Language Policy Division
Division des Politiques linguistiques

Jaanuar 2009

Keeleeksamite sidumine Euroopa keeleõppe raamdokumendiga: õppimine, õpetamine ja hindamine

Lisamaterjal standardite järjepidevuse tagamiseks eri keeltes ja kontekstides ning eri testikasutuskordadel, lähtudes õpetajate hindamisotsustest ja üksikvastuste teoorial põhinevatest skaaladest

Brian North (Eurocentres/EAQUALS)
Neil Jones (Cambridge Assessment / ALTE)

Integratsiooni ja Migratsiooni Sihtasutus
Meie Inimesed

INNOVE

Keelepoliitika osakond, Strasbourg
www.coe.int/lang

Lisamaterjali tõlkimist korraldas SA Innove keelekeskus Euroopa keeleoskusuuringu (European Survey on Language Competences, ESLC) raames, uuringut rahastas Integratsiooni ja Migratsiooni Sihtasutus *Meie Inimesed* Euroopa Sotsiaalfondi programmist "Keeleõppe arendamine 2011—2013".

Tõlkija: Kristel Weidebaum (Luisa tõlkebüroo)

Keeletoimetaja: Ethel Roosna (SA Innove)

Euroopa keeleoskusuuringu 2011 koordinaator Eestis: Kristi Mere (SA Innove)

Tallinn 2012

Sisukord

1. Sissejuhatus 1

2. Mõõtmisskaala koostamine ja tõlgendamine 2

2.1. Eristuskirja koostamine 4

2.2. Eeltestimine 4

2.3. Andmekogumine ja skaala koostamine 4

2.4. Skaala tõlgendamine 5

2.4.1. Tasemepiiride tõlgendamine 6

3. Testiväliste kriteeriumite kaasamine 7

3.1. Raamdokumendi ankurküsimused 7

3.2. Õpetajate üldhinnang 8

3.2.1. Hindamisvahendid 9

3.2.2. Õpetajate hindamisotsuste täpsus 9

3.2.2.1. Keeleoskustasemete tõlgendamine 10

3.2.2.2. „Nähtamatute" oskuste hindamine 10

3.2.2.3. Leebus ja rangus 11

3.2.2.4. Kriteerium- ja normhindamine 11

3.2.3. Keeleoskustasemete piiride määratlemine 11

3.3. Kirjelduskriteeriumid kui IRT kohased testiküsimused 12

3.3.1. Hindamisskaala 13

3.3.2. Õpetajapoolne hindamine 14

3.3.3. Enesehindamine 14

3.3.4. Keeleoskustasemete piiride määratlemine 15

4. Raamdokumendi kirjelduskriteeriumite skaala vahetu rakendamine 16

4.1. Standardimine FACETSi meetodil 17

5. Järjestuse kasutamine keeleülesel tasemepiiride määratlemisel 18

6. Kokkuvõte 20

Kasutatud kirjandus 21

1. Sissejuhatus

Standardite kehtestamisel (tasemepiiride määratlemisel) on mõistagi tähtis nende pikaajaline stabiilsus ning järjepidevus testi väljatöötamise ja korraldamise tsükli vältel. Tähtis on püüda tasemepiiride määratlemine ja nende järjepidevuse tagamine sellise tsükliga siduda. Just sellele käesolevas dokumendis keskendutaksegi.

Peale selle pakutakse siin võimalusi, kuidas kasutada tasemepiiride määratlemiseks õpetajate hindamisotsuseid ning mil viisil rakendada eri keelte vahelise seose loomiseks raamdokumendil põhinevaid kirjeldus- ja/või hindamiskriteeriume. Käsitletakse küsimust, kuidas saab reaalsete tasemepiiride määratlemiseks kasutada õpetajate hindamisotsuseid ja/või enesehindamist. Ent põhirõhk pannakse siiski sellele, et tasemepiiride määratlemist tuleb näha tasemeteskaala, eri keelte, testide väljatöötamise ja korraldamise tsüklite ning nende korduva läbiviimise kontekstis. Kõik need aspektid on seotud skaleerimisega.

On selge, et sidudes tasemepiiride määratlemise testi väljatöötamise ja korraldamise tsükliga, peab rõhuasetus aja jooksul nihkuma tasemepiiride määratlemiselt nende järjepidevuse tagamisele. See ei tähenda, et tasemepiire võib määratleda vaid üks kord ja lõplikult. Uue eksami väljatöötamisel on väga tõenäoline, et esimesed tasemepiirid on ajutised ja muutuvad. Tavaliselt kujunevad tasemepiirid välja pärast korduvat järjestikust täpsustamist. Isegi kui tasemepiirid tunduvad usaldusväärsed, on neid siiski vaja pikema aja jooksul kontrollida. Rõhk on sellegipoolest standardi täiustamisel, mitte iga kord uute tasemepiiride määratlemisel. Sellel on kaks eeldust:

- esiteks peavad meetodid olema võrreldavad. Uut testi ja uusi testitavaid võrreldakse eelmiste testide ja testitavatega. Nii võib inimliku hinnangu kese erineda sellest, mida kohaldati tasemepiiride määratlemisel;

- teiseks saaks ja tuleks rohkem panustada meetoditesse, mis võimaldavad standardeid täiustada: küsimuste andmebaasi loomisse ja skaleerimisse.

