grup que es vol avaluar perquè el criteri emprat a l'hora de respondre sigui clar.

- **Factors relacionats amb les persones**. En aquest apartat entrarien totes les característiques personals dels que respondran al qüestionari (capacitat lectora, capacitat intel·lectual, capacitat discriminativa, estat emocional, etc.). Cal fer esment especial de les situacions en què la persona està en un procés de selecció, ja que tractarà de donar una imatge distorsionada de si mateix, tractant d'adaptar-se al que pensa que el seleccionador busca.

3.2. Procediment

Encara que en els *Standards* de l'APA (1999) apareix aquest indici de validesa, amb prou feines aporten informació sobre com es pot determinar si un test té indicis de validesa basat en el procés de resposta. Dins de les alternatives que proposen hi ha tècniques com les següents:

- **L'entrevista**. Es pregunta a les persones que responen al test per les diferents estratègies emprades per a contestar cadascun dels ítems. El coneixement d'aquestes estratègies pot conduir fins i tot a l'enriquiment de la definició del constructe estudiat.

- **Tècniques de pensament en veu alta**. Es demana a les persones que emplenin el qüestionari dient en veu alta els diferents processos pels quals passen a mesura que han de contestar el test.

- **Entrevistes cognitives**. Estan dissenyades per a comprendre com les persones que responen a un test comprenen la pregunta, recuperen la informació rellevant per a respondre, avaluen la rellevància del recordat i responen a la pregunta. Emprant aquesta informació es poden identificar errors de resposta potencials i patrons d'interpretació de les preguntes. També poden aportar informació sobre els factors socioculturals que afecten la manera de respondre (Czaja i Blair, 1996).

4. Evidència de validesa basada en l'estructura interna

4.1. Concepte

Per a elaborar un test s'utilitzaran diversos ítems o preguntes. És possible que es consideri que tots els ítems són igual de rellevants per a mesurar la característica que s'estudia, en aquest cas obtindrem una puntuació total del test per la simple suma de les puntuacions que ha obtingut el subjecte en els diferents ítems.

La situació pot no ser tan senzilla quan suposem que no tots els ítems tenen la mateixa importància en la mesura del constructe, i per tant caldrà ponderar les puntuacions dels ítems, abans de procedir a la suma: en aquest cas parlarem de *puntuacions compostes*. En aquesta situació, l'estructura del test que haurem de determinar és unidimensional, ja que suposem que tots els ítems, encara que de diverses maneres, contribueixen a la mesura d'un únic aspecte de la variable criteri.

Un test també pot presentar una estructura interna multidimensional, això és, que les diferents preguntes no mesuren un sol aspecte sinó dues dimensions o més.

La tècnica estadística de l'anàlisi factorial ens servirà per a estudiar la contribució dels diferents ítems a un sol factor (estructura unidimensional) o a diversos factors (estructura multidimensional).

La tècnica de l'anàlisi factorial ens permetrà determinar k factors subjacents, a partir d'una sèrie p de puntuacions determinades pels ítems inicials del test. La idea és la cerca d'un model parsimoniós (simple) a partir d'un conjunt complex de dades.

A partir dels treballs de Spearman al principi del segle XX, i sobretot de Thurstone als anys quaranta del segle passat, l'anàlisi factorial s'evidencia com una bona eina en psicologia per a tractar d'identificar els factors que intervenen en la intel·ligència. Thurstone va proposar la utilització de l'anàlisi factorial per a explicar les correlacions que observava entre diferents ítems dels test d'intel·ligència. Així, l'ús d'aquesta tècnica li va permetre identificar i diferenciar les capacitats espacial, verbal i numèrica com a factors de la intel·ligència.

El problema d'aquesta tècnica rau en les dificultats del càlcul, sobretot a partir d'un nombre elevat d'ítems (variables). No obstant això, el desenvolupament i popularització actual dels programes estadístics ha permès la difusió d'aquesta i altres tècniques d'anàlisi de dades multivariables.

4.2. Procediments

El terme d'*anàlisi factorial* no designa un concepte unitari, sinó que reuneix diferents procediments que persegueixen la reducció inicial de múltiples variables en un nombre inferior de factors. En processos exploratoris, la tècnica més utilitzada és la de l'anàlisi en components principals, si bé hi ha altres formes d'extracció dels factors o components.

4.2.1. Unidimensionalitat

L'anàlisi en components principals parteix inicialment de la matriu de correlacions entre les diferents variables. Disposem de la matriu de correlacions obtinguda a partir de l'administració d'un test a una mostra de cinquanta-dos individus, i compost per vuit preguntes o ítems que intenten mesurar un únic constructe, en aquest cas l'autoestima dels subjectes.

Taula 2. *Correlation matrix*

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Ítem 7	Ítem 8
Ítem 1	1,000	,447	,411	,444	,533	,337	,365	,442
Ítem 2	,447	1,000	,561	,662	,707	,528	,333	,522
Ítem 3	,411	,561	1,000	,665	,699	,462	,572	,540
Ítem 4	,444	,662	,665	1,000	,682	,518	,560	,564
Ítem 5	,533	,707	,699	,682	1,000	,467	,592	,488
Ítem 6	,337	,528	,462	,518	,467	1,000	,424	,418
Ítem 7	,365	,333	,572	,560	,592	,424	1,000	,422
Ítem 8	,442	,522	,540	,564	,488	,418	,422	1,000

En aquesta situació, la matriu de correlacions presenta una distribució de valors prou uniforme, i en aquest cas no es detecten agrupacions de variables amb correlacions altes entre elles i baixes amb les altres.

És possible descompondre la variància de cada variable (ítem) en tres fonts de variació: la variància factorial comuna, que comparteixen les variables en comú, la variància específica, o no compartida per altres variables, i la variància de l'error. La variància comuna, també denominada *comunalitat* (h^2), interessa que sigui prou alta una vegada hem seleccionat els factors rellevants.

En l'inici de l'anàlisi, la comunalitat de les variables és la unitat; després de l'anàlisi, com més a prop sigui d'1, més relació indicarà amb el factor o factors extrets.

Taula 3. Comunalitats

	Inicial	Extracció
Ítem 1	1,000	,410
Ítem 2	1,000	,628
Ítem 3	1,000	,670
Ítem 4	1,000	,722
Ítem 5	1,000	,743
Ítem 6	1,000	,455
Ítem 7	1,000	,489
Ítem 8	1,000	,518

Mètode d'extracció: anàlisi de components principals

La comunalitat d'un ítem *j* està representada per:

$$h_j^2 = a_j^2 + b_j^2 + \cdots + k_j^2$$

En què a_j^2, b_j^2, ..., k_j^2 representen el quadrat dels coeficients de saturació de cada ítem amb cada factor A, *B*, ..., *K*, extrets, i el coeficient de saturació és la correlació de cada ítem amb els factors extrets.

En l'exemple les variables que presenten més comunalitat són els ítems 4 i 5.

Taula 4. Variància total explicada

Component	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acumulat	Total	% de la variància	% acumulat
1	4,636	57,947	57,947	4,636	57,947	57,947
2	,718	8,969	66,916			
3	,681	8,513	75,429			
4	,572	7,155	82,584			
5	,558	6,969	89,553			
6	,344	4,303	93,856			
7	,304	3,803	97,659			

Mètode d'extracció: anàlisi de components principals

Component	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acumulat	Total	% de la variància	% acumulat
8	,187	2,341	100,000			

Mètode d'extracció: anàlisi de components principals

A partir de *p* variables l'anàlisi factorial extreu el mateix nombre de factors, no relacionats entre ells. Cadascun dels factors es defineix com a combinació lineal de les *p* variables originals. Aquests *p* factors s'ordenen per ordre d'importància. En efecte, el primer component o factor és el millor resum de les relacions lineals que presenten les dades. El segon factor és la segona millor combinació de les variables, amb la condició que sigui ortogonal (sense relació) amb el primer, i així successivament amb la resta dels *p* factors o components.

Un criteri molt estès a l'hora d'extreure els components és el del valor propi o autovalor superior a 1, que és el més utilitzat. Un altre seria retenir els factors necessaris fins a aconseguir un percentatge adequat de variabilitat explicada pels components.

El valor propi o autovalor (λ) es defineix com la suma dels quadrats de les saturacions o correlacions de cada ítem amb el component en qüestió. Per tant, representa una mesura de la variabilitat explicada en les variables per part del component o factor.

En la taula de variància total explicada en què es desglossen els diferents components, veiem que solament hi ha un component amb valor propi (4,63) superior a 1. Per tant, confirmaria una estructura unidimensional del test, ja que totes les preguntes conflueixen en un sol component, identificable com el constructe subjacent, que es vol mesurar. En el nostre exemple, les diferents preguntes de l'escala elaborada contribuirien a la mesura d'un únic constructe psicològic que identificaríem amb l'autoestima.

Taula 5. Matriu de components[a]

	Component
	1
Ítem 1	,640
Ítem 2	,793
Ítem 3	,819
Ítem 4	,850
Ítem 5	,862

Mètode d'extracció: anàlisi de components principals. a. 1 components extrets

	Component
	1
Ítem 6	,675
Ítem 7	,699
Ítem 8	,720

Mètode d'extracció: anàlisi de components principals. a. 1 components extrets

La matriu de components indica les correlacions entre cada ítem amb el component. Són les que abans hem denominat *saturacions* (*factor loadings*). En l'exemple veiem que els valors són alts, i a més amb poca fluctuació, la qual cosa indicaria que tots els ítems tenen una importància similar en la mesura del constructe.

