EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 3, 221–233

KUIDAS HINNATA SUURE PANUSEGA TESTIDE HINDAJAID

Hille Pajupuu

Ülevaade. Suure panusega testide (ingl *high-stakes test*) osatähtsuse kasv toob kaasa vajaduse pöörata rohkem tähelepanu testi subjektiivhinnatavate osade hindamise kvaliteedile. Kasutatavate lihtsate statistiliste meetodite puhul võib jääda märkamata palju valestihindamisi juhul, kui hindajate arv on suur ning nende koolitus nõrk. Artikkel tutvustab meetodit, mille abil saab kindlaks teha valesti hinnanud hindajad ning nende hinnatud tööd õigeaegselt ümber hinnata. Meetod on mõeldud kasutamiseks kahekordse hindamise puhul, seda demonstreeritakse eesti keele algtaseme testi rääkimisosa hindamise näitel.*

Võtmesõnad: hindajate hindamine, hindajate järjekindlus, subjektiivhindamine, kahekordne hindamine, eesti keel

Sissejuhatus

Suure panusega testide all mõeldakse niisugust laadi teste, kontrolltöid, arvestusi, eksameid jms, millel on paratamatult suur mõju inimese saatusele. Seesuguste testide tähtsus on pidevalt kasvanud, nende põhjal tehakse järjest enam otsustusi testi sooritaja, aga ka õpetaja ja õpetamise kohta. Paljud neist testidest on muutunud aja jooksul kohustuslikuks (Altshuler jt 2006, Fine 2005, Helfenbein 2004). Sama tendentsi on näha ka Eestis, mõeldagu siis arvestuslikele töödele, põhikooli lõpueksamitele või gümnaasiumi riigieksamitele. Sama puudutab eesti keele kui teise keele oskuse mõõtmist. Kui 1995. aastal oli Eestis vaid üks riiklikult korraldatav standardiseeritud eesti keele kui riigikeele test – Eesti kodakondsuse taotleja eesti keele eksam –, siis 1999. aastal lisandusid sellele eesti keele tasemeeksamid (alg-, kesk- ja kõrgtase), eesti keele kui teise keele riigieksam ning võõrkeelse põhikooli eesti keele kui teise keele lõpueksam. 2000. aastal ühitati kodakondsuse taotleja eesti keele eksam tasemeeksamitega. (Vt tabel 1.)

* Artikkel on valminud Eesti Keele Instituudi baasfinantseerimise ja Alfred Kordelini Sihtasutuse Eesti Fondi toel. Tänan Riiklikku Eksami- ja Kvalifikatsioonikeskust võimaluse eest kasutada eksamitulemusi uurimistööks.

Tasemeeksami tulemustest sõltub inimese võimalus töötada nii või teistsugust keeleoskustaset eeldavatel ametikohtadel, põhikooli lõpueksamitulemustest sõltub võimalus jätkata õpinguid gümnaasiumis, riigieksamitulemustest pääs ülikooli.

Kõiki neid eksameid korraldab Riiklik Eksami- ja Kvalifikatsioonikeskus (REKK). Tasemeeksamid toimuvad üle riigi seitsme linna eksamipunktides 10 korda aastas, riigieksam ja põhikooli lõpueksam viiakse läbi koolides üks kord aastas. Eksaminandide arv on aasta-aastalt kasvanud. 2005. aastal oli eksaminande 16 124 (algtase 5075, kesktase 2357, kõrgtase 1177, riigieksam 5344, põhikooli lõpueksam 2172).

Tabel 1. Suure panusega eesti keele kui riigikeele eksamid ja nende ligikaudne vastavus Euroopa Nõukogu keeleoskustasemetele

Euroopa Nõukogu keeleoskustasemed		**Eesti keele kui riigikeele eksamid**
Vilunud keelekasutaja (*Proficient User*)	C2	(Eestis riiklikult ei testita ega nõuta)
	C1	Eesti keele kõrgtaseme eksam
Iseseisev keelekasutaja (*Independent User*)	B2	Eesti keele kesktaseme eksam Muukeelse gümnaasiumi riigikeele riigieksam
	B1	Eesti keele algtaseme eksam ühitatud Eesti kodakondsuse taotleja eesti keele eksamiga Eesti keele kui teise keele riigieksam
Algtasemel keelekasutaja (*Basic User*)	A2	(Eestis riiklikult ei testita ega nõuta)
	A1	

Kuna testi võim on suur – mõne tunni jooksul pannakse paika inimese saatus –, siis eeldab see kõigi testiarenduse ja testide kasutamisega seotud inimeste vastutustunnet (vt ka Shohamy 2001).

Suure panusega eesti keele testides annavad testi maksimaalsest punktisummast poole objektiivhinnatavate osade (kuulamine ja lugemine) eest saadavad punktid ja poole subjektiivhinnatavate osade eest (kirjutamine ja rääkimine) saadavad punktid. Seega on hindaja roll lõpliku punktisumma kujunemisel väga suur. Eksameid korraldav institutsioon peab tagama, et hindajad hindaksid võimalikult õigesti. Selleks tuleb hindajaid koolitada ning nende tööd analüüsida. Hindamisvead tuleb leida enne, kui hinded jõuavad eksaminandide tunnistustele.