Seda, et skaleerimisest on saanud tasemepiiride määratlemisel üha tähtsam abivahend, on selgesti rõhutatud käsiraamatu 6. peatükis. Tasemepiiride määratlemise meetodeid on tutvustatud enam-vähem kronoloogilises järjestuses, et lugejal oleks parem eri mõistete kasutuselevõttu jälgida. Käsiraamatu kasutajad märkasid kindlasti teatud reeglipärasust:

a) uuemate käsitlusviiside, s.t järjestusmeetodi, korvimeetodi ja terviktööde meetodi puhul vaadeldakse tasemepiire/standardeid asjaomase keeleoskustasemete skaala kontekstis: öeldakse, et tase on pigem B1 kui A2 või B2, selle asemel et öelda „arvestatud/ mittearvestatud";

b) normiks on kujunenud eeltesti käigus saadud andmete esitamine töörühmale. Tavaliselt hõlmab see teave andmeid testiküsimuste empiirilisest raskusest (teise vooru jaoks) ja arvutuste tegemist selle kohta, kuidas mõjutavad esialgsed otsused positiivse tulemuse saavutavate testitavate osakaalu (kolmanda vooru jaoks);

c) üksikvastuste teooriat (ing *Item Response Theory*, edaspidi *IRT*) kasutatakse sageli selleks, et paigutada erinevatel (eel)testidel kasutatud küsimused ühele mõõtmisskaalale, ja ka selleks, et tasemepiire oleks andmebaasi testiküsimuste skaala põhjal võimalik kindlaks määrata lõplikult, mitte iga uue testiversiooni jaoks uuesti;

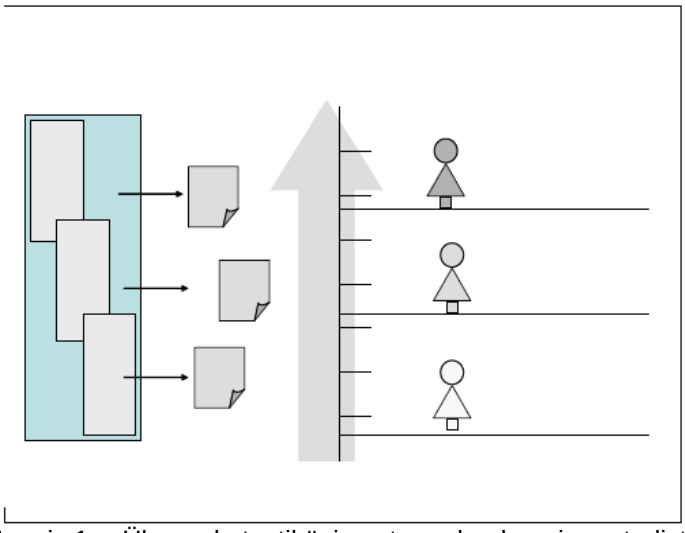
d) kõige uuem kirjeldatud meetod – järjestusmeetodi Hollandi riikliku eksamikeskuse CITO variant – mitte ainult ei hõlma kõiki kolme ülalnimetatud aspekti, vaid näeb ka ette, et

töörühma liikmed hindavad peale üksikute testiküsimuste raskuse veel *tasemepiire* IRT kohasel mõõtmisskaalal olevate tasemete vahel. Seega on selles ühendatud nii töörühma- kui ka skaalapõhine vaatenurk.

Selles dokumendis järgitakse käsiraamatu 6. peatüki käsitlusviisi ning keskendutakse skaalade koostamisele ja kasutamisele testi väljatöötamise ja korraldamise tsüklis. Küsimuste andmebaasist saab tasemepiiride määratlemise keskne teema.

2. Mõõtmisskaala koostamine ja tõlgendamine

Et ruum on piiratud, pole IRT-d võimalik käesolevas dokumendis põhjalikult tutvustada. Väga hea ja lihtsa ülevaate on andnud Baker (1997) ja detailsemalt on IRT mudeleid kirjeldatud käsiraamatu abimaterjalide osas G. Siin anname põgusa ülevaate peamistest omadustest, mis iseloomustavad IRT-l põhinevat testiküsimuste andmebaasi meetodit (vt joonist 1).



Joonis 1. Ülevaade testiküsimuste andmebaasi meetodist

Selle meetodi puhul koostatakse üks mõõtmisskaala, millele saab paigutada testiküsimused raskuse järgi ja keeleõppijad nende võimekuse järgi, samuti saab sellele paigutada soorituse kriteeriumtasemed. Skaala koostatakse võimalikult paljude testiküsimuste põhjal. Väga oluline on see, et küsimused *kalibreeritakse* (prognoositava raskuse alusel) ühele skaalale. Selleks luuakse kõikide kalibreerimisel kasutatud vastuste vahel seosed. Seoseid on võimalik luua mitmel viisil:

- uut kalibreeritavat materjali sisaldavates testides on ka testiküsimuste andmebaasis leiduvaid ja juba kalibreeritud küsimusi (nn ankurküsimusi);
- testikomplektis on kattuvaid küsimusi – nii on teste võimalik koos kalibreerida;
- ühel ja samal õpirühmal lastakse teha vähemalt kaks eri testi.

Kui skaala on koostatud, saab teste kokku panna sihtrühmale sobiva raskusega küsimustest. See tagab keeleoskuse suhteliselt tulemusliku mõõtmise. Õppijate testiskooride põhjal on võimalik välja arvutada, kuhu nad võimekuse skaalal paigutuvad. Kuigi õppija skoor erineb olenevalt testi raskusest, on tema asukoht skaalal tema võimete absoluutne näitaja, millel on ka konkreetne tähendus. See tähendus tuleneb muidugi üksnes testiküsimuste andmebaasis olevatest küsimustest. „Võimekus" ja „raskus" on teineteist määratlevad mõisted, mis kerkivad esile siis, kui õppijad puutuvad kokku testiülesannetega. Seda, mida skaala mõõdab, väljendavad just nimelt testiküsimused ja nende koht raskusjärjestuses. Tehes kindlaks need omadused, mis näivad muutvat testiküsimused raskemaks või kergemaks, saame selgust, mida õigupoolest mõõdetakse, ja nii on meil võimalik hinnata ka testiküsimuste andmebaasi valiidsust.

