Llengua i Ús
Revista Tècnica de Política Lingüística
56
1r SEMESTRE 2015
ISSN: 2013-052X
http://gencat.cat/llengua/liu

Models

1.1

L'avaluació de la competència lingüística comunicativa: principis bàsics i propietats dels exàmens

Era avaloracion dera competéncia lingüsitica comunicativa: principis basics e proprietats des examens

La evaluación de la competencia lingüística comunicativa: principios básicos y propiedades de los exámenes

Assessment of language communication skills: basic principles and properties of exams

Laura Puigdomènech Farell
Direcció General de Política Lingüística

Citació recomanada:
Puigdomènech i Farell, Laura. «*L'avaluació de la competència lingüística comunicativa: principis bàsics i propietats dels exàmens*». *Llengua i Ús: Revista Tècnica de Política Lingüística* [en línia] [Barcelona: Generalitat de Catalunya. Departament de Cultura. Direcció General de Política Lingüística], núm. 56. p. 4-14.

El català *Llengua per a tothom*
Generalitat de Catalunya
Departament de Cultura

L'avaluació de la competència lingüística comunicativa: principis bàsics i propietats dels exàmens

exàmens, constructe, validesa, fiabilitat, inferència, estàndards, dificultat, discriminació, punt de tall, barems

L'objectiu d'aquest article és aportar una mica de llum al món de l'avaluació, concretament, a les proves que avaluen la competència lingüística comunicativa. En aquest sentit, s'hi expliquen els principis bàsics en què se sostenen aquesta mena de proves i les propietats que han de complir perquè siguin vàlides i fiables.

La evaluación de la competencia lingüística comunicativa: principios básicos y propiedades de los exámenes

exámenes, constructo, validez, fiabilidad, inferencia, estándares, dificultad, discriminación, punto de corte, baremos

El objetivo de este artículo es aportar algo de luz al mundo de la evaluación, concretamente, a las pruebas que evalúan la competencia lingüística comunicativa. En este sentido, se explican los principios básicos en que se sustentan este tipo de pruebas y las propiedades que deben cumplir para ser válidas y fiables.

Era avaloracion dera competéncia lingüsitica comunicativa: principis basics e proprietats des examens

examens, constructe, validesa, fiabilitat, inferéncia, estandards, dificultat, discriminacion, punt de talh, barems

Er objectiu d'aguest article ei aportar un shinhau de lum ath mon dera avaloracion, concrètament, enes pròves qu'avaloren era competéncia lingüistica comunicativa. En aguest sentit, s'i expliquen es principis basics en qué se sostien aguesta sòrta de pròves e es proprietats qu'an de complir tà pr'amor de que siguen valides e fiables.

Assessment of language communication skills: basic principles and properties of exams

examinations, construct, validity, reliability, inference, standards, difficulty, discrimination, cut-off point, scales

The purpose of this article is to cast a little light on the world of assessment, more specifically on tests to assess language communication skills. To this end, we describe the basic principles on which such tests are based, and they properties that they must possess in order to be valid and reliable.

No deixa de ser curiós el fet que el món de l'avaluació sigui tan desconegut, quan l'avaluació, en qualitat d'avaluats, forma part integral de la nostra biografia. Intenteu fer un recompte dels exàmens que heu hagut de fer al llarg de la vostra vida: a l'escola primària i a secundària, a la universitat, a l'acadèmia d'idiomes on vau estudiar anglès, per treure-us el carnet de conduir, per poder obtenir certa habilitació professional… Us havíeu preguntat mai si els exàmens que us feien fer estaven ben dissenyats?, si avaluaven el que deien que avaluaven?, si us avaluaven d'una manera justa? L'objectiu d'aquest article és aportar una mica de llum a tot aquest univers, concretament, a les proves que avaluen la competència lingüística comunicativa. En aquest sentit, s'hi explicaran els principis bàsics en què se sostenen aquesta mena de proves i les propietats que han de complir perquè siguin vàlides i fiables.