Eesti keele testimisolukord on suhteliselt keeruline: et eksamid toimuvad üle riigi üheaegselt ja tööde hindamisperiood on väga lühike (30 päeva), tuleb hindamisse kaasata suur hulk hindajaid. Tasemeeksamihindajaid on 65, koolieksamihindajaid üle 400. Kui tasemeeksamihindajad on läbi aastate enam-vähem ühed ja samad, avaliku konkursiga valitud eesti filoloogid, siis koolieksamitel kasutatakse hindajatena kooliõpetajaid, kes võivad aastati vahetuda.

Tasemeeksami hindajaid koolitatakse 1–2 korda aastas, enamik neist hindab kindla taseme eksameid 7–8 korda aastas. Koolitustel osalemine on vabatahtlik, hindajaeksameid pole. Koolieksamite hindajaid koolitatakse üks kord aastas – vahetult enne eksamit – ning enamikul puudub suure panusega eksamite hindamiskogemus.

Suurema objektiivsuse saavutamiseks toimub riigi- ja tasemeeksamitel kahekordne hindamine: sooritust hindab sõltumatult kaks hindajat, hinnete 3-punktise vahe korral (10-punktisel skaalal) läheb töö kolmandale hindamisele.

Probleemid, mis Eestis töös hindajatega välja paistavad on hindajakoolituse vähesus, analüüsi puudumine koolituse mõju hindamiseks, puudulik tagasiside hindajatele eksamihindamise tulemustest, teadmatus hindaja järjekindlusest (ingl *rater consistency*)[1]. Põhjus, miks hindamisele ja hindajatele on seni suhteliselt vähe tähelepanu pööratud, on ühelt poolt eksami sisu ja hindamise eest vastutavate inimeste väike arv (tasemeeksamite arendamise, hindajakoolituse ja analüüsiga tegeleb REKK-is 3 inimest, eesti keele kui teise keele koolieksamitega 1 inimene), kuid teisalt pole selline olukord ainuomane ainult Eesti eksamikeskusele. Vähest tähelepanu hindamisele on Eesti kõrval täheldatud ka teistes Balti riikides, kusjuures kõigis neis kasvab aastast aastasse suure panusega testide arv; just hindamise analüüsi puudumisest võib saada eksamikvaliteedile peamine oht (vt Eckes jt 2005). Ent hindajakoolituse ja hindajaanalüüsi kohta eksamikeskustes pole ka laiemalt üksikasjalikku teavet. Eksamikeskused ei publitseeri hindajaanalüüside raporteid kuigi sageli (Brooks 2004). Küsitluste põhjal võib järeldada, et igakülgsest hindajakoolitusest ja analüüsist jääb puudu paljudes eksamikeskustes (Alderson jt 1995).

Et koolituse ja tagasiside mõju hindamisele on alles viimasel ajal uurima hakatud, siis on võimalik, et teadmisi pole jõutud veel praktikasse rakendada, eriti juhul, kui vastavaid uurimusi tehakse väljaspool eksamikeskusi. Eestis näiteks puudub teadusasutus, kelle uurimisvaldkondade hulka kuuluks keeletestimine. Keeletestimist uuritakse lühiajaliste projektide raames, ning kuna ka REKK-i enda põhiülesannete hulka teadustöö ei kuulu, siis uurimistulemuste praktikasse viimine on küllaltki pikaajaline protsess, mis takerdub sageli ka rahastamise taha.

Hindajakoolituse ja -analüüsi tähtsust ei tohi aga alahinnata: need peaksid käima käsikäes. Eksaminandi hinne ei tohi sõltuda sellest, kas teda satub hindama koolitatud hindaja või mitte. Valesti hinnatud sooritus vajab ümberhindamist enne, kui selle punktisumma jõuab tunnistusele.

Hindajakoolituse mõju

Hindaja usaldatavust tõstab järjekindel koolitamine.

Tom Lumley ja Tim F. McNamara (1995) uurimus näitab, et koolituse mõju ei kesta tingimata väga kaua – aja möödudes hakkavad mõned hindajad hindama ebajärjekindlamalt. Analüüsides hindaja tööd, võib tema koolitusvajadust õigeaegselt märgata. Sara Cushing Weigle (1998), uurides hindajate rangust (ingl *rater severity*) ja järjekindlust enne ning pärast koolitust, täheldas koolitusjärgses hindamises mitmeid muutusi, kuid peaasjalikult uute, vähese kogemusega hindajate puhul. Hindajad jäid küll ka pärast koolitust hindama erineva rangusega, ent äärmuslik

[1] Hindaja järjekindluse all mõeldakse seda, et ta hindab töid alati ühtedelt alustelt lähtudes.