Testiküsimuste andmebaasi meetod hõlbustab märkimisväärselt tasemepiiride määratlemist ja nende järjepidevuse tagamist. Skaalal olevad punktid võivad alguses olla vaid numbrid, kuid sarnaselt termomeetrile kantud numbritega on ka nende tähendus püsiv. Seega võime aja jooksul kujundada ühise arusaama sellest, milliseid järeldusi need teha võimaldavad ja milliseid samme nad toetavad. Testiküsimuste andmebaasile mitte tugineva testi sooritusе tõlgendamine on üksikakt – aja jooksul ei saa kujuneda ühist selget arusaama ning ebakindlus, mis paratamatult tasemepiiride hinnangupõhise määratlemisega kaasneb, ei kao kunagi. Asjaolu, et standardit on võimalik rakendada järjepidevalt, aitab testiküsimuste andmebaasi meetodi korral standardi mõtet järjest paremini mõista.

Testiküsimuste andmebaasi koostamine võib seega tasemepiiride määratlemisele kaasa aidata mitmel moel:

- Eri keeleoskustasemete testiversioone ning ühe ja sama keeleoskustaseme testi paralleel-versioone saab siduda andmekogumise meetodiga, kus puuduvad andmed kaetakse.

- Testiküsimuste andmebaasi laiendamisel võib aluseks võtta algse kalibreeritud testiküsimuste komplekti ning uued tulemused kanda samale skaalale. Et testiküsimuste andmebaasi jaoks kehtib vaid üks skaala, tuleb raamdokumendi tasemepiirid kasutatava(te) testiskaala(de) puhul kindlaks määrata vaid ühel korral.

- Objektiivtestidest, õpetajate hindamisotsustest ja enesehindamisest saadud andmeid on võimalik analüüsida koos; tänu sellele on lihtsam:
  - kaasata tasemepiiride määratlemise aluseks olevate raamdokumendi keeleoskustasemete tõlgendamisse palju rohkem keelespetsialiste;
  - ankurdada kirjelduskriteeriumite abil eri keelte tasemepiire samale skaalale.

- Olemasoleva skaala – näiteks skaala, millele kantakse andmed erinevatelt järjestikustel keeleoskustasemetel olevatelt eksamitelt – saab raamdokumendiga siduda ilma, et skaalale kantavad keeleoskustasemed või hinded vastaksid täpselt raamdokumendi keeleoskustasemetele. Testi sooritуspiiriks võib õigupoolest olla kirjeldus „tasemete A2+ ja B1 vahel; pigem lähemal tasemele B1, kuid mitte päris B1". Andmeid aluseks võttev skaalapõhine käsitlusviis võimaldab seda fakti täpselt hinnata ja siis arusaadavalt väljendada (nt A2++). Tänu sellele püsib nii kohaliku standardi kui ka raamdokumendi keeleoskustasemete terviklikkus, nagu nähti ette 2007. aasta keelepoliitika foorumil (Euroopa Nõukogu 2007: 14). Tuleb meeles pidada käsiraamatu 1. peatükis märgitut ja raamdokumendi sissejuhatuses rõhutatut: raamdokument on arutlemist ja suhtlemist soodustav metakeel; see ei ole ühtlustamisprojekt, kus kirjutatakse inimestele ette, millised peaksid olema nende eesmärgid (Euroopa Nõukogu 2001: xi „Raamdokumendi kasutajale").

Testiküsimuste andmebaasi koostamise konkreetne viis sõltub suuresti lähtekohast, projekti eesmärgist ja teostatavusest. Testiküsimuste andmebaasi saab kasutada nii mitut keeleoskustaset hõlmava testikomplekti puhul kui ka ühe kindla keeleoskustaseme testi puhul. Esimene neist variantidest on keerulisem, sest keeleoskustasemeti on vaja luua korralikud seosed (*vertikaalsed* seosed). Kui testiküsimusi saab hiljem uuesti kasutada, on selline andmebaasi koostamine eriti kasulik ja võimaldab ka paremat kalibreerimist. Alustada võib nullist või juba olemas olevast testimaterjalist, keskenduda võib ühele või mitmele osaoskusele ning eesmärgiks võib seada raamistiku moodustamise isegi mitme keele jaoks. Õppekavadest võivad tuleneda teatud piirangud. Eeltestide korraldamine või sihtotstarbeline andmekogumine võib mõnel juhul olla hõlpsam, mõnel aga keerulisem. Erinev võib olla ka reaalselt läbi viidud eksamite tulemuste kättesaadavus. Samuti võib erineval keeleoskustasemel olla tehniline tugi, näiteks testiküsimuste analüüsimisel või andmebaasi küsimuste haldamiseks vajaliku süsteemi ülesehitamisel. Kõikidest nendest teguritest võib sõltuda otsus mingi lähenemisviisi kasuks ja testiküsimuste andmebaasi koostamise üldine teostatavus. Testiküsimuste andmebaasi meetodi valimine võib tingida ka vajaduse muuta testide väljatöötamise seniseid tavasid.

Kuigi konkreetsete projektide ülesehituses on erinevusi, on raamdokumendiga seotud testiküsimuste skaala väljatöötamisel siiski neli põhietappi:

- eristuskirja koostamine;
- eeltestimine;
- andmekogumine ja skaala koostamine;
- skaala tõlgendamine.

2.1. Eristuskirja koostamine

Olenemata sellest, kas tegeldakse olemasoleva testiga või töötatakse välja uut, on testi seost raamdokumendiga vaja tõendada.