Constructe

El constructe és la **base teòrica** que hi ha darrere d'una prova: la descripció de la finalitat i la utilitat social que en motiva la creació, l'explicitació del model teòric de competència lingüística i d'ús lingüístic de què es parteix, la definició dels destinataris diana, la tria de l'àmbit o els àmbits d'ús lingüístic a què la prova pretén donar cobertura i la detecció de les necessitats lingüístiques dels destinataris dins d'aquests àmbits, l'adaptació del nivell del Marc europeu comú de referència per a les llengües (d'ara endavant, MECR) que es vol avaluar d'acord amb la finalitat de la prova i les necessitats lingüístiques que s'han seleccionat, si la prova estarà basada en tasques o bé en exercicis (i per què), quines habilitats i competències s'avaluaran (i per què), si les habilitats s'avaluaran de manera separada o bé de manera integrada… Sense constructe, sense fonaments, una prova és un gegant amb peus de fang.

Observació

El constructe acostuma a prendre forma de document intern de les organitzacions. Alguns organismes avaluadors, però, especialment els que tenen finalitat de lucre, posen a l'abast els seus constructes a totes les persones interessades, com a segell distintiu de transparència i de qualitat.

Validesa

La normativa internacional sobre proves educatives i psicològiques estableix que una prova és vàlida si tant la teoria com les evidències confirmen les interpretacions que fem de les puntuacions obtingudes. Parafrasejant aquesta definició un xic críptica, es tracta de demostrar aportant-hi evidències que la prova, globalment, implementa el constructe dissenyat; això comprèn els exercicis i les tasques, els criteris d'avaluació i els barems, l'administració de la prova i els procediments d'avaluació i correcció. La validesa, per tant, fa referència a l'**adequació** de tot l'instrument de mesura al que se suposa que realment mesura. En tota prova, però especialment en les proves d'alt impacte, en què els resultats tenen conseqüències importants per als examinands (passar de curs a la universitat, superar unes oposicions, obtenir un certificat de llengua que permet accedir a un lloc de treball concret, proves de llengua per obtenir permís de residència o de nacionalitat…), hem d'estar en condicions de demostrar la validesa de les nostres proves.

A partir d'ara, tots els conceptes i totes les propietats que es comentaran contribueixen a conferir **validesa** a una prova, sempre que es compleixin.

Exemple

Hem dissenyat una prova de final de curs del nivell B2. Volem que ens aporti informació sobre els aprenents que han superat el curs i que poden passar al següent (C1) i els que no. Decidim que hi inclourem les habilitats d'expressió escrita, expressió oral, comprensió oral i comprensió lectora, i elaborem tasques i exercicis adequats per avaluar el nivell B2; situem la nota de tall per superar cada àrea en 60 punts sobre 100. Un cop corregides les proves, ens adonem que el 50 % de l'alumnat supera les àrees d'expressió escrita, expressió oral i comprensió oral, mentre que l'àrea de comprensió lectora presenta un índex d'èxit molt superior, el 95 %. Les evidències apunten un possible problema de validesa en l'àrea de comprensió lectora, que el desenvolupador haurà d'indagar:

- Els textos de comprensió lectora han resultat massa fàcils (de dificultat inferior al nivell B2)?
- El format de l'exercici no permet dissenyar ítems de dificultat adequada per al nivell B2?
- Hi ha hagut ítems interdependents (la resposta correcta d'un ítem revelava la resposta correcta d'un altre)?
- La nota de tall de l'àrea és massa baixa?

Inferència

La finalitat d'una prova de llengua és obtenir prou dades lingüístiques dels examinands per poder fer **judicis** sobre la seva competència comunicativa. Es parteix que l'actuació dels examinands en una prova s'ha de poder interpretar com un indicador vàlid d'allò que poden fer en un context real. A partir del que fan en la prova i de com ho avaluem, fem inferències, extrapolacions, sobre el seu domini de la llengua i sobre les tasques i les activitats lingüístiques que seran capaços d'afrontar, i es prenen decisions.

