Azterketa informatizatu eraginkor baten bila

Looking for an effective computer test

Anaje Armendariz, Javier López-Cuadrado, Conchi Presedo*

Bilboko Ingeniaritza Eskola
(Bilboko Industria Ingeniaritza Teknikoko Unibertsitate Eskola).
Euskal Herriko Unibertsitatea

Tomás Pérez

Hizkuntza eta Sistema Informatikoak, Informatika Fakultatea.
Euskal Herriko Unibertsitatea

* anaje.armendariz@ehu.eus

DOI: 10.1387/ekaia.16368

Jasoa: 2016-05-31 Onartua: 2016-07-19

Laburpena: Gaur egun testak ohikoak dira, ezagutza ebaluatzeko garaian. Gehienetan, ebaluatzailearen esperientzian eta eskarmentuan oinarrituta hartzen dira azterketarako galdera edo itemak. Test Egokigarri Informatizatu (TEI) batean pertsona bakoitzak item ezberdinak ditu, aurreko erantzunen arabera. Hori lortu ahal izateko, item bakoitza kalibratu egin behar da, hau da, item bakoitzari balio batzuk eman behar zaizkio, prozesu jakin, zehatz bati jarraituz. Artikulu honetan azaltzen da zer egin behar den TEI bat osatzeko eta itemak kalibratzeko.

Hitz gakoak: ebaluazio, test egokigarri informatizatua, itemak, kalibraketa.

Abstract: Tests are a common method for assessing knowledge. In their classic version, they consist of fixed questions created according to the experience of the test developer. In computer adaptive tests (CAT) each person will have different items depending on her/his previous answers. To achieve this behavior, each item has to be calibrated. That is, certain characteristics have to be estimated using specific processes. This paper shows how to create a CAT and the necessary steps for calibrating the items.

Keywords: assessment, computerized adaptive test, items, calibration.

1. **SARRERA**

Gaur egun, irakaskuntza-sistema ugari ditugu, alor guztietako gaiak jorratzen dituztenak; Euskal Herriko Unibertsitatean bertan, Fakultate guztietan, eGela (Moodle) sistema erabiltzen da, materiala utzi, ikasleen artean harremana mantendu edota irakasleari zalantzak proposatzeko. Ez dira asko, ordea, ebaluazioa eskaintzen duten sistemak; batzuetan, irakaslearen konfiantza faltarengatik, eta, bestetan, berriz, sistemak berak ez duelako irakasgaiak behar duen ebaluazioa eskaintzen.

Testak dira gaur egun ebaluatzeko erabiltzen diren modurik zabalduenetariko bat. Ebaluatzeko modu hau testuinguru ezberdinetan erabiltzen da; adibidez, hizkuntzaren ezaguera neurtzen duten azterketa ofizialetan [1] edo baita on-line irakaskuntza-sisteman ere [2].

Arrazoi ugari daude ordenagailua erabiltzeko, azterketa bat egitean. Zenbait ikerketen arabera, test informatikoek abantaila asko dauzkate paperean egiten direnen aldean; besteak beste, azterketa gehiago egin daitezke, toki oso ezberdinetan eta aldi berean, horrela, kopia-kasuak saihestuz [3]. Erantzunak ere azkarrago prozesatzen dira, eta, ondorioz, ikasleek bukatu bezain laster izango dute probaren emaitza; gainera, galderak formatu egokiagotan egin daitezke, multimedia, elementu dinamikoak edota interaktiboak erabiliz. Bestalde, azterketak egiteko ordenagailua erabiltzeak ingurumenari ere laguntzen dio, papera erabiltzen ez den heinean, zuhaitz gutxiago moztu beharko dira eta.

Test informatizatu hauek bi erakoak izan daitezke: ohikoak, hau da, azterketa egingo duten pertsona guztiek azterketa bera izango dute, edo egokigarriak, hau da, azterketa egiten ari den pertsonak galderari ondo erantzuten badio, galdera zailagoa jarriko zaio; gaizki erantzuten badio, ordea, galdera errazagoa egingo zaio, duen ezagutza-maila finkatu arte. Test egokigarriak informatizatuak izaten dira eta item bakoitzak parametro batzuk eduki behar ditu, ondorioz, kalibraketa egitea ezinbestekoa da.

Artikulu honetan test baten sorrera eta bi kalibraketa-mota azaltzen dira, adituetan oinarrituriko kalibraketa eta kalibraketa psikometrikoa, bukatzeko, kasu erreal baten esperientzia kontatzen da.