Amb tots els indicadors esmentats podem comprovar que la comunalitat de cada ítem, en haver seleccionat un sol factor, simplement és el quadrat de la saturació entre ítem i factor.

Així, per a l'ítem 1, la comunalitat final és $h_1^2 = (0,64)^2 = 0,41$. Un 41% de la variabilitat del primer ítem és explicada pel component.

El valor propi del component 1 s'obté de la suma dels quadrats de les saturacions de cada ítem amb el factor.

Així en el primer component, $\lambda_1 = (0,64)^2 + (0,793)^2 + ... + (0,72)^2 = 4,63$.

Com que tenim vuit ítems, el màxim serien vuit components. Si fem el quocient 4,63/8 = 0,5795. Un 57,95% de la variabilitat total és explicada pel primer component.

Una vegada hem extret els components, disposarem de la matriu de puntuacions factorials. Aquesta matriu ens proporciona les ponderacions de cada variable per al càlcul de la puntuació de cada subjecte en els factors extrets. Les puntuacions factorials (*factor scores*) per a les dades individuals es calculen a partir de la matriu de coeficients de puntuacions factorials

$$F_i = a_1 \cdot Z_1 + a_2 \cdot Z_2 + \cdots + a_p \cdot Z_p$$

Lectura de la fórmula

a_i: coeficients de ponderació de cada variable per a cada factor

Z_i: puntuacions tipificades dels valors de cada variable obtinguts per cada individu.

Amb les dades de l'exemple, la matriu de ponderacions:

Taula 6. Matriu de coeficients per al càlcul de les puntuacions en els components

	Component
	1
Ítem 1	,138

Mètode d'extracció: anàlisi de components principals

	Component
	1
Ítem 2	,171
Ítem 3	,177
Ítem 4	,183
Ítem 5	,186
Ítem 6	,146
Ítem 7	,151
Ítem 8	,155

Mètode d'extracció: anàlisi de components principals

Per a cada subjecte és possible calcular una puntuació factorial de l'índex d'autoestima:

$$F_1 = 0,138 \cdot Z_{ítem1} + 0,171 \cdot Z_{ítem2} + \cdots + 0,155 \cdot Z_{ítem8}$$

De totes maneres, en aquest exemple es veu que les ponderacions de tots els ítems són molt similars i que la contribució de totes les preguntes en la mesura del constructe d'interès és molt similar, i per tant seria adequat optar per una puntuació simple sumant les puntuacions obtingudes en cada ítem, contra una puntuació composta ponderant els valors de cada ítem.

4.2.2. Multidimensionalitat

Sovint, encara que d'entrada intentem elaborar una escala per a mesurar un sol constructe psicològic, és possible que després de la primera administració del test, en la prova pilot, observem que en realitat els ítems s'agrupen entre ells i afecten diferents constructes subjacents.

Presentem un altre exemple en el qual s'ha elaborat un qüestionari amb la intenció de mesurar les actituds sobre idees religioses en una mostra de 870 subjectes. Els ítems s'han identificat amb el concepte principal que implicava la pregunta. La matriu de correlacions de Pearson entre els ítems es presenta a continuació:

Taula 7. *Correlation matrix*

	Sentit vida	Religió	Obedi-ència	Més enllà	Exper.	Inseguretat	Influència	Independ.
Sentit de la vida	1,000	,295	,220	,253	,226	,134	,178	,027
Religió	,295	1,000	,440	,507	,243	,099	,241	,103
Obediència	,220	,440	1,000	,339	,292	,063	,336	,054
Més enllà	,253	,507	,339	1,000	,309	,113	,278	,121

	Sentit vida	Religió	Obedi-ència	Més enllà	Exper.	Inseguretat	Influència	Independ.
Experiència	,226	,243	,292	,309	1,000	,078	,204	,049
Inseguretat	,134	,099	,063	,113	,078	1,000	,117	,169
Influència	,178	,241	,336	,278	,204	,117	1,000	,102
Independència	,027	,103	,054	,121	,049	,169	,102	1,000

Les correlacions fluctuen en un rang similar entre les diferents preguntes. Potser les que presenten correlacions inferiors són les preguntes referides a inseguretat i independència.

En aplicar l'anàlisi corresponent en components principals, i utilitzant el criteri de valor propi superior a 1, per a l'extracció dels components, obtenim la llista següent:

Taula 8. Comunalitats

	Inicial	Extracció
Sentit de la vida	1,000	,279
Religió	1,000	,554
Obediència	1,000	,513
Més enllà	1,000	,525
Experiència	1,000	,333
Inseguretat	1,000	,562
Influència	1,000	,315
Independència	1,000	,588

Mètode d'extracció: anàlisi de components principals

Taula 9. Variància total explicada

Component	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acumulat	Total	% de la variància	% acumulat
1	2,554	31,926	31,926	2,554	31,926	31,926
2	1,115	13,936	45,862	1,115	13,936	45,862
3	,901	11,258	57,120			
4	,826	10,329	67,448			
5	,787	9,840	77,288			
6	,739	9,234	86,522			

Mètode d'extracció: anàlisi de components principals

Component	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variància	% acumulat	Total	% de la variància	% acumulat
7	,632	7,906	94,428			
8	,446	5,572	100,000			

Mètode d'extracció: anàlisi de components principals

Taula 10. Matriu de components[a]

	Component	
	1	2
Sentit de la vida	,526	–,046
Religió	,735	–,115
Obediència	,685	–,208
Més enllà	,723	–,056
Experiència	,558	–,146
Inseguretat	,267	,701
Influència	,560	,044
Independència	,221	,734

Mètode d'extracció: anàlisi de components principals. a. 2 components extrets

Dels vuit components possibles, solament els dos primers compleixen el criteri d'autovalor superior a 1, encara que un tercer factor està a punt d'arribar a aquest límit ($\lambda_3 = 0{,}901$). En tot cas, s'extreuen dos components.

L'última de les taules presentades (matriu de components) ens mostra les saturacions o correlacions entre els diferents ítems i els dos components.

Recordem que la comunalitat dels ítems representa la variabilitat explicada de l'ítem pels factors extrets. Així, en l'ítem 1 ("Sentit de la vida"):

$$h_1^2 = (0{,}526)^2 + (-0{,}046)^2 = 0{,}279$$

Explicaria un 27,9% de la variabilitat de l'ítem. És la més baixa de tots els ítems del test. Potser aquest ítem saturaria amb un tercer component.

El valor propi (autovalor) de cada component es calcula amb la suma de quadrats de les correlacions (saturacions) entre ítems i component.

Component 1: $\lambda_1 = (0{,}526)^2 + (0{,}735)^2 \ldots + (0{,}221)^2 = 2{,}554$

Component 2: $\lambda_2 = (-0,046)^2 + (-0,115)^2 + ... + (0,734)^2 = 1,115$

El primer component (2,554/8 = 0,319) explicaria un 31,9% de la variabilitat presentada pels ítems, mentre que el segon (1,115/8 = 0,139) n'explicaria un 13,9%. Un 45,8% de la variància total és explicada per la combinació dels dos factors o components.

L'anàlisi de la matriu de saturacions ens permetrà intentar buscar una interpretació als dos components, en funció dels ítems que correlacionin de manera més alta.

En l'exemple observem que el primer component té saturacions elevades en els ítems 1, 2, 3, 4, 5, i 7; mentre que les correlacions són baixes en els ítems 6 i 8. En el segon component s'esdevé just el contrari: té correlacions altes amb els ítems 6 i 8 i, en canvi, baixes en els altres.

De vegades, a simple vista la solució final no presenta una interpretació tan fàcil. En aquests casos és possible utilitzar una rotació dels eixos per a aconseguir que les correlacions siguin fortes en un eix o component i baixes en els altres. Recordem que els eixos són ortogonals i no estan relacionats entre ells. Una de les tècniques matemàtiques de rotació dels eixos és la rotació varimax, que és la més utilitzada en processos exploratoris, encara que els programes estadístics n'incorporen d'altres com les rotacions quartimax, equimax, etc.

La matriu de saturacions amb la solució de les rotacions amb el mètode varimax, amb les dades de l'exemple, quedaria com segueix:

Taula 11. Matriu de components rotats[a]

	Component	
	1	**2**
Sentit de la vida	,522	,080
Religió	,742	,062
Obediència	,715	–,040
Més enllà	,715	,117
Experiència	,577	–,010
Inseguretat	,093	,744
Influència	,533	,175
Independència	,041	,765

Mètode d'extracció: anàlisi de components principals. Mètode de rotació: normalització varimax amb Kaiser. a. La rotació ha convergit en 3 iteracions.

Veiem com la solució de les rotacions confirma en aquest cas la conclusió prèvia, sis de les preguntes saturen en el primer component. Veient els sis ítems que saturen aquest component –"Sentit de la vida", "Religió", "Obediència", "Més enllà", "Experiència", "Influència"–, podem interpretar aquest component com la mesura de l'actitud sobre idees religioses, que era el motiu inicial de l'elaboració del test.

El segon component, se suposa que no esperat inicialment en l'elaboració del qüestionari, satura els ítems "Inseguretat" i "Independència". Aquest segon component es pot interpretar com una mesura de l'actitud envers la dependència personal.

Com que només es tracta de dos components, és possible representar gràficament en un espai bidimensional els dos eixos que representen els components i la situació dels ítems respecte a aquests en funció de les saturacions obtingudes.