Observació

Les inferències són inherents a tot tipus de proves: proves per poder fer de controlador de vol, proves per poder exercir de cirurgians, proves per superar la carrera de jutge… És obvi que la validesa de les inferències que fem a partir de les actuacions dels examinands en les proves que s'acaben d'esmentar és crítica, no ens podem permetre el luxe de prendre decisions equivocades. En proves de llengua, les conseqüències de fer inferències errònies no ho acostumen a ser tant, de crítiques, però ho poden arribar a ser, i molt. Un exemple flagrant són les proves de llengua que es fan en molts països europeus per poder entrar al país, aconseguir un permís de residència o obtenir la ciutadania.

Àmbit de generalització

L'àmbit de generalització (*target language use domain*) es defineix com les situacions comunicatives i els àmbits d'ús que cobreix i avalua la prova i en què l'examinand utilitzarà la llengua fora del context d'examen.

Observacións

- Les proves de nivells baixos del MECR (A1, A2, B1) adreçades a la població en general sovint cobreixen els àmbits personal i públic, mentre que les proves d'aquests nivells adreçades al col·lectiu immigrant s'acostumen a centrar en els àmbits laboral i públic, i es deixa força de banda l'àmbit personal, atès que en les seves relacions interpersonals de caràcter privat la llengua que hi preval és la d'origen.
- Arran de la globalització i la mobilitat per motius laborals i acadèmics, han emergit amb força les proves de llengua per a finalitats específiques (*language testing for specific purposes*): B2 d'alemany mèdic per a metges que volen exercir a Alemanya, C1 d'anglès acadèmic per accedir a universitats britàniques o nord-americanes, C1 d'anglès per a la comunicació aeuronàutica internacional…

Rellevància i representativitat

Si volem utilitzar les puntuacions obtingudes d'una prova per fer inferències sobre la competència lingüística dels examinands i prendre decisions vàlides, hem de poder **demostrar** com es relaciona l'actuació lingüística en la prova amb els usos lingüístics fora d'un context d'examen. Per establir aquesta correlació, cal seleccionar i dissenyar tasques, exercicis i ítems rellevants i representatius de tot l'univers de situacions comunicatives que, d'acord amb el constructe, els examinands han de poder afrontar.

Les tasques i els exercicis d'una prova són **rellevants** si les **habilitats** que avaluen són **necessàries** per dur a terme les activitats lingüístiques dins l'àmbit de generalització. D'altra banda, les tasques i els exercicis són **representatius** si constitueixen una **selecció representativa** de les activitats comunicatives pròpies de l'àmbit de generalització.

Exemples

- Si haguéssim de desenvolupar una prova de nivell B1 de català per a atenció al públic, l'expressió oral i la comprensió oral haurien de ser habilitats rellevants, mentre que la presència de l'expressió escrita hauria de ser testimonial, si no inexistent.

- Fins al 1998, la coneguda prova d'anglès acadèmic TOEFL avaluava l'habilitat d'expressió escrita mitjançant un exercici amb ítems de resposta múltiple. La representativitat de l'exercici era, doncs, nul·la.

Autenticitat de la situació i de la interacció

En sintonia amb el model de competència lingüística que el MECR propugna i l'estratègia pedagògica de l'ensenyament/aprenentatge basat en tasques, cada vegada més les proves de llengua adopten un enfocament competencial i mesuren la capacitat dels examinands per dur a terme **tasques** que requereixen l'ús de la llengua. És important que aquestes tasques compleixin dos aspectes importants: l'autenticitat de la situació i l'autenticitat de la interacció.

L'**autenticitat situacional** fa referència a la **similitud** amb què les tasques d'una prova reprodueixen activitats lingüístiques de la vida real. Es tracta, doncs, de detectar els trets característics de les tasques que l'examinand haurà d'executar i de reproduir-los en tasques simulades amb finalitats avaluatives. L'**autenticitat interaccional** es refereix a la **naturalitat** amb què l'examinand aborda i resol una tasca. Idealment, l'examinand hauria de seguir i reproduir els mateixos processos cognitius que empraria en activitats comunicatives similars de la vida real.