2. **TEST BATEN SORRERA**

Test bat sortzeko behar den lehenengo gauza item edo galdera-sorta bat da. Ataza hau ondo egitea uste baino garrantzitsuagoa eta zailagoa da, horregatik da gomendagarria zenbait aholkuri jarraitzea. Itemaren enuntziatuan *beti, inoiz ez,* edo *normalean* gisako, moduko hitzak saihestu behar dira; ez dira errepikatu behar aukera bakoitzean, enuntziatuan, behin jar daitezkeen hitzak; ez dira erabili behar bi ezezko esaldi berean; ez dira era-

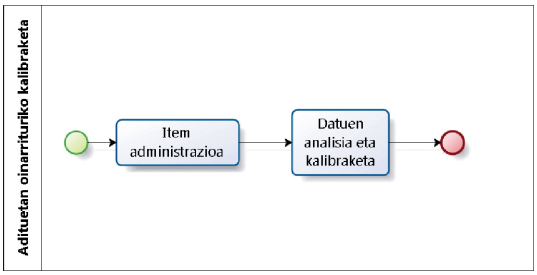
bili behar *aurreko guztiak gaizki daude* edo *aurreko guztiak ondo daude* moduko aukerak, eta luzera berdintsuko aukerak jarri behar dira [4]. Gomendio hauei jarraitzeak ez du esan nahi itema ona izango denik, baina arazoak saihesten lagunduko digu.

Test batean erabiltzeko itemak ditugunean, galdera bakoitzak ezaugarri bereizgarri batzuk eduki beharko ditu; gutxienez, zailtasuna. Ezaugarri hauei balio bat eman behar zaie, eta hori egiteko prozesuari kalibraketa deitzen zaio.

Ohiko testak Testen Teoria Klasikoan (CTT-Classic Test Theory) [5] oinarrituta daude, eta haietan beharrezkoa den parametroa itemaren zailtasuna da. Test Egokigarri Informatizatuak (CAT-Computerized Adaptive Testing), ordea, Itemaren Aurreko Erantzun-Teorian (IRT-Item Response Theory) [6] oinarrituta daude, eta lau parametro izan ditzakete: zailtasuna (zer mailatakoa den itema), diskriminazioa (ikasleek asmatzeak edo huts egiteak noraino ematen digun informazioa), asmatzea (zer probabilitate duen pertsona batek itema jakin gabe, erantzuna asmatzeko) eta hutsegitea (zer probabilitate duen pertsona batek itema jakinda, erantzuna huts egiteko). Den item bakoitzerako zenbat eta balio gehiago nahi izan, orduan eta zailagoa edo luzeagoa izango da kalibraketa-prozesua.

3. ADITUETAN OINARRITURIKO KALIBRAKETA

Kalibraketa-mota honek oinarritzat estatistika konplexua erabili beharrean, adituen esperientzia erabiltzen du, eta Testaren Teoria Klasikoa hartzen du ardatz gisa. Adituetan oinarritutako kalibraketa-prozesu osoa bi zatitan bana daiteke [7], 1. Irudian ikus daitekeen bezala: 1) itemen administrazioa eta 2) datuen analisia eta kalibraketa.



1. irudia. Adituetan oinarrituriko kalibraketa.

Item-administrazioan erabaki behar dira item bakoitzeko zenbat balorazio lortu nahi diren, zenbat adituk hartuko duten parte eta nola egingo diren

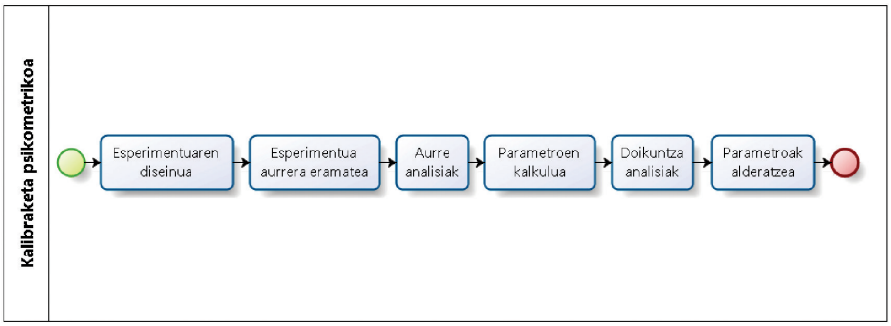
datuen banatzea eta erantzunak jasotzeko prozesua, hau da, internet bidez edo paperean. Erabaki hauek guztiak hartu ondoren, adituekin harremanetan jarri eta galdera-sortak banatu behar zaizkie [8]. Gomendatzen da adituek, galdera bakoitzeko, erantzun zuzena eta zailtasun-maila ematea.