rangus ja leebus taandusid. Koolitus mõjus tõhusamalt hindaja järjekindlusele: osa hindajad, kes enne koolitust klassifitseerusid kui ebajärjekindlad, muutusid küllaltki järjekindlaks pärast koolitust. Kuid koolitus ei muutnud järjekindlaks sugugi mitte kõiki hindajaid: võib-olla pole kõiki hindajaid võimalik üldse järjekindlaks koolitadagi ja nende kasutamisest hindajana tuleb loobuda. S. Cushing Weigle (1998) uurimus kinnitas varasemaid Mary D. Lunzi, Benjamin D. Wrighti ja John M. Linacrei (1990) tulemusi, et hindajakoolitus ei muuda hindajaid üksteise duplikaatideks, kuid koolitus võib muuta hindajad järjekindlamaks. Järjekindlate hindajate pandud hinded on eeldatavalt õigemad. Seega on koolitus tingimata vajalik uutele hindajatele nii järjekindluse saavutamiseks kui ka äärmuslikult range või leebe hindamise vältimiseks; ühtlasi aitab koolitusega kaasnev hindajaanalüüs õigeaegselt välja selgitada need hindajad, kes vaatamata koolitusele jäävad ebajärjekindlaks. Ka William J. Bonk ja Gary J. Ockey (2003) rõhutavad koolituse tähtsust, aga ka kogemuse ja tagasiside vajadust hindaja järjekindlamaks muutmisel, olgugi et isegi väga intensiivne koolitus ei kaota hindajate erinevusi täiel määral.

Hindajate püsivat ranguserinevust koolitusele vaatamata on märkinud ka Kimi Kondo-Brown (2002): sarnase hariduse ja hindajakogemusega hästi koolitatud hindajad hindavad ühe ja sama sooritusе eri aspekte erineva rangusega, ja kuigi see hindajatevaheline erinevus on väike, on ta piisavalt oluline, et mõjutada lõpphinde kujunemist.

Koolitust vajavad ka kogenud hindajad, et tõlgendada hindamisskaalat sarnasemal viisil (Lumley 2002).

Seega on koolitus oluline õiglasema hindamise saavutamiseks, kuid kuna koolitus ei pane hindajaid hindama ühesuguselt, on pidev hindajaanalüüs ja asjakohane tagasiside hindajale vajalik.

Ülesanne

Eesmärk on leida meetod, mis lubab tuvastada valesti hinnanud hindajad võimalikult lihtsal ja kergesti rakendataval viisil ja mida sobiks kasutada eksamikeskustes, kus eksameid ja hindajaid on palju, kasutatakse kahekordset hindamist ning kõikide hindajate järjekindel koolitamine on mingitel põhjustel raskendatud (hindajad paiknevad üle riigi või isegi üle maailma laiali, hindajaid vajatakse ajutiselt ja nende hulgas on palju uusi hindajaid, eksamid toimuvad tihedalt ja nende vahel ei jää koolituseks aega, hinnata tuleb palju töid lühikese ajavahemiku jooksul, hindajakoolitajatest on puudus, hindajakoolituseks pole raha vms).

Hindajate hindamine ja probleemid

Subjektiivhindamisel on oluliseks hindaja järjekindlus, mille tähtsam näitaja on reliaablus. Kui hindajaid on üks, siis peetakse teda reliaabseks juhul, kui ta annab ühe ja sama sooritusе eest eri olukordades samad punktid (ingl *intra-rater relaibility*). Mitme hindaja puhul peetakse hindajaid reliaabseteks, kui eri hindajate ühele ja samale sooritusele antud hindepunktid langevad kokku (ingl *intera-rater reliability*) (vt ka Bachman 1990: 178–181, Luoma 2004: 179–184).

Hindajakoolitusega püütakse saavutada olukord, kus hindaja pandud hinded langeksid kokku teda koolitava eksperthindaja omadega (Alderson jt 1995: 105–112). Enamasti alustatakse koolitust hindaja reliaabluse kindlakstegemisest: arvutatakse hindaja korrelatsioon iseendaga (ingl *intra-rater correlation*). Selleks lastakse hindajal hinnata pikemate vaheaegade järel samu töid uuesti ning korreleeritakse pandud hinded. Tugeva positiivse korrelatsiooni puhul (korrelatsioonikoefitsient üle 0,8) usutakse, et hindajal on kujunenud välja arusaam, kes on tugev ja kes nõrk keeleoskaja, ning ta hindab sama tööd alati ühesugustelt alustelt lähtudes. Tugev positiivne korrelatsioon ei välista siiski seda, et hindaja hindab töid erinevatel kordadel erineva rangusastmega. Et rangusastet välja selgitada, võidakse arvutada iga korreleeritava hindamiskorra hinnete keskmine ja vaadata, kus see hindamisskaalal paikneb.

Lisaks hindaja korrelatsioonile iseendaga, korreleeritakse koolitustel hindaja hindeid eksperthindaja omadega (ingl *inter-rater correlation*) ning arvutatakse mõlema hinnete keskmine ning standardhälve. Eksperthindajaga ühesugust hindamist näitab 0,8-st suurem korrelatsioonikoefitsient. Keskmise järgi näeb, kas hindaja on eksperthindajast rangem või leebem. Kui keskmine on märgatavalt madalam, siis on tegu range hindajaga, kui kõrgem, siis leebemaga (Alderson jt 1995: 132).