Observació

L'autenticitat interaccional s'aconsegueix si

- les tasques estan ben contextualitzades (identitat de l'emissor i rol que adopta, finalitat per comunicar, tipus de text, audiència o lectors potencials…);

- l'examinand aplica estratègies com ara la planificació i la revisió;

- l'examinand coneix per endavant els criteris amb què serà avaluat.

Sobrerepresentació i infrarepresentació del constructe

Un constructe està sobrerepresentat si hi ha exercicis o tasques d'una prova que mesuren essencialment les mateixes habilitats. Paral·lelament, un constructe està infrarepresentat si exercicis o tasques que considerem rellevants i representatius de l'àmbit de generalització no

tenen cabuda a la prova. Tant la sobrerepresentació com la infrarepresentació poden posar en dubte la validesa de les inferències que fem sobre la competència lingüística dels examinands en els àmbits d'ús que avalua la prova.

Exemple

Hem dissenyat una prova de català per a l'àmbit acadèmic (capacitat dels examinands per poder seguir cursos universitaris que tinguin el català com a llengua vehicular). Per avaluar l'habilitat de comprensió oral, hem decidit que la tasca consistirà a escoltar una conversa entre dos amics parlant de les vacances. Aquí tenim un problema d'infrarepresentació del constructe, perquè una conversa informal no és representativa ni rellevant de l'àmbit de generalització.

Establiment d'estàndards

L'establiment d'estàndards per a les àrees de producció (expressió escrita i expressió oral) sempre comença identificant i definint els trets distintius de les actuacions dels aprenents fronterers (els aprenents que presenten el nivell de competència mínim que es considera acceptable per al nivell del MECR que s'avalua), dels aprenents bons i dels aprenents fluixos. La fixació d'estàndards ha de ser fruit del consens d'un equip d'experts, i això s'aconsegueix estandarditzant judicis (ens posem d'acord en què constituiria una bona mostra, una mostra fronterera…) i, seguidament, estandarditzant mostres (mostres avaluades i comentades que exemplifiquen actuacions fluixes, frontereres i bones). Aquests estàndards també queden reflectits en els descriptors dels barems que utilitzem per avaluar expressions escrites i expressions orals. Finalment, cal fixar el punt de tall, la nota de tall per a cada tasca i àrea, que en proves en què s'avalua per criteris s'acostuma a situar entre el descriptor del barem corresponent a l'aprenent fronterer i el descriptor corresponent a l'aprenent fluix.

Si les habilitats receptives (comprensió lectora i comprensió oral) o bé els coneixements gramaticals i lèxics es volen avaluar a partir d'exercicis amb ítems discrets, també s'han d'establir estàndards. Cal que un grup de persones implicades en desenvolupament d'exàmens i amb un bon coneixement del MECR emetin judicis qualitatius sobre la dificultat dels exercicis dissenyats: d'una banda, han de determinar si els diferents ítems i exercicis de la prova pertanyen al nivell que es vol avaluar o si més aviat corresponen a l'anterior o al posterior. Un cop decidit, la fase següent consisteix a mesurar-los, és a dir, determinar-ne el grau de dificultat. Una manera de fer-ho és formulant a l'equip d'experts la pregunta següent: quin percentatge de probabilitats hi ha que un examinand fronterer del nivell que es vol avaluar contesti correctament l'ítem X?; i així, ítem per ítem. Una altra manera més pràctica de fer-ho és demanar al grup d'experts que analitzin cada exercici i que estimin quants ítems, com a mínim, consideren que un examinand amb el nivell adequat hauria de poder resoldre. Aleshores es fixa un punt de tall provisional de l'exercici.