Datuen analisian eta kalibraketan, lehendabizi, datuak iragazi behar dira, adibidez, adituek gaizki erantzundako galderak kenduz edo adituen ustez arazoak dituzten itemak baztertuz. Ondoren, zailtasunaren balioa lortuko da item bakoitzeko, eta bukatzeko, datu estatistikoen bidez analisiak egingo dira, adibidez, fidagarritasun-indizeak kalkulatuko dira, baita batez-bestekoa, desbideratze tipikoa edo bariantza ere. Horrela, itemak bakarka zein multzoka aztertzea lortzen da [9].

4. KALIBRAKETA PSIKOMETRIKOA

Kalibraketa psikometrikoa pertsona askok egingo dituzten azterketetarako pentsatuta dago, Itemaren Aurreko Erantzun-Teorian oinarrituta baitago. Item-banku handi bat edukiz gero, item bakoitza askotan azaltzea saihesten da, baita item gutxi erabiliz, azterketa egitera doazenen maila jakitea eta emaitzak oso zehatzak izatea ere. Baina horretarako, prozesu jakin bati jarraitu behar zaio, eta prozesua zaila ez den arren, luzea eta astuna izan daiteke, hortik, prozesu asko hasi ere ez izana [10].

Prozesu honek sei pauso ditu, 2. irudian ikusten den bezala: 1) esperimentuaren diseinua, 2) esperimentua aurrera eramatea, 3) aurre analisiak, 4) parametroen kalkulua, 5) doikuntza-analisiak eta 6) parametroak alderatzea.



2. irudia. Adituetan oinarrituriko kalibraketa.

Esperimentuaren diseinuan hainbat erabaki hartu behar dira, besteak beste, 1) zein izango diren kalibraketan parte hartuko duten itemak, 2) zen-

bat parametro kalkulatu nahi diren (bat, gutxienez, eta lau, gehienez, normalean hiru izaten diren arren), 3) zenbat pertsona beharko diren kalibraketa aurrera eramateko (200 erantzun item bakoitzeko, parametro bat lortzeko [11], eta 500 gutxienez, 2 parametro edo gehiago lortzeko [12]), 4) galdetegi bakoitzak zenbat item edukiko dituen (item asko badira, galdetegi ezberdinak egin beharko dira eta ainguratze-diseinu bat egin beharko da [13], hau da, item batzuk galdetegi guztietan sartu beharko dira, gero, parametroak alderatu ahal izateko) eta 5) zenbat galdetegi ezberdin egingo diren (kontuan hartu beharko da pertsona batek ezin diela item askori ongi erantzun; adibidez, ohiko itemak badira, hau da, galdera arrunt bat eta aukera motzeko lau erantzun badituzte 60 item inguru gomendatzen da galdetegi bakoitzeko).

Esperimentua aurrera eramateko orduan, proba nola egin erabaki beharko da: paperean, ordenagailu bidez edo modu konbinatuan. Arrazoi ugari daude ordenagailu bat erabiltzeko proba pasatzeko orduan [3], lehen aipatu den bezala.

Aurre analisietan bi alderdi aztertzen dira: alde batetik, administrazioen analisia eta bestetik, item bakoitzaren portaera. Lehenengo kasuan, hauek izango dira begiratu beharreko gauzak: bukatu gabe geratu diren testak, patroiak eduki dituzten probak (adibidez, erantzun guztiak *a* erantzuna badute), proba egiteko denbora gutxiegi edo gehiegi erabili dutenak, galderaren batean denbora gehiegi erabili dutenak, galdera guztiak ondo ala denak gaizki egin dituztenak. Bigarren kasuan, ordea, estatistika-indizeak sakonago aztertzen dira, adibidez, zer maiztasunarekin aukeratu den erantzun bat, item bati askotan erantzun zuzena eman zaion edo erantzun gabe utzi den, edo barneko sendotasun-indizeak diren Cronbach-en alfa [14], Spearman-Brown [15] [16] eta Kunder-Ridcharson (KR20 eta KR21 esaten dietenak) [17].