Standardhälbe järgi näeb, kui suures ulatuses hindaja hindamisskaalat kasutab, kas ta kasutab seda sarnaselt eksperthindajaga, kas ta julgeb vajadusel panna kõrgeid ja madalaid hindeid. Standardhälbe põhjal võib eksamisituatsioonis hindaja kohta teha järeldusi siiski vaid juhul, kui hinnatavate tööde hulk on suur. Kui hinnatavaid töid on vähe, võib väikse standardhälbe põhjus olla hinnatavate juhuslikult ühesuguses tasemes, mitte hindaja oskamatuses või kartuses hindamisskaalat kasutada.

Kas eksamil kasutada ühe- või kahekordset hindamist, on olnud mõneti problemaatiline. Kahekordne hindamine on kallis, sest selleks vajatakse rohkem hindajaid, ning see on ka aeganõudvam (Brooks 2004). Kahekordne hindamine on õigustatud subjektiivhinnatavate osade puhul, kus sedasi pandud hinne on eeldatavasti usaldusväärsem. Eksameid korraldavale institutsioonile annab kahekordne hindamine võimaluse kontrollida hindamisprotsessi (hindamise kokkulangevuse ulatuse analüüs) ja aidata kaasa hindamise standardiseerimisele. Kahekordne hindamine annab tagasisidet ka hindajatele: hindajate arutlusi pandud hinnete üle võib käsitada ka kui hindajakoolitust (Cannings jt 2005).

Eksamikeskustes, kus subjektiivhinnatavaid testiosi hindab sõltumatult kaks hindajat, kontrollitakse hindamise reliaablust valdavalt hindajatevahelise korrelatsiooniga või arvutatakse kahe hindaja pandud hinnete vahed. Nõrga korrelatsiooni või hinnete suure vahe puhul hinnatakse tööd ümber kas kolmanda hindaja poolt või siis püüavad kaks paarishindajat jõuda kompromissini. Mida koolitatumad ja järjekindlamad on hindajad, seda usaldatavamad on taolised lihtsad statistilised näitajad.

Kui aga ollakse olukorras, kus hindajakoolitus pole väga heal järjel ja kus hindaja järjekindluse kohta pole andmeid, ei saa välistada olukorda, et paaris hindama satuvad kaks kvalifitseerimata hindajat. Sellisel juhul pole eelpoolnimetatud lihtsatest statistilistest näitajatest kuigi palju abi: korrelatsioon võib olla tugev ka siis, kui hindajad hindavad ühtemoodi valesti, nii nagu ei tule märkimisväärset

hinnetevahet siis, kui paaris on hinnanud kaks leebet või kaks ranget hindajat. Seega, kuigi statistilised näitajad võivad olla head, võib eksaminandi siiski olla hinnatud ebaõigesti.

Olukorras, kus hindaja pädevusest pole selget ülevaadet, ent on vaja kindlaks teha ebakvaliteetselt hinnanud hindajad, oleks otstarbekas leida võimalus hindajate võrdlemiseks eksperthindaja(te)ga.

Hindaja sarnasus eksperthindajaga – hindaja kvaliteediindeks

Hindaja (H) töö kvaliteedi hindamise meetodi väljatöötamisel lähtuti eeldusest, et eksperthindajad (E) on hindajad, kes on korralikult koolitatud, kes on järjekindlad (korrelatsioon iseendaga üle 0,8), kes kasutavad hindamisskaalat õigesti (piisava arvu hinnatud tööde korral hinnete keskmine u 6,0 10-punktilisel skaalal, standardhälve võimalikult suur), kes saavad ühtemoodi aru, kes on tugev, kes nõrk keeleoskaja (hindajatevaheline korrelatsioon kahekordsel hindamisel üle 0,8, kel omavahel paaris hinnates keskmised hinded sarnased ning hinnete vahe väiksem kui 3 p 10-punktilisel skaalal).

Analoogiliselt hindajakoolituse eesmärgiga – õpetada hindaja hindama eksperthindajaga sarnaselt, püstitati ka hindaja kvaliteedi määramisel eesmärk leida hindaja sarnasus eksperthindajaga, ja seda ka juhtudel kui hindaja eksperthindajaga ise paaris ei hinda. Hindajakoolituste põhjal võib eeldada, et mida sarnasemalt on hindaja hinnanud eksperthindajaga, seda parem on tema töö kvaliteet ja mida erinevamalt on ta hinnanud eksperthindajast, seda enam on põhjust kahelda tema töö kvaliteedis. Viimasel juhul on hindaja eksperthindajast kas leebem, rangem või ebajärjekindlam.

Meetodi kirjeldus

Hindaja sarnasuse leidmiseks eksperthindajaga arvutatakse kaks indeksit:

- paari sarnasusindeks S (ingl *similarity index*) – näitab kahe paarishindaja sarnasust;
- hindaja kvaliteediindeks K (ingl *quality index*) – näitab hindaja sarnasust eksperthindajaga.