En tots dos processos, els estàndards fixats sempre s'han de validar amb aprenents de característiques conegudes (és a dir, amb aprenents el nivell dels quals ja coneixem). D'aquesta manera, tindrem una idea força ajustada del percentatge d'examinands que superarien la prova i també ens ajudarà a establir els punts de tall. Un cop fets els reajustaments necessaris, podrem fixar el punt de tall definitiu en exercicis, àrees i el conjunt de la prova.

Manteniment d'estàndards

Cal vetllar perquè els estàndards fixats es mantinguin de convocatòria en convocatòria; és a dir, cal garantir que les diferents versions d'una prova siguin comparables (similars) quant a dificultat, discriminació i fiabilitat. Això també contribuirà que el punt de tall establert en la nova prova sigui estable. Si les condicions d'examen i les propietats psi-

comètriques no són comparables (p. ex., una versió de la prova ha resultat més difícil que l'altra), el punt de tall fixat inicialment per a la prova perd tota la validesa, les puntuacions obtingudes pels examinands en una prova i en l'altra no són equiparables (perquè els uns ho van tenir més difícil que els altres) i, per tant, es vulnera el principi bàsic de l'avaluació: la justícia dels resultats. El manteniment d'estàndards s'assegura, entre d'altres coses, amb l'assaig previ d'ítems i tasques (pretest), analitzant els resultats psicomètrics d'una prova *a posteriori*, creant especificacions de prova i de tasques detallades i atenint-nos-hi en el procés cíclic d'elaboració de versions de prova, redactant orientacions i criteris per als elaboradors de proves i oferint formació continuada a correctors i examinadors.

Pretest

El pretest consisteix a fer un assaig previ d'ítems d'una prova ja estandarditzada abans d'una convocatòria amb aprenents de característiques similars a les dels destinataris de la prova. Aquest assaig serveix per comprovar si els ítems «funcionen»: que avaluïn l'estructura prevista, que tinguin el grau esperat quant a dificultat, que discriminin… Paral·lelament, l'anàlisi psicomètrica dels resultats del pretest permet mantenir els estàndards fixats per a àrees i exercicis; per exemple, si detectem que un exercici ha resultat ser massa difícil, podem decidir eliminar-lo i substituir-lo per un altre. Penseu que una informació psicomètrica bàsica es pot obtenir fàcilment amb programes de càlcul com ara Excel. Si el pretest s'ha dut a terme amb un nombre suficient d'examinands perquè sigui estadísticament informatiu, els ítems, juntament amb les seves propietats psicomètriques, es poden desar dins un banc d'ítems, una eina de gestió que facilita l'elaboració de proves estandarditzades. Quant a les àrees de correcció subjectiva (expressió escrita i expressió oral), també és recomanable assajar prèviament les tasques per verificar de manera qualitativa que generen les actuacions esperades.

Les organitzacions petites, però, no acostumen a tenir la infraestructura ni els recursos necessaris per dur a terme pretests, per la qual cosa analitzen les propietats psicomètriques de l'examen després d'haver-lo administrat i corregit, i tenen en compte la informació obtinguda a l'hora d'elaborar-ne noves versions. De tota manera, encara que sigui a petita escala, sempre és recomanable provar abans els ítems i les tasques.

> **Observació**
>
> Per més que una prova hagi estat elaborada de manera meticulosa per persones expertes, mai no es pot saber com funcionarà un ítem o una tasca fins que no s'hagi assajat. Us convido que en feu la comprovació: passeu un ítem entre diferents elaboradors i pregunteu-los quin nivell de dificultat presenta i què és el que pretén avaluar. Us sorprendran les discrepàncies que apareixen.

Dificultat

La dificultat d'una prova depèn de diversos paràmetres: *a*) dels ítems o les tasques concretes que conté, és a dir, dels continguts; *b*) dels barems dissenyats per avaluar les tasques de producció; *c*) del format de les tasques i els exercicis; *d*) de les propietats concretes dels *inputs* de les tasques (la informació de partida que té l'examinand), i *e*) de les propietats concretes dels *outputs* de les tasques (el que demanem a l'examinand que produeixi).