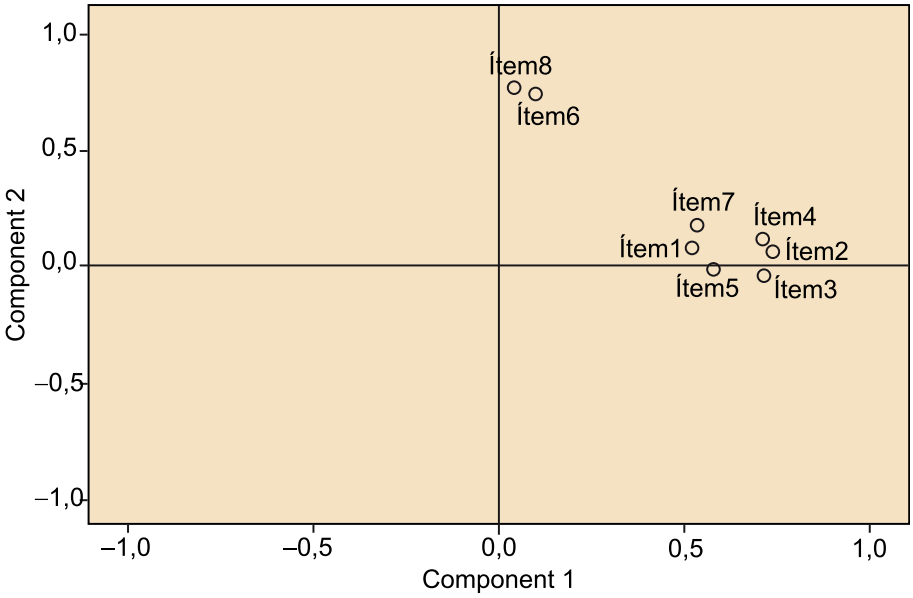

Figura 5. Gràfic de components en espai amb rotació

El gràfic dóna una informació visual ràpida de l'agrupació dels ítems en els dos components.

Una vegada obtinguts els factors terminals, podrem calcular els valors o puntuacions factorials de les dimensions teòriques que hem associat als factors respectius, idees religioses (component 1) i dependència (component 2).

Taula 12. Matriu de coeficients per al càlcul de les puntuacions en els components

	Component	
	1	**2**
Sentit de la vida	,210	,009

Mètode d'extracció: anàlisi de components principals. Mètode de rotació: normalització varimax amb Kaiser.

	Component	
	1	2
Religió	,304	–,032
Obediència	,305	–,118
Més enllà	,287	,018
Experiència	,243	–,076
Inseguretat	–,047	,635
Influència	,204	,090
Independència	–,072	,660

Mètode d'extracció: anàlisi de components principals. Mètode de rotació: normalització varimax amb Kaiser.

Per a cada subjecte *i* és possible calcular les puntuacions factorials de les noves variables *Idees religioses* i *Dependència*:

$$Idees\ religioses_i = 0{,}21 \cdot Zi_{ítem1} + 0{,}304 \cdot Zi_{ítem2} + \cdots + 0{,}204 \cdot Zi_{ítem7} - 0{,}072 \cdot Zi_{ítem8}$$

$$Dependència_i = 0{,}009 \cdot Zi_{ítem1} - 0{,}032 \cdot Zi_{ítem2} + \cdots + 0.{,}09 \cdot Zi_{ítem7} + 0{,}66 \cdot Zi_{ítem8}$$

Els diferents paquets estadístics permeten el càlcul automàtic i la generació d'aquestes noves variables generades i que representarien la mesura dels constructes subjacents.

5. Evidència de validesa basada en la relació amb altres variables

5.1. Concepte

En el procés de validació d'una nova prova psicològica ens podem ajudar d'altres instruments de mesura del constructe d'interès que estiguin contrastats i considerats vàlids i fiables. En aquest procés parlarem de validesa convergent, o correlació entre puntuacions del test amb altres mesures del mateix constructe fetes a partir de diferents tècniques o indicadors.

Les diferents tècniques estadístiques de relació entre variables ens serviran per a determinar el coeficient de validació entre les dues variables. Així, el més utilitzat serà el coeficient de correlació de Pearson, en cas que les dues variables siguin quantitatives, però també qualsevol de les seves variacions, com els coeficients de Spearman, biserial puntual, biserial, phi, tetracòrica, etc., en funció de com siguin les dues variables que s'han de relacionar.

Taula 13

Variable A	Variable B	Coeficient correlació
Numèrica (interval o raó)	Numèrica (interval o raó)	*r* de Pearson
Numèrica (interval o raó)	Numèrica (ordinal)	r_s Spearman o *τ* de Kendall
Numèrica (ordinal)	Numèrica (ordinal)	r_s Spearman o *τ* de Kendall
Qualitativa	Qualitativa	*V* de Cramer
Qualitativa (dicotòmica)	Numèrica (interval o raó)	r_b biserial o r_{bp} biserial puntual
Qualitativa (dicotòmica)	Qualitativa (dicotòmica)	Φ phi o r_t tretacòrica

Per poder conèixer el nivell de benestar físic i psicològic en gent gran, ens interessa validar un nou qüestionari que hem elaborat, per determinar el grau d'independència en les activitats bàsiques de la vida diària (ABVD). Amb aquest objecte podem utilitzar algunes de les proves que ja hi ha en el mercat i que es troben prou contrastades. En una mostra de tres-cents subjectes més grans de setanta anys, i ingressats en centres geriàtrics, administrem la nova prova elaborada conjuntament amb l'escala de mesura d'independència funcional (MIF o FIM) Keith, Granger, Hamilton i Sherwin, 1987) i l'escala de grau d'autonomia de Barthel (Mahoney i Barthel, 1965).

En la taula següent es mostra la matriu de correlacions de Pearson (dades simulades) entre les tres proves administrades:

Taula 14

	ABVD	FIM	Barthel
ABVD	1		
MIF	0,69	1	
Barthel	0,67	0,77	1

Els valors de la correlació entre la nova prova (ABVD) i les escales MIF i Barthel presenten valors prou alts (069 i 0,67, respectivament), la qual cosa indica una validesa concurrent alta de la nova prova elaborada amb les tècniques prèvies per a mesurar el grau d'autonomia de les persones grans analitzades.

5.2. Evidència de decisió (sensibilitat i especificitat)

En situacions en què la prova feta tingui com a objectiu el diagnòstic o classificació dels subjectes en dos grups (diagnòstic negatiu - diagnòstic positiu), parlarem de la validesa de decisió, quan comparem aquesta nova prova amb un altre mètode de diagnòstic anterior prou contrastat. En la validesa de decisió podem distingir dos processos: d'una banda, la sensibilitat de la prova, definida com la capacitat d'aquesta per a detectar veritables positius; i, d'altra banda, l'especificitat, com la capacitat per a determinar diagnòstics negatius veritables.

Taula 15

		Diagnòstic prova inicial		
		Positiu	**Negatiu**	**Total**
Diagnòstic nova prova	**Positiu**	Decisió correcta (f_{11})	Fals positiu (f_{12})	$f_{1.}$
	Negatiu	Fals negatiu (f_{21})	Decisió correcta (f_{22})	$f_{2.}$
	Total	$f_{.1}$	$f_{.2}$	n

Una mesura de l'acord assolit per mitjà de les dues proves diagnòstiques consistirà a calcular el percentatge d'acord (P_c) entre totes dues tècniques, a partir de la raó entre la suma de decisions correctes i el total de decisions.

$$P_c = \frac{f_{11} + f_{22}}{n}$$

La sensibilitat de la nova prova l'obtindrem a partir de la proporció de subjectes classificats correctament com a veritables positius.

$$Sensibilitat = \frac{Diagnòstics\ positius\ de\ la\ prova}{Total\ diagnòstics\ positius} = \frac{f_{11}}{f_{.1}}$$

Mentre que l'especificitat s'obté mitjançant el quocient dels diagnosticats sense trastorn per la prova entre el total de diagnòstics negatius.

$$Especificitat = \frac{Diagnòstics\ \ negatius\ \ de\ \ la\ \ prova}{Total\ \ \ diagnòstics\ \ \ negatius} = \frac{f_{22}}{f_{.2}}$$

Un índex global per a valorar la validesa, el proporciona el càlcul del coeficient kappa, establert inicialment com a indicador de l'acord entre dos observadors. L'avantatge que presenta és que és fàcil d'interpretar, similar a la d'altres indicadors de relació entre variables. En efecte, el seu valor fluctua entre 0 (cap acord) i 1 (màxim acord). La fórmula de càlcul és senzilla:

$$K = \frac{F_c - F_a}{n - F_a}$$

en què

$$F_c = f_{11} + f_{22} \qquad y \qquad F_a = \frac{f_{1.} \cdot f_{.1} + f_{2.} \cdot f_{.2}}{n}$$

Taula 16. Criteris Alman per a interpretar kappa

Valor	Relació
0-0,20	Inexistent
0,21-0,40	Molt baixa
0,41-0,60	Moderada
0,61-0,80	Bona
0,81-1	Intensa

En una consulta psicològica es vol validar una nova prova, més simple que les tradicionals, per al diagnòstic del trastorn de depressió dels pacients atesos. En una mostra de cinc-cents pacients atesos al centre s'administren dues proves (tradicional i versió breu) per al diagnòstic del trastorn de depressió.