Esimene samm hindaja kvaliteediindeksi leidmiseks on arvutada reaalselt paaris hinnanud hindajate omavahelised sarnasusindeksid. Selleks arvutatakse kõigepealt hinneteerinevuse ruutude keskmine (1) ning seejärel sarnasusindeks (2):

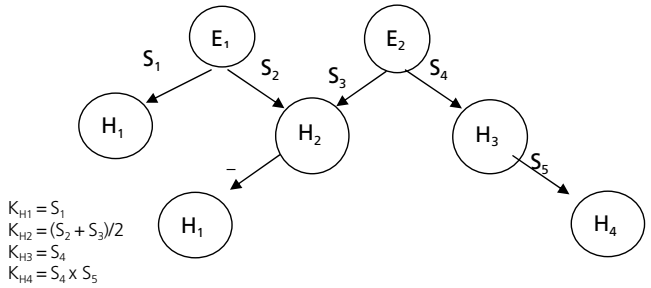
(1) $\overline{H}_{\Delta} = (H_1 - H_2)^2$

(2) $S = 1 - \frac{\overline{H}_{\Delta}}{H_{\max}^2}$,

kus $\overline{H}_{\Delta}$ hinneteerinevuse ruutude keskmine, H_{max} maksimaalne võimalik hinne, S on paari sarnasusindeks. Sarnasusindeks jääb vahemikku 0–1, kus 1 näitab paariliste täpselt sarnast hindamist ja 0 täiesti erinevat hindamist.

Eksperthindaja kvaliteediindeksiks võetakse 1. Ülejäänud hindajatele arvutatakse kvaliteediindeks (K).

Kui hindaja on hinnanud paaris eksperthindaja(te)ga, siis tema kvaliteediindeks on eksperthindaja(te)ga paaris hindamisel saadud sarnasusindeksite keskmine (vt joonis 1, hindajad H_1, H_2, H_3).



Joonis 1. Hindaja kvaliteediindeksi tuletuspuu

Hindajatele, kel pole olnud võimalust hinnata paaris eksperthindaja(te)ga, arvutatakse tuletuslik kvaliteediindeks pärast seda, kui eksperthindajaga paaris hinnanud hindajad on kvaliteediindeksid saanud. Tuletuslik kvaliteediindeks arvutatakse eksperthindajaga sarnaselt hinnanud hindajate kaudu sarnasusindeksite kaalutud keskmisena. Kaalukoefitsiendina kasutatakse eksperthindajaga sarnaselt hinnanud hindaja kvaliteediindeksit (vt joonis 1, hindaja H_4).

Kvaliteediindeksi arvutamisel rakendatakse kahte piirangut:

1. Sarnasusindeksit ei arvutata, kui paaris hinnatud töid on kaks või vähem.
2. Hindaja tuletusliku kvaliteediindeksi arvutamisel võetakse arvesse ainult tema need paarilised, kelle kvaliteediindeks on suurem etteantud koefitsiendist R. R-i väärtus valitakse selline, et see oleks maksimaalne kvaliteediindeks, mille korral saab arvutada võimalikult paljudele hindajatele tuletusliku kvaliteediindeksi. Kui hindajal R-ist suurema kvaliteediindeksiga paarilist pole olnud, jäetakse talle kvaliteediindeks arvutamata.

Kvaliteediindeksite arvutamist jätkatakse iteratiivselt, kuni on nõudeid täitvaid paare.

Katseuurimuse materjal

Meetodit katsetati ja kontrolliti Riikliku Eksami- ja Kvalifikatsioonikeskuses eesti keele algtaseme kolme eksami suulise osa hindamise andmetel. 2005. a märtsi, aprilli ja mai algtaseme eksamil hinnati kokku 1567 suulist sooritust, hindajaid oli 54. Kasutati kahekordset hindamist. Kuna hindajate paaridesse jagamist nendel eksamitel ette ei planeerita, siis on ka hindaja hinnatud tööde ning tema paariliste arv väga erinev, kõikudes 1 hinnatud tööst kuni 209 tööni ja 1 paarilisest 19 paariliseni (vt tabel 2).

REKK kasutab hinnete usaldusväärsuse kontrollimiseks vahede arvutamise meetodit: ümber hinnatakse tööd, mille hindeerinevused kahekordsel hindamisel on 3 või enam punkti 10-punktisel hindamisskaalal.

Tabel 2. Hindaja kood, temaga paaris hinnanud hindajate arv ja hinnatud tööde arv. Poolpaksu kirjaga on tähistatud eksperthindajad

Hindaja kood / paariliste arv / hinnatud tööde arv				
001 / 6 / 92	022 / 1 / 1	038 / 1 / 1	063 / 2 / 52	100 / 3 / 51
002 / 8 / 119	023 / 3 / 40	043 / 1 / 10	064 / 1 / 19	101 / 2 / 29
004 / 10 / 139	**024 / 19 / 139**	045 / 1 / 15	065 / 1 / 15	104 / 4 / 68
005 / 3 / 42	027 / 2 / 33	046 / 3 / 56	067 / 3 / 45	107 / 1 / 16
007 / 1 / 16	028 / 1 / 14	051 / 2 / 24	069 / 1 / 17	109 / 1 / 16
008 / 7 / 84	030 / 8 / 139	054 / 3 / 43	070 / 1 / 16	110 / 3 / 53
009 / 7 / 127	031 / 1 / 15	055 / 3 / 46	074 / 1 / 16	122 / 2 / 27
012 / 10 / 157	033 / 1 / 14	057 / 3 / 49	081 / 3 / 52	123 / 1 / 7
013 / 3 / 41	034 / 8 / 130	059 / 14 / 209	086 / 9 / 152	124 / 1 / 13
014 / 10 / 144	035 / 3 / 50	061 / 2 / 33	087 / 7 / 136	125 / 2 / 34
017 / 1 / 17	037 / 1 / 23	062 / 3 / 48	**091 / 10 / 138**	129 / 3 / 50