Parametroen kalkuluan zenbait teknika erabil daitezke, gehienezko sinesgarritasun baldintzatua [18], gehienezko sinesgarritasun bateratua [19] edo gehienezko sinesgarritasun marjinala [20], adibidez. Bestalde, modu errazean aplika daitezke gaur egun teknika hauek, zenbait programa informatikori esker, Xcalibre edo Bilog programei esker, esate baterako.

Doikuntza analisien helburua da ikustea aukeratu den eredua eta lortu diren emaitzak bat datozela. Hori frogatzeko egiten den ikerketa bat unidimentsionalitatearen analisia da. Analisi honen ondorioz, ohikoa da item batzuk baztertu behar izatea.

Ainguratze-diseinua badago, *Parametroak aldaratzea* bakarrik egiten da. Pausu honetan parametroen balio guztiak eskala berean jartzen dira.

5. **KASU ERREAL BATEN ESPERIENTZIA**

2012an, Donostiako IRALEn, ebaluazio-lanetan ibiltzeko talde bat sortu zen. Testu baten ulertze-maila lortuko zuen proba bat egin nahi zen. Taldearen eginkizuna betetzeko, Test Egokigarri Informatizatu (TEI) bat sortzea erabaki zen. Horretarako, 1) item-multzo bat sortu, 2) kalibraketa egin eta 3) testak egiteko programa garatu behar zen.

Item-multzoa sortzean, IRALEko lan-talde batek mota bereko 127 item sortu zituen; lau, AAG (Aukera Anitzeko Galderak) erakoak, zeinetan enuntziatuaren testua luzea zen. 127 item horiek lortzeko, zenbait pauso eman ziren: lehendabizi, 90 item sortu ziren, ondoren, erantzun zitzaten, 100 ikasleri pasa zitzaizkien, gero, zuzentzaileak begiratu zituzten eta, bukatzeko, 9 baztertzea erabaki zen. Item gutxi zirela ikusiz, beste 50 sortu ziren, eta horietatik 4 kendu ostean, 127 itemekin geratu zen item-bankua.

Kalibraketa egitean, [12] erabaki garrantzitsu bat hartu zen: bi kalibraketa-mota erabiliko ziren: adituetan oinarritutako kalibraketa, lehenengo, eta kalibraketa psikometrikoa, ondoren, 3 parametro lortu ahal izateko. Prozesu honetan lehen kalibraketa eduki zen kontuan azpitestak egiterako orduan.

Aditu bidezko kalibraketan 9 adituk hartu zuten parte. Guztiak ziren hizkuntza-irakasleak, baina, halere, bi taldetan sailkatu ziren. Alde batetik, beheko mailetan irakasten ohituak zeudenak eta bestetik, goi mailatan aritzen zirenak. Itemak ere bitan banatu ziren, eta talde bakoitzari zegokion multzoa eman zitzaion. Prozesuari jarraitu ondoren, 127 itemak Europako Erreferentzia Esparru Bateratuak jasotzen dituen 5 hizkuntza-gaitasun mailetan geratu ziren sailkaturik.

Kalibraketa psikometrikoa egiteko lehen aipatutako sei pausuei jarraitu zitzaien. Lehendabizi, esperimentuaren diseinua egin zen, hau da, kalibraketa aurrera eramateko erabakiak hartu ziren. 3 parametro kalkulatuko ziren, beraz, 500 erantzun beharko ziren item bakoitzeko; enuntziatuak luzeak zirenez, 25 item edukiko zituen galdetegi bakoitzak. Ainguratze-diseinu bat behar zen, eta 8 itemez osaturik egongo zen, beraz, 7 galdetegi ezberdin erabiliko ziren. Galdetegi hauek 11 eta 18 urteko ikasleen artean, ordenagailu bidez egitea erabaki zen; ahalik eta profil heterogeneoenetan eta Euskal Herriko toki askotan egingo ziren. Horrela eginez, 3.475 erantzun jaso ziren. Horietatik ez zen bat bera ere baztertu, bai, ordea, itemak: aurre analisiak egin ondoren, 15 item baztertzea erabaki zen. 112 item horien 3 parametroak kalkulatu ziren, doikuntza-analisiak ondo atera ziren eta, bukatzeko, parametroak alderatu ziren.

Programa garatzeko, enpresa bat kontratatu eta Hizeba sistema (irakurketa.hizeba.eu) garatu zen, 112 item kalibratuak bertan txertatuz. Sistema irekia da, edonork egin dezake proba eta, dagoeneko, milatik gora test egin dira.