Normalment les organitzacions calculen per a ítems i exercicis l'índex de dificultat (valor *p*), que és la proporció d'examinands que contesten correctament un ítem. Cal tenir present que com més alt es l'índex de dificultat (que va de 0 a 1) més fàcil és l'ítem. En general, es considera que el valor «ideal» de dificultat dels ítems és 0,50, perquè faciliten una major gamma de variació entre examinands individuals i també són els que millor discriminen.

Observació

El grau de dificultat d'una prova sempre s'estableix a priori (mitjançant processos d'establiment d'estàndards) i queda reflectit en els diferents punts de tall que fixem per a exercicis, tasques, àrees i el conjunt de la prova. En proves en què s'avalua per criteris, els exercicis i les tasques han de tenir el **grau just de dificultat** perquè puguin discriminar entorn de l'aprenent fronterer: es tracta que els aprenents fluixos trobin la prova difícil, que els aprenents bons trobin la prova assequible i que els aprenents fronterers tinguin problemes per resoldre-la però que, més o menys, se'n surtin.

Discriminació

Els índexs de discriminació indiquen la capacitat d'un ítem per classificar, per distingir entre els examinands bons i fronterers dels fluixos i molt fluixos. L'índex de dificultat no és suficient a l'hora de valorar la qualitat d'un ítem. Si bé la capacitat discriminatòria d'un ítem té molt a veure amb el seu grau de dificultat, també hi pot influir el format amb què el presentem, la tria encertada dels distractors (en el cas d'exercicis d'elecció múltiple), el fet que hi pugui haver ítems interdependents (p. ex., que algun ítem doni pistes per resoldre'n un altre), el fet que alguns examinands tinguin coneixements previs del tema del text (si l'ítem és de comprensió lectora), la presència d'ítems o d'enunciats ambigus, que l'encert de l'ítem sigui degut a l'atzar o a la casualitat…

L'índex de discriminació no tan sols es calcula per a cada ítem, sinó també per a cada exercici, cada tasca i cada àrea de la prova. Els índexs de discriminació són tant o més importants que els de dificultat. En aquest sentit, si una àrea, un exercici, una tasca… no demostra tenir poder de discriminació, d'entrada no hauria de tenir cabuda en la prova.

Hi ha diferents índexs per calcular la discriminació (que van de 0 a 1); tot i que depèn de l'índex utilitzat, en general es considera que ítems amb valors de 0,30 en amunt tenen una capacitat discriminatòria bona. Cal tenir en compte que els ítems molt fàcils i molt difícils mai no poden ser, per definició, gaire discriminatoris.

Exemple

Considerem una situació hipotètica en què la meitat dels examinands responen correctament un ítem i l'altra meitat el responen incorrectament. Basant-nos només en l'índex de dificultat, podríem concloure que es tracta d'un ítem ideal. Però si descobrim que la meitat que ha respost correctament l'ítem està formada pels examinands fluixos i molt fluixos i la meitat que l'ha respost incorrectament està formada pels examinands bons i fronterers, la consideració sobre l'adequació d'aquest ítem canviarà: es tracta d'un ítem amb un índex de dificultat bo, però amb un poder de discriminació nul.

Fiabilitat

La fiabilitat és un concepte clau en qualsevol sistema d'avaluació. Es refereix al grau en què les puntuacions i els resultats d'un examen són estables, consistents i lliures d'error de mesura. En altres paraules: fins a quin punt podem confiar en els resultats d'un examen per prendre decisions correctes i adequades? Per exemple, podem estar segurs que la puntuació que ha obtingut un examinand és la que realment li correspon o bé és una puntuació inferior a la *real*?, o superior? Podem assegurar que un examinand que s'hagués presentat a les convocatòries 2012 i 2013 d'un mateix certificat obtindria puntuacions similars? Podem garantir que dues actuacions d'expressió escrita de característiques similars obtindran la mateixa puntuació si són avaluades per dos correctors diferents, o podria passar que amb un corrector un exa-

minand fos considerat apte en expressió escrita i amb un altre no? Podem estar segurs que un corrector atorgaria la mateixa puntuació a un text si el tornés a corregir passades unes quantes setmanes? En qualsevol d'aquests casos, si no en podem estar segurs, la decisió presa pot ser injusta. Per tant, una prova, a més de vàlida, ha de ser fiable.