Esimene samm on seega koostada väga üksikasjalik raamdokumendiga seotud eristuskiri asjaomas(t)e oskus(t)e kohta igal keeleoskustasemel. Selle juures võib abi olla tutvumisest käsiraamatu 4. peatükiga, kus kirjeldatakse eristuskirja koostamist, ja sealsamas käsitletud sisuanalüüsi tabelitega.

Uue testimaterjali väljatöötamiseks tuleb moodustada töörühm. Raamdokumendi keeleoskustasemete üksmeelse tõlgenduse leidmiseks peab töörühm läbima raamdokumendiga tutvumise ja standardimise koolituse, nagu on kirjeldatud käsiraamatu 3. ja 5. peatükis. Seejärel lastakse inimestel omaette või väiksemates rühmades koostada teste või testiküsimusi nende konkreetsete keeleoskustasemete jaoks, millega neil on vajalikud kogemused olemas.

2.2. Eeltestimine

Teiseks tuleks testiküsimuste komplekte väikese, ent esindusliku ja sobiva(te)l keeleoskustaseme(te)l olevate testitavate valimiga eeltestida. Praktikas võib küsimuste andmebaasi koostamine pelgalt uurimisprojekti raames, väljaspool eksamikorraldustsüklit, olla keeruline. Kui tavapärases testi koostamise tsüklis eeltestimist ei toimunud, on küsimuste andmebaasi meetodit juba olemas oleva eksami kontekstis keeruline rakendada. Seega võib olla vajalik see etapp lisada.

Kui eeltestimine juba toimub, tuleb loomulikult kaaluda paljusid aspekte, nagu käsiraamatu osas 7.2.3 on öeldud. Kõige ilmsem neist on testiküsimuste kordumatuse kaitse – kas eeltestimises osalevad õppijad peavad päris eksamil tõenäoliselt vastama samadele testiküsimustele.

2.3. Andmekogumine ja skaala koostamine

Järgmisena tuleb testiküsimused esialgseks kalibreerimiseks erinevatesse omavahel seotud testidesse jagada. Kõige praktilisem on ehk lisada ankurtest, mis on kõigile ühe keeleoskustaseme testiversioonidele ühine. Seostamiseks on ka keerukamaid viise, kuid need võivad raskendada andmekogumise korraldust.

Eriti hoolikas tuleb olla seoste loomisel eri keeleoskustasemete vahel, sest seda on testikorralduse tsüklisse keeruline lõimida ja see eeldab erilist organiseerimist. Õige sihtrühma valimine on raske ning seega on vigade vältimiseks parem valida suhteliselt lai võimete skaala. Hiljem tuleb olla valmis jätma välja vastused, mis ei ole eesmärgistatud (on kas väga tugevad või väga nõrgad). Vertikaalne seostamine toimib üldjuhul paremini siis, kui seda tehakse vahetult pärast iga keeleoskustaseme küsimuste kalibreerimist, kasutades selleks küsimusi, mis on ükshaaval välja valitud raskuse ja statistilise toimivuse (sobivuse) pärast.

Mitut keeleoskustaset hõlmavat küsimusekomplekti on teoreetiliselt võimalik kalibreerida üheainsa analüüsi käigus, kui see hõlmab kõiki vastuseandmeid. Aga ehkki see on võimalik, tuleks tegutseda

ettevaatlikult ja kontrollida andmeid korduvalt, et usaldusväärseima tulemuse saamiseks need puhastada. Kui otsustatakse selle käsitlusviisi kasuks, on kõige kindlam kasutada CITO meetodit: ankurdada iga test 50% ulatuses üks keeleoskustase kõrgema ja 50% ulatuses üks keeleoskustase madalama testiga, nii et iga testiküsimus on ankurküsimus, välja arvatud 50% küsimustest kõrgeima ja madalaima keeleoskustaseme testiversioonis.

Praktikas tuleb testiküsimuste andmebaasi loomiseks vajalikku analüüsi ja kalibreerimist korrata ning see kestab kaua. Kui sellest saab aga lahutamatu etapp testikorraldustsüklist, muutub protsess muidugi pidevaks.

Ükskõik, kuidas kalibreerimiseks andmeid kogutakse, kvaliteedi kontrollimise põhinõuded jäävad samaks:

- testitavate valim peaks üldkogumit piisavalt hästi esindama;
- valim peaks olema küllaldase suurusega (näiteks Raschi mudelit kasutades 100);
- vältida tuleks vastuseid, mis ei ole eesmärgistatud, s.t väga suuri ja väga väikseid skoore; sellised vastused tuleks kalibreerimisandmetest välja jätta;
- vältida tuleks mõju, mis põhjustab ette teada olevalt kallutatust, näiteks ankurtesti ja reaalse testi tegemisel nõutava pingutuse erinevat määra või ajalist survet, mille tõttu näivad testi lõpus olevad küsimused raskemad;
- kui selline mõju siiski ilmneb, tuleb püüda see kindlaks teha ja kalibreerimisandmetest kõrvaldada;
- ankurküsimuste valimisel tuleb olla hoolikas.

Heade ankurküsimuste sobivus ja eristusjõud on *keskmine* ning need ei toimi peamistes huvirühmades erinevalt.

2.4. Skaala tõlgendamine

Selles punktis käsitletakse eeskätt küsimust, kuidas tõlgendada eri keeleoskustasemeid (nt A2–C1) hõlmavat skaalat, ent see on asjakohane ka ühe keeleoskustaseme (nt B1-tase, mis tähistab praktikas vahemikku A2+ kuni B1+) skaalade puhul. Tõlgendamist käsitatakse protsessina, mis viiakse lõpule ühe kokkusaamise käigus, ehkki nagu eespool märgitud, võib skaala koostamine ja tõlgendamine väga vabalt venida pikemaks ning osutuda ka korduvaks toiminguks.