6. **ONDORIOAK**

Test on bat prestatzea ez da zaila kalibratutako item-banku bat badago, baina, hain zuzen ere, prozesu hori astuna eta luzea suertatzen da, zaila ez izan arren. Ikusi den bezala, pausu batzuei jarraitu ostean, ez da zaila kalibraketa hori lortzea, baina baliabide ugari behar dira horretarako, batez ere, pertsonak; adituak, lehenengo kalibraketaren kasuan, eta testa erantzungo duten pertsonak, bigarrenean.

Azpimarratu beharra dago, egindako ahalegina handia izan arren, kalibraketak merezi izan duela, Hizeba gisako sistema bat garatzea lortu baita eta, aldi berean, euskaraz irakurtzeko gaitasun-maila aztertzen laguntzen duen tresna.

7. **ESKER ONAK**

Gure eskerrik beroena IRALEko Pello Aranbururi eta Antton Peñalbari, haiekin lan egitea oso aberasgarria izan baitzen eta 127 item kalibratzea lortu baitzen.

8. **BIBLIOGRAFIA**

[1] HICKS, M. 1989. *The TOEFL Computerized Placement Test: Adaptive Conventional Measurement. TOEFL Research Report No. 31*. Educational Testing Service, Princeton, New Jersey (USA).

[2] LÓPEZ-CUADRADO, J., PÉREZ, T.A., ARRUABARRENA, R., VADILLO, J.Á. eta GUTIERREZ, J. 2002. «Generation Of Computerized Adaptive Tests In An Adaptive Hypermedia System. » Educational Technology - Information Society And Education: Monitoring A Revolution. **2**, 674-678.

[3] OLEA, J., PONSODA, V. eta PRIETO, G. 1999. *Tests informatizados: fundamentos y aplicaciones. Colección «Psicología»*. Ediciones Pirámide. Madrid.

[4] MUÑIZ, J. 1997. *Introducción a la teoría de respuesta a los ítems*. Ediciones Pirámide. Madrid.

[5] NOVICK, M. R. 1966. «The Axioms and Principal Results of Classical Test Theory.» *Journal of mathematical psychology*. **3**, 1-18.

[6] LORD, F. M. 1952. *A Theory of Test Scores. Psychometric Monograph 7*. Psychometric Corporation, Richmond, VA.

[7] PRESEDO, C., ARMENDARIZ, A., LOPEZ-CUADRADO, J. eta PEREZ, T. 2015. «Sistema de ayuda para la calibración de ítems por el procedimiento basado en el juicio de expertos». *Revista Internacional de Tecnologías de la Educación*. **2**, 1-17.

[8] ARRUABARRENA, R. 2010. *E-learning y la calibración de ítems de test: Teoría de Respuesta al Ítem versus calibración basada en juicios de expertos. Un estudio empírico. (Tesis doctoral)*. Universidad del País Vasco (UPV/ EHU), San Sebastián.
[9] MUÑIZ, J. 2000. *Teoría clásica de los test*. Ediciones Pirámide. España.
[10] CRAIGH, M eta STOCKING, M.L. 1995. *Practical issues in large-scale high-stakes computerized adaptive testing*. Educational Testing. Princeton.
[11] WRIGHT, B.D. eta STONE, M.H. 1979. Best *Test Design. Rasch Measurement*. MESA Press. Chicago.
[12] RENOM, J. eta DOVAL, E. 1999. *Tests adaptativos informatizados: estructura y desarrollo*. Ediciones Piramide. Madrid.
[13] ARMENDARIZ, A. 2014. *Calibración psicométrica guiada por un sistema experto*. Editorial Académica Española. Madrid.
[14] CRONBACH, L.J. 1951. «Coefficient alpha and the internal structure of tests». *Psychometrika*. **16,** 297-334.
[15] SPEARMAN, C. 1910. «Correlation calculated from faulty data». *British Journal of Psychology*. **3,** 271-295.
[16] BROWN, W. 1910. «Some experimental results in the correlation of mental abilities». *British Journal of Psychology*. **3,** 296-322.
[17] KUNDER, G.F. eta RICHARDSON, M.W. 1937. «The theory of the estimation of test reliability». *Psychometrika*. **2,** 151-160.
[18] LORD, F.M. 1980. *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates. New Jersey.
[19] BIRNBAUM, A. 1968. *Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores*. Addison-Wesley: capters 17-20. Reading (USA).
[20] BOCK, R.D eta AITKIN, M. 1981. «Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm». *Psychometrika*. **35,** 179-197.