REKK-i algtaseme eksami peaspetsialisti soovitusel käsitati eksperthindajatena hindajaid koodiga 014, 024, 091 (tabelis 2 poolpaksus kirjas). Eksperthindajate järjekindlust pole uuritud, nende koolitus on sama mis teistel hindajatel, ent neil on suurem kogemus ning peaspetsialisti usaldus seniste hindamiste põhjal.

Soovitatud eksperthindajatel lasti sõltumatult hinnata 16 tööd ning arvutati pandud hinnete korrelatsioonid esmase hinnangu saamiseks hindamise kokkulangemise ulatuse kohta (vt tabel 3).

Tabel 3. Eksperthindajate hinnete vahelised korrelatsioonid 16 ühesuguse töö hindamise põhjal

Ekspert	E 014	E 091
E 014	1	
E 091	0,839	1
E 024	0,913	0,844

Eksperthindajate kvaliteediindeksiks võeti 1 ja arvutati ülejäänud 51 hindaja kvaliteediindeksid. Tuletusliku kvaliteediindeksi arvutamisel kasutati koefitsiendi R väärtust 0,97.

Tulemused ja järeldused

Hindajate kvaliteediindeksid on esitatud kahanevas järjestuses tabelis 4. 9 hindajale polnud võimalik kvaliteediindeksit arvutada, sest nad olid hinnanud 2 või vähem tööd või ei olnud hinnanud paaris ühegi eksperthindajaga või eksperthindajasarnase hindajaga.

Mida väiksem on hindaja kvaliteediindeks, seda erinevamalt ta eksperthindajast hindab ja seda enam on alust kahelda pandud hinnete usaldusväärsuses. Alates pingerea viimasest, tuleks alustada tööde uuesti hindamist, ning seda mitte juhusliku hindaja poolt, vaid eksperthindaja või temaga sarnaselt hinnanud hindajate poolt (s.t tabeli esimesed hindavad üle viimaseid).

Tabel 4. Hindajate kvaliteediindeksid

Hindaja kood	Kvaliteedi-indeks	Hindaja kood	Kvaliteedi-indeks	Hindaja kood	Kvaliteedi-indeks
014	1	031	0,976	057	0,951
024	1	122	0,976	086	0,948
091	1	104	0,973	002	0,948
030	0,995	035	0.973	109	0,940
087	0,992	129	0,972	107	0,936
059	0,991	001	0,971	027	0,928
023	0,991	005	0,971	061	0,923
043	0,991	008	0,970	123	0,891
012	0,990	017	0,969	022	–
067	0,990	054	0,967	028	–
100	0,989	081	0,967	037	–
013	0,989	051	0,965	038	–
110	0,988	004	0,963	064	–
062	0,988	034	0,963	065	–
046	0,987	101	0,958	069	–
124	0,986	009	0,958	070	–
055	0,984	063	0,957	074	–
007	0,982	045	0,955		
125	0,980	033	0,955		

Meetodi kontrollimiseks võeti eksamirühm, kus sooritust olid hinnanud H_{061} ja H_{001}. Kvaliteediindeksilt paiknes H_{061} pingerea lõpuosas ja H_{001} pingerea keskel (vt tabel 4). Hindajate eksamil pandud hinded ja hinnete vahed on esitatud tabelis 5. Suure hinnetevahe tõttu suunas REKK kolmandale hindamisele 6 tööd.

Eksperthindajal E_{091} lasti hinnata samu töid (vt tabel 5).

Eksperthindaja hinded erinevad H_{061} hinnetest oluliselt 5 juhul (vt tabelis 5 H_{061}–E_{091}), neist 2 ei tulnud REKK-is kasutatava hinnete vahede arvutamise meetodiga välja ja läksid ebaõigetena eksaminandi tunnistusele: üks eksaminand sai põhjendamatult madala hinde ja teine põhjendamatult kõrge hinde (vt tabelis 5 eksaminandid 2 ja 6).

Pingerea keskel oleva H_{001} hinnetest erinesid eksperthindaja hinded oluliselt 1 juhul (vt tabelis 5 H_{001}–E_{091}), mis läks ka eksaminandi tunnistusele ebaõigena (vt tabelis 5 eksaminand 2).