Taula 17

		Diagnòstic tradicional depressió		
		Positiu	**Negatiu**	**Total**
Versió breu escala Hamilton	**Positiu**	125	50	175
	Negatiu	25	300	325
	Total	150	350	500

$$P_c = \frac{125+300}{500} = 0,85$$

Les dues proves presenten un percentatge d'acord (85%) elevat.

Així mateix, els valors de sensibilitat i especificitat indiquen que la nova prova té una bona capacitat per a detectar subjectes amb trastorn depressiu (sensibilitat = 0,83), i per a detectar subjectes sense depressió (especificitat = 0,86).

$$Sensibilitat = \frac{125}{150} = 0,83$$

$$Especificitat = \frac{300}{350} = 0,86$$

$$F_c = 125+300 = 425 \qquad \text{i} \qquad F_a = \frac{175 \cdot 150 + 325 \cdot 350}{500} = 280$$

$$K = \frac{425-280}{500-280} = 0,66$$

El càlcul de l'índex kappa d'acord entre les dues proves (K = 0,66) indica una bona relació entre aquestes. Per tant, sembla adequat pensar que l'economia de temps i esforç (tant per al pacient com per al terapeuta) justificaria la nova escala de diagnòstic de depressió, en funció dels resultats obtinguts en la validesa de decisió.

5.3. Evidències convergents i/o discriminants

Fins ara, en aquest apartat, ens hem referit a escales que volen mesurar un sol constructe psicològic. Si ens referim a proves formades per ítems que mesuren diferents constructes (trets múltiples), podrem diferenciar dos tipus de validesa. D'una banda, la validesa convergent (enunciada anteriorment), o sigui, la validesa que determinen diferents proves que mesuren el mateix constructe, i, d'altra banda, la validesa discriminant, que està determinada per la mesura de diferents constructes dins de la mateixa prova.

La matriu de correlacions entre les puntuacions dels diferents trets obtinguts a partir de les diferents escales (matriu multitret-multimètode; Campbell i Fiske, 1959) ens servirà per a determinar els diferents valors de validesa convergent i discriminant.

Imaginem la situació en què disposem de tres escales diferents, formades per ítems que mesuren els mateixos dos constructes subjacents: escales *A*, *B* i *C*, que mesuren els constructes 1 i 2. Per a cada subjecte analitzat tindrem, per tant, sis puntuacions diferents obtingudes per la combinació de cada prova i cada tret analitzat.

Taula 18

	***A*1**	***A*2**	***B*1**	***B*2**	**C1**	**C2**
***A*1**	**Fi**					
***A*2**	Vd	**Fi**				
***B*1**	Vc		**Fi**			
***B*2**		Vc	Vd	**Fi**		
C1	Vc		Vc		**Fi**	
C2		Vc		Vc	Vd	**Fi**

En la taula anterior, les diferents escales es representen per les lletres *A*, *B* i *C*, i dins de cada escala 1 i 2 representen els dos constructes que es volen analitzar.

Si observem la matriu multitret-multimètode (MTMM), trobem en la diagonal principal valors de fiabilitat de les escales (valors 1 si s'obtenen en una única administració).

Els valors de validesa convergent es troben en les combinacions dels mateixos trets i diferents escales (per exemple, casella *A*1 i *B*1), esperant que aquests valors de validesa siguin prou alts, la qual cosa indica convergència de les diferents maneres de mesurar el constructe i aporta evidència real de l'existència del constructe.

Els valors de validesa discriminant seran els coeficients de correlació obtinguts dins de la mateixa escala per les puntuacions dels diferents trets. En aquest cas, esperem que els diferents constructes siguin prou independents entre ells perquè les correlacions siguin properes a 0.

En una mostra de sis-cents subjectes s'han utilitzat tres proves diferents de personalitat (tests 1, 2 i 3), formades cada una per ítems referits als tres mateixos constructes de la personalitat (trets *A*, *B* i *C*).

La taula següent mostra els valors de la matriu de correlacions entre les nou variables obtingudes de la combinació de les tres proves i els tres trets (3 × 3).

Taula 19

	***A*1**	***B*1**	**C1**	***A*2**	***B*2**	**C2**	***A*3**	***B*3**	**C3**
***A*1**	1								
***B*1**	*0,03*	1							
C1	*0,28*	*0,17*	1						
***A*2**	0,73	0,14	0,22	1					
***B*2**	0,15	0,69	0,03	*0,18*	1				
C2	0,12	0,04	0,81	*0,03*	*0,15*	1			
***A*3**	0,77	0,11	0,09	0,77	0,21	0,16	1		

	A1	*B1*	*C1*	*A2*	*B2*	*C2*	*A3*	*B3*	*C3*
B3	0,21	0,75	0,18	0,21	0,68	0,03	*0,15*	1	
C3	0,19	0,05	0,78	0,22	0,09	0,72	*0,14*	*0,09*	1

Els valors de validesa convergent són els valors destacats amb el subratllat. Els valors de validesa discriminant es destaquen amb la cursiva.

Si ens fixem en el tret *B*, es detecta convergència a partir de la mesura per mitjà de diferents proves:

$$r\,(B1 - B2) = 0{,}69$$

$$r\,(B1 - B3) = 0{,}75$$

$$r\,(B2 - B3) = 0{,}68$$

Si ens fixem en l'escala 1, observem que hi ha prou independència entre les diferents mesures dels tres trets mesurats:

$$r\,(A1 - B1) = 0{,}03$$

$$r\,(A1 - C1) = 0{,}28$$

$$r\,(B1 - C1) = 0{,}17$$

5.4. Evidències basades en les relacions test-criteri

De vegades, un test o prova psicològica construït per a mesurar un determinat constructe psicològic pot estar relacionat amb una altra variable d'interès que es denomina *criteri*.

Per exemple, imaginem que hem elaborat una prova vàlida que ens permet mesurar la capacitat de raonament numèric de les persones (test), i observem que presenta una correlació molt alta amb els resultats que obtenen els subjectes en una determinada prova de matemàtiques (criteri).

Podem distingir tres tipus de situacions:

- Validesa concurrent o simultània
- Validesa predictiva
- Validesa retrospectiva

5.4.1. Validesa concurrent o simultània

En aquest cas el test i el criteri es mesuren de manera simultània. Obtindrem validesa concurrent en obtenir valors alts de coeficients de correlació entre les puntuacions del test i del criteri. Per tant, ens permet validar el test, inicialment elaborat per a mesurar una altra variable, per a mesurar el criteri.

En funció del tipus d'escala de mesura utilitzat tant per a les puntuacions del test com del criteri, utilitzarem un tipus o un altre de coeficient per a mesurar la correlació.

Com a exemple, vegem la situació següent, en la qual s'ha administrat un test de raonament numèric a un grup de vint subjectes just abans de fer una determinada prova de matemàtiques:

Taula 20

Test raonament	Nota matemàtiques
100	10
100	9
100	9
90	8
90	8
80	9
70	7
70	7
70	7
70	5
70	4
70	4
50	3
40	3
40	2
40	2
30	1
20	1
20	3
20	2

Una visió del gràfic de dispersió ens donarà una idea de si hi ha relació lineal entre totes dues proves, o no:

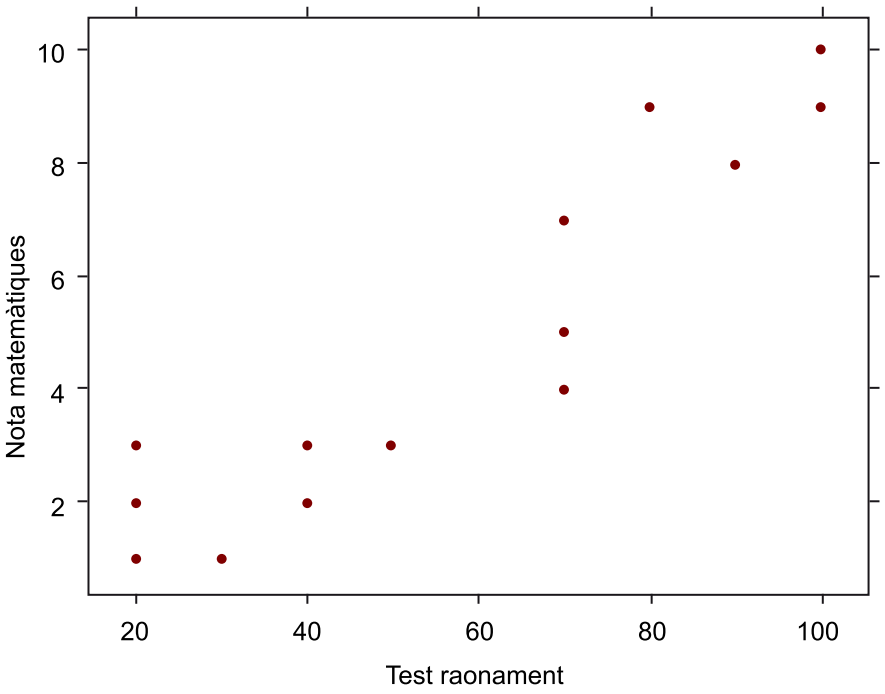

Figura 6. Gràfic de dispersió (núvol de punts)

En calcular el coeficient de correlació de Pearson:

```
> rcorr.adjust(Dades[,c("Nota.Matemàtiques","Test.Raonament")], type="pearson")
                      Nota. Matemàtiques     Test.Raonament
Nota. Matemàtiques             1.00              0.92291
Test.Raonament               0.92291                 1.00
n= 20
```

S'obté un valor de validesa concurrent igual a 0,92291, que indica una forta relació directa i propera a 1.