Les estimacions de fiabilitat es fonamenten en una teoria sobre la mesura, la teoria de la puntuació certa. Aplicada als exàmens de llengua, aquesta teoria assumeix que la puntuació obtinguda en un examen (puntuació observada) sempre és la suma de dos components: una puntuació deguda a la capacitat real de l'examinand (puntuació certa) i una puntuació deguda a altres factors, no atribuïbles al seu domini de la llengua (puntuació error).

Exemple

D'acord amb aquesta teoria, un examinand que ha obtingut 70 punts en un examen (puntuació observada) podria realment merèixer 75 punts (puntuació certa); és a dir, podria ser que l'examinand fos millor que el que la puntuació observada indica. En aquest cas, la puntuació error seria de –5. Potser això ha estat així perquè aquell dia estava cansat, o no havia esmorzat, o li va corregir l'examen un corrector molt estricte, o la tasca estava mal dissenyada, o la instrucció era ambigua, o l'enregistrament de la prova de comprensió oral tenia poca qualitat, o la versió d'examen d'aquella convocatòria tenia més dificultat… I també, perquè no hi ha cap instrument de mesura perfecte.

Fiabilitat intercorrectors i intracorrectors

Quan en la correcció de la prova intervé la valoració humana, sempre hi ha el perill que es cometin errors de judici respecte als estàndards establerts. Anomenem *fiabilitat intercorrectors* el grau d'estabilitat i consistència en les correccions de tot el col·lectiu de correctors (si tots els correctors corregissin la mateixa mostra d'expressió escrita, atorgarien la mateixa puntuació en tots els conceptes avaluatius?), mentre que anomenem *fiabilitat intracorrectors* el grau de consistència en les correccions que fa una mateixa persona (si un corrector corregís una mateixa mostra durant cinc anys seguits —sense ser-ne conscient—, sempre hi atorgaria la mateixa puntuació?).

La fiabilitat de les puntuacions pot quedar compromesa per diferents causes. Les més freqüents entre els avaluadors acostumen a ser les següents:

- Indulgència: tendència a assignar puntuacions altes a tots els examinands.
- Severitat: tendència a assignar puntuacions baixes a tots els examinands.
- Efecte halo: la valoració d'un concepte (p. ex., correcció) influeix en la valoració d'un altre concepte (p. ex., sofisticació lingüística), tant a l'alça com a la baixa.
- Efecte biaix: l'avaluador és particularment indulgent o particularment sever respecte a un concepte avaluatiu concret, respecte a un grup d'examinands concret…
- Tendència central: sobreutilització de les bandes centrals dels barems i evitació de les bandes extremes.

Observació

La fiabilitat intracorrectors i intercorrectors s'assegura:

- elaborant i validant criteris i barems efectius i fàcils d'interpretar i d'utilitzar;
- fent una bona selecció del col·lectiu de correctors i proporcionant-los formació perquè entenguin, interpretin i apliquin els barems d'una manera correcta i consistent;
- establint processos de segona i tercera correcció de les proves.

Variància irrellevant per al constructe

En una prova vàlida, les diferències d'actuació entre els examinands, que queden reflectides en la puntuació final obtinguda, idealment haurien de ser un mirall de les seves diferències **reals** pel que fa a la seva competència lingüística comunicativa. Parlem de *variància irrellevant* quan la prova presenta variables que no formen part del constructe i que emmascaren, distorsionen, la competència real dels aprenents i, per tant, també les puntuacions finals. Les fonts de variància irrellevant poden ser molt diverses.