Tabel 5. Eksamil hinnanud hindajapaari hinded ja hinnete vahed; eksperthindaja hinded ning tema ja hindajate hinnetevahed

Eksaminand	H_{061}	H_{001}	H_{061}–H_{001}	Kommentaar	E_{091}	H_{061}–E_{091}	H_{001}–E_{091}	Kommentaar
1	9	7	2		8	1	–1	
2	4	4	0		7	–3	–3	tunnistusele läks ebaõige hinne
3	10	7	3	hinnati uuesti	6	4	1	
4	9	6	3	hinnati uuesti	6	3	0	
5	7	7	0		7	0	0	
6	10	8	2		7	3	1	tunnistusele läks ebaõige hinne
7	5	4	1		5	0	–1	
8	5	5	0		4	1	1	
9	7	4	3	hinnati uuesti	5	2	–1	
10	7	7	0		6	1	1	
11	6	3	3	hinnati uuesti	5	1	–2	
12	9	6	3	hinnati uuesti	6	3	0	
13	7	6	1		8	–1	–2	
14	10	6	4	hinnati uuesti	8	2	–2	
15	9	7	2		8	1	–1	
16	9	7	2		8	1	–1	
$\bar{x}$	7,7	5,9			6,5			
SD	2,0	1,5			1,3			

Tabel 6. Hindajate ja eksperthindaja pandud hinnete vaheline korrelatsioon

	H_{061}	H_{001}	E_{091}
H_{061}	1		
H_{001}	0,735	1	
E_{091}	0,505	0,592	1

Näitest on näha, et hindajapaari hindevahe arvutamine pole piisav, et üles leida ebaõigesti hinnatud töid. Hindevahede arvutamisel läheb kolmandale hindamisele töid, mis kolmandat hindamist ei vajaks, samas jääb osa ebaõigesti hinnatud töid uuesti hindamata.

Tabelis 6 esitatud hindajate ja eksperthindaja hinnetevaheline korrelatsioon näitab hindajapaari omavahelist suuremat kooskõla kui eksperthindajaga. See viitab asjaolule, et pelgalt korrelatsioonile tuginedes ei saa otsustada hindamise õigsuse üle.

Hindaja kvaliteediindeksi arvutamine aitab mõlema laialt kasutatud meetodi (korrelatsiooni ja hinnete vahe arvutamise) puudusi vältida.

Diskussioon

Hindaja kvaliteediindeksi leidmise meetod on kasutatav vaid kahekordse hindamise korral. Kahekordne hindamine pole väga levinud, kuigi tõstab hindamise usaldusväärsust (Cannings jt 2005, Brooks 2004). Suure panusega eksamitel, millest sõltub inimese tulevik, on aga iga võte hindamise usaldusväärsuse tõstmiseks vajalik ning õigustab sellega kaasnevaid kulutusi.

Kuid kulutused pole ainus põhjus, miks kahekordset hindamist omaks ei võeta. Üheks põhjuseks on peetud kahekordse hindamise puhul harjumuspärasest ühekordsest hindamisest erinevat filosoofiat: kui ühekordsel hindamisel teostab (pistelist) kontrolli eksperthindaja, kelle pandud hindeid loetakse usaldusväärsemateks nii tema suurema hindamiskogemuse kui ka selle tõttu, et paljud neist on ise osalenud hindamisskaalade väljatöötamisel, siis kahekordsel hindamisel, kus "õige hinne" kujuneb kahe hindaja pandud hinnete keskmisest või siis hinneteerinevuse korral paariliste või rohkemate hindajate kompromissist, on eksperthindaja roll taandunud. See pole aga hindamismetodoloogias kergelt omaksvõetav muutus (Brooks 2004).

Kvaliteediindeksi leidmise meetod toetub ühekordse hindamise filosoofiale ja hindajakoolituse põhimõtetele, mis peaks hindajatele olema mõistetav ja vastuvõetav. Ent ka see meetod eeldab tööde ümberhindamisel mõtteviisi muutust: meetod ei tegele ebaõigete üksikhinnete leidmisega (nagu kahe hindaja pandud hinnete vahede arvutamine), vaid hindajaga ja tema töö kvaliteediga. Kuid ka siin võib näha analoogiat hindajakoolitusega: ka hindajakoolituses pole olulised üksikud eksperthindajast erinevalt hinnatud tööd, vaid üritatakse selgust saada hindaja järjekindlusest ning rangusest. Kvaliteediindeksi järgi leitakse valesti hinnanud hindajad, ent valestihindamise põhjus selgub alles siis, kui eksperthindaja või temaga sarnane hindaja hakkab töid uuesti hindama.

Uuesti hindamine on otstarbekas siduda koolitusega: ümberhinnatavate tööde hindajat tuleb ebaõnnestunud hindamisest teavitada ja koos temaga arutleda võimalike põhjuste üle.

Meetod tõstab oluliselt eksperthindaja rolli. See tähendab aga eksperthindaja kavakindlat koolitamist, tema järjekindluse uurimist ja ka atesteerimist. Olukord, kus eksperthindaja töö kvaliteet pole eksamikeskuses teada ja tugineb vaid oletusel ja usaldusel või ammusel analüüsil, on lubamatu. Isegi järjekindlad ja kogenud hindajad muutuvad ajapikku ebajärjekindlamaks (Lumley, McNamara 1995, Lumley 2002).