Una mesura de la bondat de l'ajust lineal entre les dues variables es defineix per r^2, o sigui el valor del quadrat de la correlació, en el nostre exemple $r^2 = (0{,}92291)^2 = 0{,}8518$. Aquest valor, que es denomina *coeficient de determinació*, multiplicat per 100, indica el percentatge de variabilitat en la variable criteri, que és explicat per la relació amb la variable independent. Per tant, el 85,18% de la variabilitat que presenten les puntuacions obtingudes en la nota de matemàtiques estaria explicada per la relació que presenta amb els valors que s'han obtingut en el test de raonament numèric.

5.4.2. Validesa predictiva

Si sabem que un determinat test i una variable criteri es troben altament relacionats, serà possible utilitzar els valors obtinguts en el test per a la predicció o pronòstic dels valors que s'obtindran en el criteri. Parlarem en aquest cas de la validesa predictiva que té el test respecte a la variable criteri.

Per exemple, si volem seleccionar un candidat per a un determinat lloc de treball, podem utilitzar determinades proves que tinguin una alta validesa predictiva respecte al rendiment futur dels candidats en el lloc de treball. O, utilitzant l'exemple esmentat més amunt, podem fer un pronòstic de la nota que obtindran els subjectes en una determinada prova de matemàtiques a partir de les puntuacions que van obtenir al seu moment en el test de raonament numèric.

Quan tant les puntuacions del test com el criteri són puntuacions numèriques, i hem calculat el coeficient de correlació de Pearson corresponent, que és significatiu estadísticament, és possible establir un model de regressió lineal per a poder fer el pronòstic dels valors del criteri.

Les puntuacions en el test constitueixen la variable independent del model (variable predictora), mentre que el criteri representa la variable dependent.

Regressió lineal simple

És el cas més senzill i solament disposem d'una variable independent (X) i una variable depenent (Y).

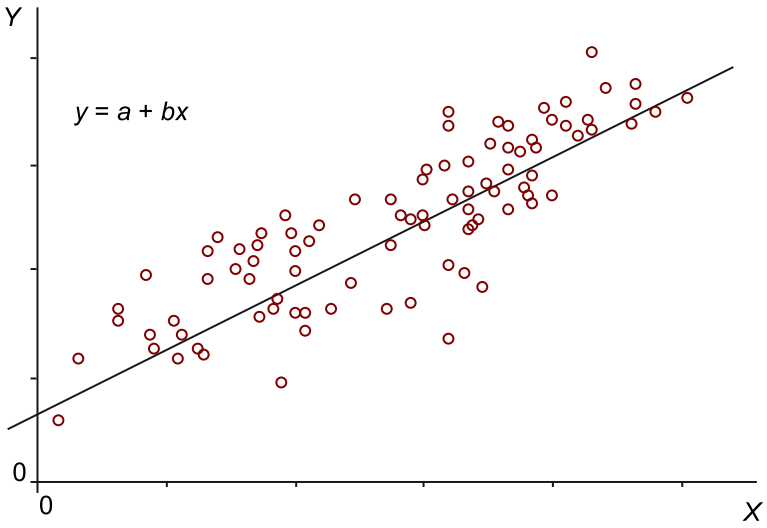

Figura 7. Gràfic de dispersió amb recta de regressió

La regressió lineal descriu una relació lineal entre Y i X, això és, en el gràfic de dispersió representa la recta que millor s'ajusta al núvol de punts.

L'expressió d'una línia recta és $y = a + bx$, en què b representa el pendent de la recta, o sigui, el canvi que es produeix en y a partir del canvi que es produeixi en x, i a es denomina *intersecció* o *intercepta*, i y és el valor que pren y quan x és igual a 0.

Per a trobar els coeficients de la regressió, a i b, usem un mètode d'estimació molt conegut en estadística, el mètode de mínims quadrats, el qual minimitza la suma dels quadrats de les diferències (o residus) entre els valors y_i i els valors estimats segons la recta de regressió $y'_i = a + bx_i$.

A partir de les dades (x_i, y_i), $i = 1, \ldots, n$, estimem els coeficients a i b de la recta de regressió. Així doncs, tenim:

- Pendent:

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{s_x^2}$$

Lectura de la fórmula

S_{xy}: covariància entre x i y.
s^2_x: variància de x.

- Intersecció:

$$a = \overline{y} - b \cdot \overline{x}$$

Comparant les fórmules del pendent b i del coeficient de correlació r, tenim la relació següent:

$$b = r_{xy} \cdot \frac{s_y}{s_x}$$

És possible verificar o validar el model de regressió a partir del coeficient de determinació r^2. Recordem que més amunt l'hem definit com una mesura de bondat d'ajust, o mesura de la proximitat dels punts a la recta estimada. Representa la proporció de variància de la variable dependent explicada per la recta de regressió.

El valor $1 - r^2$ quantifica la proporció de variància que no és explicada per la regressió. A partir d'aquests dos valors podem calcular un estadístic de contrast:

$$F_{EC} = \frac{r2}{(1-r2)/n-2}$$

Aquest estadístic de contrast F es distribueix seguint una distribució F de Snedecor, amb 1 grau de llibertat en el numerador i $n - 2$ graus de llibertat en el denominador.

Les hipòtesis que s'han de contrastar seran:

- H_0: el model no és vàlid, no hi ha relació.

- H_1: sí que hi ha relació, per tant el model sí que és vàlid.

Seguint amb l'exemple del test de raonament numèric i el criteri de la nota de matemàtiques, el resultat amb el programa *R* és el següent:

```
> summary(RegModel.1)


Call:
lm(formula = Nota.Matemàtiques ~ Test.Raonament, data = Dades)
```

```
Residuals:
   Min       1Q    Median    3Q      Max
-2.01102 -0.96970 -0.03857 0.98898 2.05785

Coefficients:
                 Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)      -1.085399  0.673899   -1.611    0.125
Test.Raonament 0.101377  0.009969   10.170    6.89e-09 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.201 on 18 degrees of freedom

Multiple R-squared: 0.8518, Adjusted R-squared: 0.8435
F-statistic: 103.4 on 1 and 18 DF, p-value: 6.89e-09
```

Un cop consultada la sortida del programa, podem especificar el model de regressió:

$$\textit{Nota matemàtiques} = -1{,}085 + 0{,}1014 \times \textit{Test} + \textit{Residual}$$

També observem que el model està verificat, ja que el valor de p (grau de significació) que acompanya el valor de l'estadístic de contrast ($F = 103{,}4$) és tendent a 0. Per tant, res no s'oposa a rebutjar la hipòtesi nul·la, i el model està validat.

Si representem la recta de regressió en el gràfic de dispersió:

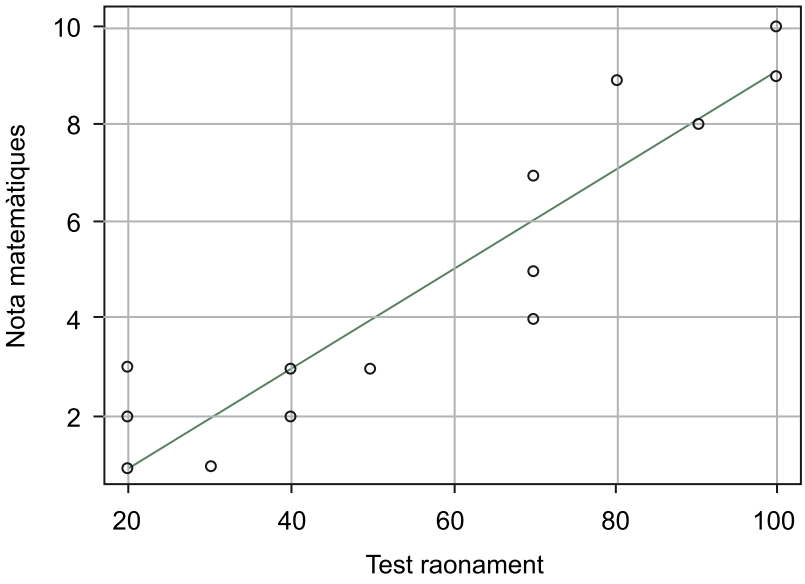

Figura 8. Gràfic de dispersió amb ajust de la recta

A partir de l'expressió de la recta de regressió, podem fer el pronòstic, per a cada subjecte, del valor de la nota de matemàtiques en funció del test, com també el càlcul del residual mitjançant la diferència entre la puntuació real obtinguda i la puntuació pronosticada.

Taula 21

Test raonament	Nota matemàtiques	Pronòstic	Residual
100	10	9,0546	0,9454
100	9	9,0546	–0,0546
100	9	9,0546	–0,0546
90	8	8,0406	–0,0406
90	8	8,0406	–0,0406
80	9	7,0266	1,9734
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	5	6,0126	–1,0126
70	4	6,0126	–2,0126
70	4	6,0126	–2,0126
50	3	3,9846	–0,9846
40	3	2,9706	0,0294
40	2	2,9706	–0,9706
40	2	2,9706	–0,9706
30	1	1,9566	–0,9566
20	1	0,9426	0,0574
20	3	0,9426	2,0574
20	2	0,9426	1,0574

```
> numSummary(Dades[,c("Nota.Matemàtiques", "Pronòstic", "Residual")],
+  statistics=c("mean", "sd","var"))
                    mean     sd        var     n
Nota.Matemàtiques 5.2000  3.036619  9,221    20
Pronòstic         5.2014  2.803138  7,855    20
Residual         -0.0014  1.169176  1,367    20
```

Tal com hem indicat més amunt, el coeficient de determinació es defineix com el quocient entre la variància explicada per la regressió i la variància total de la variable criteri:

$$r^2 = \frac{s^2_{y'}}{s^2_y} = 1 - \frac{s^2_{y-y'}}{s^2_y}$$

Lectura de la fórmula

s^2_y: variància de la variable dependent *Y*.