Skaala tõlgendamisel on esimene ülesanne teha kindlaks, millise osa eri keeleoskustasemetele mõeldud testiküsimused skaalal enda alla võtavad, ilma et kattuks järgmisel keeleoskustasemel olevate küsimustega, ning samal ajal võrrelda testiküsimusi hoolikalt eristuskirjadega. Tuleb otsustada, kuhu tuleks kattuvatel aladel tõmmata esialgsed tasemepiirid. Selleks on kolm võimalust ja ideaalis tuleks otsused teha korduvas protsessis, võttes arvesse neid kõiki. Kui testiküsimused on hästi koostatud ja eesmärgistatud, peaksid need kolm võtet moodustama terviku:

- kaalutletud otsuse tegemiseks sobitage kõik kattuvatele aladele jäävad küsimused üksikasjalike eristuskirjade ning raamdokumendi kirjelduskriteeriumitega eelmise ja järgmise keeleoskustaseme kohta;

- paigutage tasemepiir alguses eri keeleoskustasemetele määratud testiküsimuste kattuva osa peal täpselt keskele, nagu soovitati järjestusmeetodi ja selle CITO variandi käsitluses käsiraamatu osades 6.8 ja 6.9;

- määrake tasemepiir kõigil kattuvatel aladel selliselt, et igale keeleoskustasemele logitiskaalal kuuluv osa on oma proportsioonilt loogiline. See tähendab, et keeleoskustasemete laius ei tohiks olla juhuslik.

Selle käigus on kasulik analüüsida kõrvalekalduvaid tulemusi andnud testiküsimusi (konkreetse keeleoskustaseme jaoks kavandatud testiküsimusi, mis praktikas paigutusid mõnele teisele keeleoskustasemele) ja otsida selle põhjuseid. Mõnel juhul võib põhjuseks olla teksti ja testiküsimus(t)e kokkusobimatus, teistel juhtudel võib aga ülesanne tunduda mõistlik ja hästi kalibreeritud, ent nõuda väiksemat või suuremat pingutust mingis kitsamas osaoskuses, mis eelmisel või järgmisel keeleoskustasemel on eesmärgina põhjendatud. Selliste küsimuste korral on vaja otsustada, kas jätta need alles või mitte. Üldjuhul tuleb kahtlevalt suhtuda kõikidesse küsimustesse, mille kavakohane ja empiiriline raskus erinevad suurel määral.

Otsuse võib koostaja või analüüsija teha ka üksinda, kuid mida rohkem inimesi selles osaleb, seda parem. Võimalik, et väikse töörühma korral oleksid tulemused esinduslikumad. Igal juhul tuleks kogu protsess täpselt dokumenteerida ja selle kohta aruanne koostada.

Selles punktis kujutatud toimingud sarnanevad küsimuse ja kirjelduskriteeriumi sobitamise meetodi (käsiraamatu osa 6.7) ja järjestusmeetodiga (osa 6.8) ning selle Hollandi riikliku eksamikeskuse variandiga (osa 6.9). Erinevus seisneb selles, kes määrab kindlaks testiküsimuste keeleoskustaseme ja millal seda tehakse. Siin kirjeldatu kohaselt moodustavad testiküsimuste koostajad küsimusi täpselt eristuskirja järgi ning tasemepiirid määratakse selle põhjal kindlaks peaaegu mehaaniliselt. Kui kasutatakse selliseid tasemepiiride määratlemise meetodeid nagu küsimuse ja kirjelduskriteeriumi sobitamise meetod ning järjestusmeetod, teeb hindajarühm midagi samalaadset pärast tasemepiiride määratlemist. Mõlemal juhul on täiendav välisvalideerimine siiski ülimalt soovitatav. Kogemused keeletestimisega, millesse suhtutakse tõsiselt ja mille küsimuste koostamisega tegeldud väga põhjalikult, näitavad, et oskuslikud testiküsimuste koostajad suudavad kokku panna valiidse ja kavandatud keeleoskustasemele hästi vastava testi. Kuid sellest üksi ei piisa, et toetada vertikaalvõrdsustamist või täpset hindamist. Tuleb kasutada IRT mudelit, skaleerimist ja mõnda üldtuntud standardit (nt keeleoskustase).

2.4.1. Tasemepiiride tõlgendamine

Punktis 2.4 kirjeldatu võimaldab testiküsimuste andmebaasil põhinevale skaalale määrata tasemepiiri, mis eraldab võimalikult täpselt kahe järjestikuse keeleoskustaseme kirjeldamiseks mõeldud testiküsimusi. Kuid see tasemepiir ei võimalda veel väita, et testitav on konkreetse keeleoskustaseme saavutanud. Testitav, kes paigutub täpselt mingi testiküsimuse, näiteks B1-taseme esimese ja lihtsaima testiküsimuse keeleoskustasemele, vastab sellele testiküsimusele õigesti 50% tõenäosusega. Mida keerukam on testiküsimus, seda väiksem on õigesti vastamise tõenäosus.

Sellise piiripealse testitava skoor testis, mille küsimused paiknevad skaalal tasemepiiride A2/B1 ja B1/B2 vahel, oleks üsna väike – liiga väike, et näidata selle keeleoskustaseme saavutamist. Nagu käsiraamatu osas 6.8 on selgitatud, peame määrama vastuse tõenäosuse, mis oleks oluliselt suurem kui 50%, näiteks 80%. Seejärel peame nihutama tasemepiiri kõrgemale, et sellele paigutuv testitav vastaks 80% tõenäosusega õigesti kõige lihtsamale B1-taseme küsimusele ja saavutaks selle keeleoskustaseme küsimustest koosnevas testis mõistliku skoori.