Mida rohkem suudetakse eksperthindajaid välja koolitada, seda enamate hindajatega saab neid paaris hindama panna ja seda enam saab infot teiste hindajate kohta. Võrreldes kõikide hindajate koolitamisega on eksperthindajate koolitamine igati otstarbekam: üheks eksamisessiooniks palgatud hindajat ei saa mõnepäevase koolitusega muuta järjekindlaks hindajaks, küll aga on võimalik välja koolitada inimesi, kes pidevalt on hindamise ja eksamitega seotud. Koolitatud ja atesteeritud eksperthindaja on võimeline ise hindajaid koolitama, mis omakorda toob kaasa hindajate töö kvaliteedi tõusu.

Meetod eeldab hoolikat hindajapaaride planeerimist. Võimalikult paljud hindajad peaksid saama hinnata paaris eksperthindajaga, hindajale hindamiseks antavate tööde hulk peaks olema piisavalt suur kvaliteediindeksi arvutamiseks,

hindajapaarid ei tohiks olla fikseeritud, eksperthindajaid ei peaks omavahel paaris hindama (eriti siis, kui eksperthindajaid on vähe). Kui paaride valik on juhuslik ja eksperthindajatega paaris hinnanud hindajaid vähe, siis ei ole võimalik kõigile hindajatele kvaliteediindeksit arvutada.

Kirjandus

Alderson, Charles J.; Clapham, Caroline; Wall, Dianne 1995. Language Test Construction and Evaluation. Cambridge: Cambridge University Press.

Altshuler, Sandra J.; Schautz, Tresa 2006. No Hispanic students left behind: The consequences of "High-Stakes" testing. – Children & Schools 28 (1), 5–14.

Bachman, Lyle F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bonk, William J.; Ockey, Gary J. 2003. A many-facet Rasch analysis of the second language group oral discussion task. – Language Testing 20 (1), 89–110.

Brooks, Val 2004. Double marking revisited. – British Journal of Educational Studies 52 (1), 29–46.

Cunnings, Rebecca; Hawthorne, Kamila; Hood, Kerenza; Houston, Helen 2005. Putting double marking to the test: A framework to assess if it is worth the trouble. – Medical Education 39, 299–308.

Cushing Weigle, Sara 1998. Using FACETS to model rater training effects. – Language Testing 15 (2), 263–87.

Eckes, Thomas; Ellis, Melanie; Kalnberzina, Vita; Pižorn, Karmen; Springer, Claude; Szollás, Krisztina; Tsagari, Constance 2005. Progress and problems in reforming public language examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. – Language Testing 22 (3), 355–377.

Fine, Michelle 2005. High Stakes Testing and Lost Opportunities. The New York State Regents Exams. – Encounter 18 (2), 24–29.

Helfenbein, Robert 2004. New Times, New Stakes: Moments of Transit, Accountability, and Classroom Practice. – Review of Education, Pedagogy & Cultural Studies 26 (2/3), 91–109.

Kondo-Brown, Kimi 2002. A FACETS analysis of rater bias in measuring Japanese second language writing performance. – Language Testing 19 (1), 3–31.

Lumley, Tom 2002. Assessment criteria in a large-scale test: What do they realy mean to the rater? – Language Testing 19 (3), 246–76.

Lumley, Tom; McNamara, Tim F. 1995. Rater characteristics and rater bias: Implications for training. – Language Testing 12, 54–71

Lunz, Mary E.; Wright, Benjamin D.; Linacre, John M. 1990. Measuring the impact of judge severity on examination scores. – Applied Measurment in Education 3, 331–45.

Luoma, Sari 2004. Assessing Speaking. Cambridge: Cambridge University Press.

Shohamy, Elana 2001. Democratic assessment as an alternative. – Language Testing 18 (4), 373–91.

Hille Pajupuu (Eesti Keele Instituut) uurimisvaldkondadeks on kõneakustika, keeletestimine, kultuuridevaheline kommunikatsioon.
hille.pajupuu@eki.ee

HOW TO ASSESS THE RATERS OF HIGH-STAKES TESTS

Hille Pajupuu
Institute of the Estonian Language

In a situation where, among all tests, the proportion of high-stakes tests is constantly growing, while their results are increasingly used to pass judgements not only on examinees but also on teachers and teaching quality, and, with time, many of those tests have become obligatory, high demands should certainly be set on the sense of responsibility of anyone involved in test development or test use.

Special attention should be paid to the quality of subjective ratings as the writing and speaking parts of a test may often account for half of the total score. If a testee should score lower or higher than their competence is worth, it may change their life as well as that of other people.

The commonly used simple statistics (calculation of differences between the marks awarded by two raters, inter-rater correlation) may actually fail to take account of quite a lot of wrong credits if the raters are many and inadequately prepared.

In order to reduce unfair assessment a method is suggested to identify poorly performing raters and to reassess their results in good time. The method is meant to be used in the case of double marking. Notably, a quality index is computed to show the degree of similarity between the credits given by the rater to be asessed and an expert rater, even if the two have never worked in a pair. It is assumed that the higher the similarity the fairer the credits.

The article describes the general principles of the method suggested, pointing out its advantages over some other simple methods used for the same purpose.

Keywords: rater evaluation, rater consistency, rating, double marking, Estonian