$s^2_{y'}$: variància dels pronòstics obtinguts mitjançant l'equació de regressió.

$s^2_{y-y'}$: variància dels errors produïts.

Amb les dades del nostre exemple:

$$r^2 = \frac{7,855}{9,221} = 1 - \frac{1,367}{9,221} = 0,8518$$

Per tant, a partir dels valors del coeficient de determinació i de la variància de la variable criteri, és possible, aïllant l'expressió, obtenir el valor de la variància dels errors:

$$s^2_{y-y'} = s^2_y(1 - r^2)$$

La desviació típica dels errors o error típic o estàndard de l'error ens ajudarà en l'estimació per interval de nous valors desconeguts.

$$s_{y-y'} = s_y\sqrt{1 - r^2}$$

En efecte, si necessitem fer un pronòstic en el criteri a partir d'un nou valor en el test, o variable predictora, és possible fer-lo de manera puntual, però, donada una probabilitat, aconseguirem estimacions millors si es fa per interval.

$$IC^{1-\alpha} \rightarrow y' \pm t_{n-1;\alpha/2} \cdot s_{y-y'}$$

Lectura de la fórmula

1 – α: nivell de confiança de l'interval construït.

t: valor de la distribució *t* de Student-Fisher tabulat en funció de α i dels graus de llibertat (*n* – 1).

Imaginem que un nou subjecte obté una puntuació igual a 60 en el test de raonament numèric, l'estimació puntual del valor en la nota de matemàtiques serà:

$$Nota\ matemàtiques = -1,085 + 0,1014 \cdot 60 = 4,999$$

Si fem una estimació per interval, amb un nivell de confiança del 95%:

$$IC^{0,95} \rightarrow 4,999 \pm 2,093 \cdot 1,169 = [2,552 \quad 7,446]$$

Amb una probabilitat de 0,95, el valor en el criteri d'un subjecte que obtingui una puntuació de 60 en el test estarà entre 2,552 i 7,446 punts.

Regressió lineal múltiple

El model lineal general planteja que una variable dependent (criteri) sigui funció de diverses variables independents, situació, d'altra banda, força més habitual. En el cas de dues variables independents, l'expressió que relaciona les tres variables serà la fórmula d'un pla, en les situacions en què hi hagi més de dues variables independents, situacions multivariants, parlarem de l'hiperplà de regressió.

$$Criteri = a + b_1 \cdot Test_1 + b_2 \cdot Test_2 + b_3 \cdot Test_3 + \cdots + b_p \cdot Test_p + Residual$$

Recuperant l'exemple anterior, imaginem que als vint subjectes als quals s'ha administrat un test de raonament juntament amb un test de càlcul mental, abans de fer una prova de matemàtiques, que representa el criteri que més endavant volem pronosticar.

Taula 22

Test raonament	Test càlcul	Nota matemàtiques
100	9	10
100	8	9
100	8	9
90	8	8
90	7	8
80	9	9
70	6	7
70	5	7
70	6	7
70	6	5
70	4	4
70	4	4
50	5	3
40	4	3
40	4	2
40	5	2
30	3	1
20	2	1
20	3	3
20	3	2

En calcular la matriu de correlacions de Pearson:

Taula 23

	Test raonament	Test càlcul	Nota matemàtiques
Test raonament	1		
Test càlcul	0,891763351	1	
Nota matemàtiques	0,922905961	0,925261006	1

Observem que el test de càlcul mental també està altament correlacionat amb la variable criteri (nota de matemàtiques). Així mateix, veiem una alta correlació entre les dues variables independents, test de raonament numèric i test de càlcul mental ($r = 0{,}89176$).

L'anàlisi de regressió múltiple ens ajudarà a determinar si la incorporació d'aquesta nova variable augmenta, significativament, la variabilitat explicada per la regressió en la variable criteri.

L'anàlisi de regressió es basarà en l'anàlisi de la relació conjunta entre la variable criteri i el conjunt de les dues variables independents. El quadrat d'aquesta correlació múltiple serà el nou coeficient de determinació.

La verificació del model es farà a partir de l'expressió:

$$F_{EC} = \frac{r2 \Big/ p}{(1-r2) \Big/ n-p-1}$$

Lectura de la fórmula

p és igual al nombre de variables independents en el model.

L'estadístic de contrast *F* es distribueix seguint una distribució *F* de Snedecor, amb *p* graus de llibertat en el numerador i $(n - p - 1)$ graus de llibertat en el denominador.

A continuació es presenta la sortida del programa obtinguda en aquest exemple mitjançant l'ús del programa *R*:

```
> RegModel.2 <- lm(Nota.Matemàtiques~Test.Numèric+Test.Raonament, data=Dades)

> summary(RegModel.2)

Call:
lm(formula = Nota.Matemàtiques ~ Test.Numèric + Test.Raonament, data = Dades)

Residuals:
Min      1Q    Median    3Q      Max
-1.72686 -0.64006 -0.00108 0.52167 1.74005
```

```
Coefficients:

                    Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)         -1.91579  0.62646    -3.058    0.00711 **
Test.Càlcul          0.70883  0.23719     2.988    0.00825 **
Test.Raonament       0.05246  0.01835     2.858    0.01088 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 17 degrees of freedom
Multiple R-squared: 0.9028, Adjusted R-squared: 0.8914
F-statistic: 78.96 on 2 and 17 DF, p-value: 2.481e-09
```

El valor de coeficient de determinació és 0,9028, per tant un 90,28% de la variància de la variable criteri s'explica per la regressió entre aquesta variable i la combinació de les dues variables independents.

L'equació del pla de regressió es troba verificada globalment, ja que el valor de l'estadístic de contrast (F = 78,96) indicia una probabilitat tendent a 0 (*p value*) que no hi hagi relació entre les variables.

També és important observar si es troben significats en els diferents coeficients de la regressió (b). En aquest cas, tant el coeficient que afecta el test de raonament (*p value* = 0,01) com el que afecta el test de càlcul mental (p = 0,008) estan verificats, ja que els graus respectius de significació associats són propers a 0.

Per tant, l'especificació de l'equació resultant del model de regressió quedarà de la manera següent:

Nota matemàtiques = –1,916 + 0,709 · *Test_càlcul* + 0,052 · *Test_raonament* + Residual

Així mateix, la sortida del programa informa del valor de l'error típic o estàndard de l'error ($s_{y-y'}$ = 1,001), valor necessari per a l'estimació per interval dels valors del criteri. En efecte, per a un nou subjecte que obtingués una puntuació igual a 6 en el test de càlcul mental i 60 en el test de raonament numèric, l'estimació puntual del valor en la nota de matemàtiques serà:

$$Nota \quad matemàtiques = -1,916 + 0,709 \cdot 6 + 0,052 \cdot 60 = 5,458$$

Si fem una estimació per interval, amb un nivell de confiança del 95%:

$$IC^{0,95} \rightarrow 5,458 \pm 2,093 \cdot 1,001 = [3,363 \quad 7,553]$$

Amb una probabilitat de 0,95, el valor en el criteri d'un subjecte que obtingui una puntuació de 6 en la prova de càlcul i de 60 en el test de raonament estarà entre 3,363 i 7,553 punts.

Altres tècniques estadístiques d'anàlisi multivariable

En funció del tipus d'escala de mesura utilitzada per a les variables criteri i les variables predictores, serà necessari aplicar alguna de les diferents tècniques d'anàlisi de dades multivariables.

Per exemple, si disposem d'una variable criteri mesurada en escala nominal, i per tant de tipus qualitatiu o categòric, mentre que les variables independents són de tipus quantitatiu o numèric, podem utilitzar una tècnica de classificació, com l'anàlisi discriminant. En efecte, suposem que la variable criteri nota de matemàtiques, de l'exemple que hem utilitzat, estigués codificada en suspens, aprovat, notable, excel·lent, o que els subjectes estiguessin directament simplement dividits entre aprovats i suspesos.

Aplicar una anàlisi discriminant ens permetria determinar la millor funció discriminant que aconsegueixi la classificació dels subjectes, en funció de les puntuacions obtingudes en les proves predictores del càlcul mental i el raonament numèric. La funció discriminant establirà l'estimació dels pesos (coeficients) i la combinació lineal de les variables independents (discriminants), de manera que els grups siguin, des del punt de vista estadístic, tan diferents com sigui possible.

Una altra opció pot ser que tant la variable criteri com les variables independents estiguin codificades en categories. En aquest cas seria aplicable la tècnica del model lògit. Aquest model, basat en els models lineals logarítmics, pretén, seguint l'enfocament de la regressió múltiple, trobar l'expressió d'associació entre la variable criteri i les variables independents, tenint en compte també la interacció entre les variables independents.