Exemples

- En una tasca d'expressió oral demanem als examinands que parlin sobre un tema del qual no tenen coneixements previs. Probablement, l'execució de la tasca se'n ressentirà i la puntuació que obtindran molts examinands en aquesta tasca serà més baixa que la que haurien obtingut amb un tema més al seu abast. El coneixement previ del tema de la tasca no formava part del nostre constructe però, en canvi, ha acabat afectant les puntuacions finals dels examinands.

- La complexitat lingüística emprada en un enunciat d'una tasca de nivell B1 pertany més aviat a un nivell C1. En aquest cas, hi ha el risc que molts examinands no entenguin bé l'enunciat i que això influeixi en la seva actuació.

Biaix

El biaix és una font de variància irrellevant per al constructe. Quan les diferències d'actuació entre examinands no són degudes a la seva habilitat real sinó al fet de pertànyer a algun col·lectiu en concret (sexe, ètnia, edat, nivell d'estudis, llengua primera…), es diu que hi ha biaix. Per tant, hi ha biaix quan un ítem, una tasca, un tema… perjudica un col·lectiu concret d'examinands i no els altres col·lectius.

Exemple

L'Agència Catalana de Cooperació al Desenvolupament necessitava contractar a Moçambic un tècnic en cooperació per coordinar els seus projectes a la zona. A l'aspirant, que era moçambiquès, se li va administrar una prova de català de nivell C1. La prova contenia, entre d'altres, l'ítem de derivació següent:

No t'oblidis de portar la mona al teu _____ (fill).

L'ítem, òbviament, presentava un problema de biaix cultural.

Nota final

Vetllar per la validesa i la fiabilitat de les nostres proves i, per tant, per l'equitat de l'instrument de mesura i per la justícia dels resultats és, al capdavall, una qüestió d'ètica. I això és, en si mateix, un gran repte.

Bibliografia

Alderson, J. Charles; Clapham, Caroline; Wall, Diane. *Exámenes de idiomas. Elaboración y evaluación*. Madrid: Cambridge University Press, 1998.

Bachman, Lyle F. *Statistical analyses for language assessment*. Cambridge: Cambridge University Press, 2004.

Bachman, Lyle F.; Palmer, Adrian S. *Language testing in practice*. Oxford: Oxford University Press, 1996.

Bachman, Lyle F.; Palmer, Adrian S. *Language assessment in practice*. Oxford: Oxford University Press, 2010.

Brown, James Dean; Hudson, Tom. *Criterion-referenced language testing*. Cambridge: Cambridge University Press, 2002.

Consell d'Europa. *Relating language examinations to the Common European Framework of Reference for Languages: A Manual* [en línia]. Estrasburg: Language Policy Division, 2009. <http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf> [Consulta: 4 febrer 2015]

Consell d'Europa. *Manual for language test development and examining*. Estrasburg: Language Policy Division. Council of Europe, 2011. També disponible en línia a: <http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf> [Consulta: 4 febrer 2015]

Martínez Arias, M. Rosario; Hernández Lloreda, M. Victoria; Hernández Lloreda, M. José. *Psicometría*. Madrid: Alianza Editorial, 2006.

Noijons, José; Bérešová, Jana; Breton, Gilles; Szabó, Gábor. *Relating language examinations to the Common European Framework of Reference for Languages (CEFR): highlights from the Manual* [en línia]. European Centre for Modern Languages: Council of Europe Publishing, 2011. <file:///C:/Documents%20and%20Settings/pollpf/Mis%20documentos/Downloads/2011_10_10_relex._E_web.pdf> [Consulta: 4 febrer 2015]

Viladrich, M. Carme; Doval Diéguez, Eduardo; Prat i Santolària, Remei; Llovet Vall-llovera, Montse. *Psicometría*. Barcelona: Editorial UOC, 2005.

Weir, Cyril J. *Language testing and validation: an evidence-based approach*. Oxford: Palgrave, 2005. També disponible en línia a: <http://pbi.fkip.untad.ac.id/wp-content/uploads/2014/09/Cyril_Weir_Language_Testing_and_Validation_An_E.pdf> [Consulta: 4 febrer 2015]