Teisisõnu – tasemepiir A2/B1 on ühtlasi A2-taseme ülempiir. B1-taseme piiripealne testitav pole mitte üksnes A2-tasemeni küündinud, vaid valdab seda täielikult. Ta vastab 80% tõenäosusega õigesti ka kõige raskemale A2-taseme küsimusele ja kergemate küsimuste puhul on õigesti vastamise tõenäosus veelgi suurem.

3. Testiväliste kriteeriumite kaasamine

Punktis 2 käsitleti skaalade väljatöötamise tavapäraseid etappe: a) konstrukti määratlemist ja testiküsimuste koostamist; b) konstrukti ja testiküsimuste kontrollimist eeltestimise käigus; c) andmekogumist ja skaala koostamist; d) tulemuse tõlgendamist ja esialgsete tasemepiiride kehtestamist. Et kinnitada sel moel kehtestatud raamdokumendi tasemepiiride õigsust, võiks aga teha veel teatud toimingud. Kõige ilmsem oleks ristvalideerimisuuringu läbiviimine, s.t mõne muu tasemepiiride määratlemise meetodi kasutamine. Selleks sobib mõni käsiraamatu 6. peatükis või alajaotuses 7.5 pealkirja „Tasemepiiride määratlemise väline valideerimine" all käsitletud meetod, näiteks seal kirjeldatud näide sellest, kuidas kasutada testivälise kriteeriumina õpetajate hindamisotsuseid.

Skaalapõhine käsitlusviis võimaldab sellised testivälised kriteeriumid appi võtta juba andmekogumisel. Kui kasutatakse raamdokumendil põhineva olemasoleva testi/skaala küsimusi või testitavakeskset käsitlusviisi ja õpetajate hindamisotsuseid või isegi enesehindamist, saab neid väliseid vaatenurki arvesse võtta juba esialgsete tasemepiiride määratlemisel. Kindlasti ei ole vaja aga tasemepiiride määratlemist ja välist valideerimist või ristvalideerimist ajaliselt eraldada. Nende mõlema sidumine esialgse andmekogumisetapiga on väga soovitatav, sest see aitab sidumisuuringut kogu protsessi vältel õigel kursil hoida ja nii ei teki ohtu, et tulemused on vastuolus traditsioonilise testivälise valideerimisuuringu tulemustega.

Testivälise kriteeriumi saab protsessi üldstruktuuri lõimida vähemalt kolmel allkirjeldatud moel. Hea oleks kasutada rohkem kui üht võimalust, sest mida rohkem teavet tasemepiiride määratlemiseks kaasatakse, seda parem.

- **Raamdokumendi ankurküsimused:** andmekogumiseks mõeldud testiversioonidesse lõimitakse asjaomaste keelte raamdokumendi sooritusnäidised.

- **Õpetajate üldhinnang:** kaasatakse testitavate õpetajad, kes peaksid hindama, millisel raamdokumendi keeleoskustasemel on nende õpilaste vastav(ad) osaoskus(ed), lähtudes koolitusel omandatud teadmistest ja sobivast raamdokumendil põhinevast hindamisvahendist.

- **Kirjelduskriteeriumid testiküsimustena:** õpetajapoolsel hindamisel ja/või enesehindamisel kasutatakse kontroll-lehel eraldi testiküsimustena asjakohaseid üksikuid raamdokumendi kirjelduskriteeriumeid.

Kui õpetajad sel moel projekti kaasatakse, saab osaga neist läbi viia tasemepiiride määratlemie seminarid (kasutades selleks mõnda käsiraamatu 6. peatükis kirjeldatud meetodit). Nii saab veenduda, et andmekogumiseks valitud testiküsimused vastavad tõepoolest hästi nende õpilaste keeleoskustasemete tüüpilisele ulatusele.

3.1. Raamdokumendi ankurküsimused

Raamdokumendi sooritusnäidiste appivõtmine asjakohas(t)e osaoskus(t)e hindamiseks on enesestmõistetav. Seda on võimalik teha kahel moel ja projekti eri etappides. Need kaks meetodit ei välista üksteist ja ideaalses projektis võiks oma koht olla mõlemal:

- **eeltestimine:** selle käigus on raamdokumendi sooritusnäidiste abil võimalik kontrollida, kas oletused kohalike testiküsimuste ja raamdokumendi keeleoskustasemete seose kohta on üldjoontes õiged ning kas kohaliku testi mingi keeleoskustaseme küsimused a) hõlmavad skaalal sooritusnäidistega sarnast logitivahemikku ja b) tõendavad keskväärtuste võrreldavust;

- **andmekogumine:** testiküsimuste andmebaasi jaoks põhiandmeid kogudes võiks raamdokumendi sooritusnäidiseid kasutada järjestikuste keeleoskustasemete teste siduvate ankurküsimustena, sest see aitab kaasa reaalse skaala loomisele. Kui kalibreerimiseks kogutakse testiküsimuste vastuste kohta andmeid päris eksamil, tuleb mõelda, kuidas saaks raamdokumendi ankurküsimusi testi lisada. Üks võimalus on lasta testitavatel päris eksamiga umbes samal ajal läbi teha ka kalibreeritud küsimustest koosnev ankurtest. Sellisel juhul võivad testitavad ankurtesti suhtuda aga vähem tõsiselt kui päris eksamisse ja see tooks tõenäoliselt kaasa nihke.

Raamdokumendi sooritusnäidiste kaasamise kasulikkust on tõestanud Szabo (2007). Ta leidis nimelt, et kuigi uuringu ajal olemas olnud raamdokumendi sooritusnäidised ei olnud kaugeltki täiuslikud, sai neid kasutada n-ö võrdlusküsimustena, mille abil tõendada, et uute testiküsimuste ja sooritusnäidiste skaalal ei esinenud raskuse ulatuses mingeid märkimisväärseid erinevusi.