5.4.3. Validesa retrospectiva

La validesa concurrent entre un o diversos tests i el criteri, que pot ser útil per a la predicció futura de la variable criteri, en certes situacions i donades certes conseqüències mesurades per mitjà del criteri, també pot servir per a trobar les causes als valors obtinguts.

En aquest cas, la variable criteri s'ha registrat abans de les variables predictores. Per exemple, en psicologia és habitual aplicar diferents proves psicològiques que permetin donar una explicació a una determinada conducta d'un subjecte.

5.5. Generalització de la validesa

El concepte de *generalització de la validesa* es refereix al fet d'estendre la validesa establerta entre test i criteri a altres situacions o a grups de subjectes diferents dels utilitzats inicialment en el càlcul.

Per exemple, imaginem que es troben validades determinades proves per a la selecció correcta de personal per a determinats llocs d'administratiu en un banc *A*. Si és possible utilitzar les mateixes proves per a la selecció d'administratius en un altre banc *B*, podrem considerar que la validesa s'ha generalitzat. En aquest cas es tractaria de subjectes diferents dels utilitzats inicialment.

Un altre cas podria ser que les proves de selecció per a administratius de banca s'utilitzessin per a seleccionar personal administratiu en companyies d'assegurances. En aquest cas parlaríem de generalització també a situacions diferents.

6. Evidència de validesa basada en les conseqüències de l'aplicació

Quan es prenen decisions a partir de l'aplicació d'un test, i no es tracta només de descriure o interpretar sense que se'n derivin accions, s'ha de pensar en les conseqüències que té aplicar aquest qüestionari (Shepard, 1997). Els tests s'han d'usar quan es maximitzin les conseqüències positives (beneficis) i es minimitzin les negatives (costos) derivades de la seva aplicació.

Els tests s'apliquen esperant aconseguir algun tipus de benefici (poder seleccionar el millor tractament terapèutic, situar els treballadors d'una empresa en el lloc més adequat, millorar les tècniques didàctiques emprades, etc.) de la informació obtinguda. Un dels propòsits fonamentals de la validació és indicar en quins casos es poden obtenir aquests beneficis.

Dins d'aquest concepte cal diferenciar entre evidències que són rellevants per a la validesa i evidències que són importants per a les polítiques socials però se situen fora del concepte de *validesa*. Aquesta diferència es fa més important quan les conseqüències que es deriven del test són diferents per a diferents grups. Per exemple, si se sap que hi ha diferències entre homes i dones en les puntuacions d'un test emprat per a la selecció de personal, això afectarà l'ús del test, però no es podria afirmar res sobre les evidències de validesa basades en les conseqüències de l'aplicació. Per a això s'ha de fer un estudi més detallat de les conseqüències. Si les diferències es deuen al fet que l'aspecte avaluat es distribueix de manera diferent entre els grups en la població, les diferències obtingudes no impliquen que les decisions que es derivin de l'aplicació del test manquin de validesa. El problema sorgeix quan aquestes diferències es deuen al fet que s'estan valorant habilitats que no estan relacionades amb la tasca que duran a terme els seleccionats o quan el test és sensible a algunes característiques dels candidats que no es vol que estiguin relacionades amb el constructe a mesurar. En el primer cas, no es pot concloure la manca d'indicis de validesa respecte a les conseqüències si bé en les dues últimes situacions sí. Però, evidentment, les tres situacions serien inadequades dins de les polítiques socials d'igualtat de gènere (APA, 1999).

7. Factors que afecten la validesa

Tal com s'ha comentat més amunt, un dels indicis de validesa que es poden calcular (o s'han de calcular) és la correlació entre el test i un criteri aliè a aquest. Aquesta correlació es pot veure afectada per múltiples factors com són la fiabilitat de totes dues mesures, la longitud del test i la variabilitat (dispersió) de la mostra emprada per a obtenir les puntuacions. A continuació es veuran aquests aspectes.

7.1. Fórmules d'atenuació

Quan es tracta de calcular la correlació entre un test i un criteri es parteix de les puntuacions empíriques que s'han obtingut en tots dos qüestionaris. Aquestes puntuacions estan compostes, segons el model lineal de Spearman, per la puntuació veritable i l'error de mesura que és aleatori. Per tant:

$$X = V_x + e_x$$
$$Y = V_y + e_y$$

Lectura de la fórmula

X: puntuació obtinguda en el test.

V_x: puntuació veritable en el test.

e_x: error de mesura (aleatori) en el test.

Y: puntuació obtinguda en el criteri.

V_y: puntuació veritable en el criteri.

e_y: error de mesura (aleatori) en el criteri.

Com es pot comprovar, en correlacionar les puntuacions X i Y també es correlacionen els dos errors de mesura entre ells. Aquests errors són aleatoris, la qual cosa significa que la correlació entre tots dos ha de ser igual a 0. Per això, com més importància tinguin els errors en la puntuació obtinguda (o, el que és el mateix, com més baixa sigui la fiabilitat del test i el criteri emprats) més baixa serà la correlació entre X i Y. En definitiva, es pot trobar una regla segons la qual com més fiables siguin el test i el criteri, més gran serà la correlació entre tots dos. Per saber en quin grau ho fa s'empren les fórmules d'atenuació (Spearman, 1907).

7.1.1. Estimació del coeficient de validesa en el supòsit en què el test i el criteri tinguin una fiabilitat perfecta

En el cas en què se suposa que el test i el criteri tenen una fiabilitat perfecta s'assumeix que els errors de mesura són iguals a 0. En aquesta situació (en què el coeficient de fiabilitat és igual a 1), la puntuació empírica (X o Y segons es tracti del test o del criteri) és igual a la veritable.

En aquest cas, la nova correlació es pot calcular mitjançant:

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

Lectura de la fórmula

ρ_{xy}: coeficient de validesa obtingut en correlacionar les puntuacions del test i el criteri.

$\rho_{xx'}$: coeficient de fiabilitat del test.

$\rho_{yy'}$: coeficient de fiabilitat del criteri.

Exemple

La correlació entre un test d'ansietat i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79 i la del criteri de 0,90. Quina és l'estimació d'aquesta correlació si se suposa que tots dos tenen una fiabilitat perfecta?

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} = \frac{0,63}{\sqrt{0,79}\sqrt{0,90}} = 0,75$$

La correlació que s'estima entre el test i el criteri si tots dos tinguessin una fiabilitat perfecta passa de 0,63 a 0,75.

7.1.2. Estimació del coeficient de validesa en el supòsit en què el test tingui una fiabilitat perfecta

En el cas en què només el test tingui una fiabilitat perfecta (igual a 1), l'estimació de la nova correlació queda determinada per:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}}$$

En l'exemple anterior, si només el test tingués la fiabilitat perfecta el resultat seria:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}} = \frac{0,63}{\sqrt{0,79}} = 0,71$$

La correlació que s'estima entre el test i el criteri en suposar que el test té una fiabilitat perfecta canvia de 0,63 a 0,71.

7.1.3. Estimació del coeficient de validesa en el supòsit en què el criteri tingui una fiabilitat perfecta

Si és el criteri el que se suposa que té una fiabilitat perfecta, l'estimació de la correlació queda determinada per:

$$\rho_{x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}}$$

En l'exemple anterior:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}} = \frac{0,63}{\sqrt{0,90}} = 0,66$$

La correlació que s'estima entre el test i el criteri en suposar que el test té una fiabilitat perfecta canvia de 0,63 a 0,66.

7.1.4. Estimació del coeficient de validesa en el supòsit en què s'hagi millorat tant la fiabilitat del test com la del criteri

La situació més freqüent és la situació en què es millora la fiabilitat del test, la del criteri o la de tots dos però sense arribar a 1 (aquest esdeveniment és més teòric que pràctic). A continuació es veurà com s'estima la correlació entre test i criteri quan es milloren les fiabilitats de tots dos. Posteriorment es tractaran les altres opcions.

$$\rho_{XY} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

Lectura de la fórmula

$\rho_{XX'}$: fiabilitat millorada del test.

$\rho_{YY'}$: fiabilitat millorada del criteri.

Exemple

La correlació entre un test d'ansietat i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79 i la del criteri de 0,90. Afegint ítems s'aconsegueix incrementar la fiabilitat del test fins a 0,83 i la del criteri fins a 0,92. Quina és l'estimació d'aquesta correlació després d'haver millorat la fiabilitat de tots dos?

$$\rho_{XY} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} = \frac{0{,}63\sqrt{0{,}83}\sqrt{0{,}92}}{\sqrt{0{,}79}\sqrt{0{,}90}} = 0{,}65$$

L'estimació de la correlació entre el test i el criteri després d'haver millorat la fiabilitat de tots dos passa de 0,63 a 0,65.

7.1.5. Estimació del coeficient de validesa en el supòsit en què s'hagi millorat la fiabilitat del test

Quan només es millora la fiabilitat del test, l'estimació del nou coeficient de correlació queda determinada per:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}}$$

En l'exemple anterior, si només es millora la fiabilitat del test:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}} = \frac{0{,}63\sqrt{0{,}83}}{\sqrt{0{,}79}} = 0{,}645$$

L'estimació de la correlació entre el test i el criteri després d'haver millorat la fiabilitat del test passa de 0,63 a 0,645.