3.2. Õpetajate üldhinnang

Õpetajatest võib olla palju abi nii testitavaid puudutava lisateabe hankimisel kui ka hindamisotsuste paigutamisel raamdokumendi keeleoskustasemetele, et neid saaks seejärel kasutada tasemepiiride kehtestamiseks mõõtmisskaalal.

Jones, Ashton ja Walker (2007) on Ühendkuningriigi projektile Asset Languages tuginedes toonud näite sellest, kui olulised on õpetajate hindamisotsused tasemepiiride kehtestamisel ja vertikaalse skaala koostamisel. Selles projektis kasutati hindamisotsuseid astmetena raamdokumendil põhineval „keelteredelil". Õpetajatelt paluti hindamisotsuseid kahes järgmises etapis.

- **Eeltestimise** etapis pidid õpetajad andma üldhinnangu, lähtudes Inglismaa riikliku õppekava keeleoskustasemetest, mis vastavad üldjoontes „keelteredeli" astmetele. Riikliku õppekava keeleoskustasemed võeti aluseks seetõttu, et need leiti põhi- ja keskkooliõpetajate jaoks keeleoskustasemetest kõige tuttavamad olevat.

  Eeltestimise etapis said õpetajad anda iga testitava keeleoskustaseme kohta ühe üldise hinnangu. Need hinnangud olid vastavuses enne määratletud skaalale paigutatud võimekusega. Seega saab IRT meetoditega analüüsitud eeltesti andmed ankurdada skaalale testitavate hinnangulise võimekuse kaudu.

- **Andmekogumine –** põhiandmete kogumise etapis paluti õpetajatel anda hinnanguid kuulamis- ja lugemisoskuse kohta, lähtudes „keelteredelist" või raamdokumendist ja kasutades abivahendit, mis põhineb mõlemal skaalal.

  Reaalsete testide koostamisel saab testiküsimuste sel moel ankurdatud raskusi kasutada selleks, et prognoosida testi eri skooridele vastavat võimekust ja seega ka skaalal olevate hinnete künnistele vastavaid skoore.

Enne määratletud skaala on mall, mida kohaldatakse projekti Asset Languages raames vaikimisi kõikide keelte ja kõikide objektiivtestitavate osaoskuste (lugemine, kuulamine) suhtes. See annab ettekujutuse, millised peaksid empiiriliselt koostatud skaala proportsioonid välja nägema. See skaala on seotud Cambridge Assessmenti ESOLi eksamite aluseks oleva ühtse empiiriliselt koostatud skaalaga, mis näitab, et keeleoskustasemed ilmnevad õpipingutuse ja nähtavate õpitulemuste funktsioonina (Jones, Ashton ja Walker 2007). Aja jooksul peaks empiiriline vertikaalsidumine võimaldama kõiki skaalasid täiustada, aga kui mõne keele puhul seda veel toimunud ei ole, aitavad keeleoskustasemeid tõhusalt ankurdada just õpetajate hindamisotsused.

Asset Languages töötab 25 keelega kuuel keeleoskustasemel. Nii paljude osiste koondamine terviklikuks raamistikuks ei ole tavapäraste testikesksete tasemepiiride määratlemise meetoditega

võimalik. Peale selle võib väita, et sisemise sidususe saab kõige paremini tagada siis, kui tervikliku raamistiku tasandist liigutakse selle väiksemate osiste poole, mitte vastupidi, s.t ei alustata üksikuid keeli, keeleoskustasemeid ja testiküsimusi puudutavatest sidumata tasemepiiride määratlemise otsustest.

Lisaks sellele, et õpetajate üldhinnanguid saab kasutada tasemepiiride määratlemisel abivahendina, nagu selles punktis kirjeldatud, võib neid rakendada ka välisvalideerimise uuringus, et saada kinnitust mõne muu meetodiga määratud tasemepiiridele. Näiteks Northi (2000b) järgi on õpetajate üldhinnanguid „keelesüsteemi tundmise" ja kirjutamisoskuse kohta kasutatud selleks, et anda keeleõppe intensiivkursusel riigis, kus asjaomast keelt kõneldakse, eri keelte esialgsele grammatika- ja sõnavaraküsimuste andmebaasile kindel alus. Seda meetodit on kirjeldatud käsiraamatu osas 7.5.3, mis kuulub testivälise valideerimise teema alla.

3.2.1. Hindamisvahendid

Õpetaja üldhinnang võib olla mitmes vormis.

a) **Üldskaala** on nagu ühtne, terviklik hindamisotsus, mis vastab kõige sobivamate raamdokumendi (all)skaalade või lihtsa üldskaala (mis on esitatud käsiraamatu lisa tabelis C1) toel küsimusele, millisel keeleoskustasemel on konkreetne testitav. Seega annab üldskaala vastuse ühele küsimusele.

b) **Analüütiline tabel** kujutab endast paljusid raamdokumendi skaaladele tuginevaid hindamisotsuseid erinevate suhtluse keeletoimingute ja suhtluspädevuse aspektide kohta. Hindamisotsuse aluseks võib valida RD skaalatabelitel põhinevad hindamisvahendid või nendest tuletatud analüütilised tabelid, nagu on näidatud käsiraamatu lisas C (tabelid C2–C4). Analüütiline tabel annab seega vastuse vähemalt viiele-kuuele küsimusele. Tulemuse põhjal saab seejärel leida keskmise, ühe üldise tulemuse, mis vastaks ülalkirjeldatud ühtsele terviklikule hindamisotsusele.