7.1.6. Estimació del coeficient de validesa en el supòsit en què s'hagi millorat la fiabilitat del criteri

L'estimació del coeficient de correlació quan només es millora la fiabilitat del criteri queda determinada per:

$$\rho_{xY} = \frac{\rho_{xy}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}}$$

En l'exemple anterior, si només es millora la fiabilitat del criteri la resposta seria:

$$\rho_{xY} = \frac{\rho_{xy}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}} = \frac{0.63\sqrt{0,92}}{\sqrt{0,90}} = 0,637$$

L'estimació de la correlació entre el test i el criteri després d'haver millorat la fiabilitat del criteri passa de 0,63 a 0,637.

7.1.7. Valor màxim que pot assolir el coeficient de correlació entre test i criteri

Com es pot apreciar, a mesura que incrementem el coeficient de correlació del test, del criteri o de tots dos, el coeficient de correlació augmenta. Això només passa fins a cert punt, ja que el coeficient de correlació entre test i criteri sempre és més baix o igual que el seu índex de fiabilitat ($\rho_{xv} = \sqrt{\rho_{xx'}}$).

Matemàticament es pot representar de la manera següent:

$$\rho_{xy} \leq \rho_{xv}$$

Per tant, en l'exemple anterior, el màxim coeficient de correlació que es pot obtenir entre el test i el criteri és:

$$\rho_{xv} = \sqrt{\rho_{xx'}} = \sqrt{0,79} = 0,89$$
$$\rho_{xy} \leq 0,89$$

Així doncs, el màxim valor del coeficient de correlació que es pot obtenir en aquest test és de 0,89.

7.2. Efecte de la longitud del test sobre el coeficient de correlació test-criteri

Un dels mitjans pels quals es pot incrementar el coeficient de correlació test-criteri és augmentant el nombre d'ítems que componen el test. La relació entre el nombre d'ítems i aquesta correlació és directa, és a dir, a mesura que s'incrementi el nombre d'ítems la correlació augmentarà (i es reduirà si es treuen ítems).

La relació entre tots dos queda determinada per:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}}$$

Lectura de la fórmula

ρ_{xy}: valor inicial de la correlació test-criteri.

$\rho_{xx'}$: coeficient de fiabilitat del test.

n: nombre de vegades que s'augmenta el test.

Exemple

La correlació entre un test d'ansietat de vint ítems i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79. Estimeu el valor de la correlació test-criteri si s'afegeixen deu ítems més.

$$n = \frac{20+10}{20} = 1{,}5$$

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}} = \frac{0{,}63\sqrt{1{,}5}}{\sqrt{1+(1{,}5-1)\,0{,}79}} = 0{,}65$$

En afegir deu ítems als vint originals, el coeficient de correlació s'incrementa des de 0,63 fins a 0,65.

Una altra possibilitat és el fet de voler arribar fins a un coeficient de correlació que es vulgui i, per tant, calgui calcular el nombre d'ítems que s'han d'afegir al qüestionari per a poder aconseguir-ho. Per a això:

$$n = \frac{\left(1-\rho_{xx'}\right)\rho^2_{Xy}}{\rho^2_{xy} - \rho^2_{Xy}\rho_{xx'}}$$

Lectura de la fórmula

ρ^2_{Xy}: quadrat del coeficient de correlació desitjat.

Exemple

La correlació entre un test d'ansietat de vint ítems i un criteri (depressió) és de 0,63. La fiabilitat del test és de 0,79. Quants ítems caldrà afegir al test si es vol aconseguir un coeficient de correlació test-criteri de 0,67?

$$n = \frac{\left(1-\rho_{xx'}\right)\rho^2_{Xy}}{\rho^2_{xy}-\rho^2_{Xy}\rho_{xx'}} = \frac{(1-0{,}79)0{,}67^2}{0{,}63^2-0{,}67^2 0{,}79} = 2{,}23$$

El test s'hauria d'incrementat 2,23 vegades. Atès que el test original té vint ítems, 20 × 2,23 = 44,6. Evidentment, no es poden tenir decimals (estem parlant d'ítems, no es pot tenir una porció d'ítem en el test), per la qual cosa ho hem d'ajustar. Quan afegim ítems l'ajust sempre s'ha de fer cap a l'enter superior, és a dir, en aquest cas hauria d'haver-hi quaranta-cinc ítems. En cas que el test sigui excessivament llarg i no ens importi eliminar ítems fins a arribar a un coeficient de correlació test-criteri més baix (passar de 0,63 a 0,50 per exemple), l'ajust s'haurà de fer a l'enter inferior.

7.3. Efecte de la variabilitat de la mostra en la correlació test-criteri

El coeficient de correlació es veu molt afectat per la dispersió de la mostra en què estigui calculat. La relació entre la dispersió i la correlació és directa: com més dispersió s'obtindrà més correlació.

En alguns camps de la psicologia és molt freqüent que només es pugui calcular la correlació entre el test i el criteri en una petita mostra de persones. L'exemple més clar és el de la selecció de personal. Després d'haver emprat un test per a seleccionar els candidats més adequats al lloc de treball, només

es pot correlacionar la puntuació que s'ha obtingut en el test amb un criteri com el de rendiment laboral. En aquest cas, la dispersió dels seleccionats serà menor que la del total de candidats, ja que se seleccionen les persones que tenen característiques molt similars (i que s'ajusten més a les que es busquen per al lloc de treball).

Després de calcular el coeficient de correlació en la mostra de seleccionats, pot interessar estimar quin seria si s'hagués calculat sobre la totalitat d'aspirants. Per a això s'ha de partir de dos supòsits: a) el pendent de la recta de regressió és el mateix per als dos grups (admesos i aspirants) i b) l'error típic d'estimació també és el mateix per a tots dos grups. Matemàticament:

$$a) \frac{\rho_{xy}\sigma_y}{\sigma_x} = \frac{\rho_{XY}\sigma_Y}{\sigma_X}$$

$$b) \sigma_y\sqrt{1-\rho_{xy}^2} = \sigma_Y\sqrt{1-\rho_{XY}^2}$$

En què les lletres majúscules fan referència al grup d'admesos, i les minúscules al total d'aspirants.

Per a estimar el valor de la correlació test-criteri en el total d'aspirants després d'haver-la calculat en el d'admesos només cal aplicar:

$$\rho_{xy} = \frac{\sigma_x \rho_{XY}}{\sqrt{\sigma_x^2 \rho_{XY}^2 + \sigma_X^2 - \sigma_X^2 \rho_{XY}^2}}$$

Lectura de la fórmula

ρ_{XY}: coeficient de correlació test-criteri en la mostra d'admesos.

σ_x^2: variància en el test del total d'aspirants.

σ_X^2: variància en el test dels admesos.

Exemple

Es va aplicar un test d'assertivitat a mil persones per a seleccionar deu sobrecàrrecs de vol. La desviació típica que es va obtenir en el test pel total d'aspirants va ser de 15, i en la mostra d'admesos, de 4. Després d'un quant temps treballant es va comprovar que la correlació entre les puntuacions en el test i l'execució laboral (valorada pels superiors) va ser de 0,36. Quin seria el coeficient de correlació test-criteri si s'hagués calculat sobre el total dels aspirants?

$$\rho_{xy} = \frac{\sigma_x \rho_{XY}}{\sqrt{\sigma_x^2 \rho_{XY}^2 + \sigma_X^2 - \sigma_X^2 \rho_{XY}^2}} = \frac{15 \times 0{,}36}{\sqrt{15^2 \times 0{,}36^2 + 4^2 - (4^2 \times 0{,}36^2)}} = 0{,}82$$

Com es pot veure, hi ha un increment notable en el valor del coeficient de correlació a causa de la dispersió diferent que tenen els grups.

Bibliografia

AERA, A. N. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

AERA, APA, i NCME (1966). *Standards for educational and psychological test and manuals*. Washington, DC: AERA.

AERA, APA, i NCME (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Anastasi, A. (1954). *Psychological testing*. Nova York: Macmillan.

Beck, A., Rush, A. J., Shawn, B. F., i Emery, G. (1979). *Cognitive therapy of depression*. Nova York: Guilford Press.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. Oxford, Anglaterra: Harpers.

Campbell, D. T. i Fiske, A. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cronbach, L. J. (1971). Test validation. A R. L. Thorndike. *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. i Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Czaja, R. i Blair, J. (1996). *Designing Surveys*. Thousand Oaks, CA: Sage Publications.

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16, 640-647.

Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 427-439.

Hamilton, M. (1960). A rating scale for depression. *Journal Neurol.Neurosurg. Psychiatry*, 23, 56-62.

Kane, M. (2006). Content-Related Validity Evidence in Test Development. A S. M. Downing, i T. M. Haladyna. *Handbook of test development* (pp. 131-153). Nova Jersey: Erlbaum.

Leighton, J. P. (2004). Avoiding misconception, missed use, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6-15.

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

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

Prieto, G. i Delgado, A. R. (2010). Fiabilidad y Validez. *Papeles del Psicólogo*, 31, 67-74.

Ramos-Brieva, J. C. (1986). Validación de la versión castellana de la escala de Hamilton para la depresión. *Actas Luso-Esp. Neurol. Psiquiatr.*, 22, 21-28.

Scott, W. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology*, 61-66.

Shepard, L. (1993). Evaluating test validity. A L. Darling-Hammond. *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-13.

Shepard, L. A., Camilli, G., Linn, R., i Bohrnstedt, G. (1993). *Setting performance standards for achievement tests*. Stanford, CA: National Academy of Education.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research*, 45, 83-117.

Sireci, S. G. (1998). Gathering and evaluating content validity data. *Educational Assessment*, 5(4), 299-321.

Spearman, C. (1907). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.