c) **Kontroll-leht,** mis koosneb 30–50 kirjelduskriteeriumist, võib samuti vahel olla hindamisotsuse aluseks. Selliseid kontroll-lehti kasutati Šveitsi uurimisprojektis, mille käigus kalibreeriti kirjelduskriteeriumeid ja töötati välja raamdokumendi keeleoskustasemed. Nüüd on need kõigile kättesaadavad valideeritud Euroopa keelemappide keeleõppeloo osas. Kontroll-lehe puhul saavad vastuse 30–50 küsimust. Üks skooride kasutamise võimalus on anda kõigi heakskiidetud *Can Do*-nendingute alusel üksainus RD-le tuginev hinnang. Milline protsent tõendaks asjaomase keeleoskustaseme saavutamist? Sellele küsimusele ei ole lihtsat vastust. Euroopa keelemappide puhul on künnis tavaliselt 80%. Kui kirjelduskriteeriumite valik on piisavalt esinduslik, võib asjaomase keeleoskustaseme määrata õppijale, kes oskab (*can do*) 67% sellest, mida on kirjelduskriteeriumites nimetatud. Kui ta saavutab aga 85% või 90%, paigutub ta tõenäoliselt juba järgmisele keeleoskustasemele, ehkki ei valda keelt sel keeleoskustasemel veel täielikult. Teisisõnu – järgmise keeleoskustaseme kontroll-lehel võiks ta paljude kirjelduskriteeriumite puhul öelda, et ta oskab.

Nagu järgmises punktis kirjeldatud, võib selliseid andmeid kasutada ka nii, et iga kirjelduskriteeriumit võetakse kui eraldi testiküsimust.

3.2.2. Õpetajate hindamisotsuste täpsus

Ükskõik, millist lähenemisviisi õpetajad hindamisel kasutavad (skaalat, tabelit, kontroll-lehte), on heaks tavaks suhtuda hindamisse kui testiprotsessi, mis peab tõendama teatud valiidsust. Tuleb silmas pidada,

- kas õpetajad saavad keeleoskustasemetest aru;
- kas õpetajad on tõepoolest suutelised kõnealuseid osaoskusi hindama;
- kas standardi kehtestamist mõjutab õpetajate leebus või rangus;
- kas õpetajad lähtuvad tugevamate ja nõrgemate õppijate hindamisel kriteerium- või normhindamisest.

3.2.2.1.Keeleoskustasemete tõlgendamine

Õpetajaid tuleb koolitada võimalikult hästi, ideaalis peaks see toimuma RD-ga tutvumise ja standardimiskoolitusena, nagu on kirjeldatud käsiraamatu 3. ja 5. peatükis. Kõigi projektis osalevate õpetajate korrakindel koolitamine võib aga osutuda keeruliseks, sest koostöö juba niigi hõivatud õpetajate vahel sõltub sageli nende heast tahtest. Pealegi võib õpikodade jaoks sobiva aja leidmine olla keeruline ja nii ei pruugi osal õpetajatel ühel või teisel põhjusel olla võimalik koolitusel osaleda. Sel juhul võib välja pakkuda kaugõppe (nt veebilehel CEFTrain.net olevate näidete uurimise), kuid siis on vaja ka süsteemi, et kontrollida, kas õpetajad tõepoolest koduse töö ära teevad.

Üks võimalus seda probleemi lahendada on märkida üles, kas õpetajad läbisid koolituse või mitte, ja andmetes esinevate probleemide korral jätta võimalus vähendada rühma, nii et alles jäävad usaldusväärsemad vastajad.

3.2.2.2.„Nähtamatute" oskuste hindamine

Teine probleem on see, kas õpetajad on tõepoolest suutelised kõnealuseid osaoskusi täpselt hindama. Meie arusaamas õppija keeleoskustasemest on ilmselgelt kõige tähtsamal kohal just suuline ja kirjalik väljendusoskus. Et mõtteline tegevus nagu lugemine ja kuulamine on vaid kaudselt vaadeldavad, on neid palju raskem täpselt hinnata, isegi kui õpetaja tunneb oma õpilasi hästi.

On tõenäoline, et õpetajate antavaid hinnanguid võivad tugevamini mõjutada just väljendusoskused. Kui palju see õpetajate hinnangute kasutamisel probleeme valmistab, on raske öelda. Kindlasti ei tekitaks see mingeid muresid siis, kui saaksime olla veendunud, et kõigi õppijate sooritus eri osaoskustes on stabiilne, kuid seda tuleb ette harva – enamasti on õpilased ühes osaoskuses tugevamad kui teises. Stabiilse soorituse (mitte teksti vastuvõtu oskuste palju kõrgema keeleoskustaseme) eeldamine oleks kindlasti väga küsitav paljudes oludes, näiteks inglise keele oskuse hindamisel Põhja-Euroopa riikides. Grin (1999/2000) on tõendanud, et tublisti üle 20% Šveitsi saksa keele kõnelejate inglise keele oskusest ei ole õpingute, töö, suhete või reisimise tulemus. See on lihtsalt miski, mis „on olemas" ja arvatavasti on see seotud pigem teksti vastuvõtu kui tekstiloome oskustega.[1]

Seega on õpetajate üldhinnangute puhul võtmeküsimus see, kas keskenduda:

- tekstiloome-, s.t rääkimis- ja kirjutamisoskusele;
- üldisele raamdokumendi keeleoskustasemele;
- muljel põhineva hinnangu andmisele selle kohta, milline on testitava (nähtamatu) võimekus uuritavates osaoskustes (kuulamises, lugemises, keeleteadmistes).

[1] Šveitsi saksa keele kõnelejate prantsuse keele oskus ja Šveitsi prantsuse keele kõnelejate inglise ja saksa keele oskus seevastu on peaaegu täielikult tingitud nendest neljast tegurist, mida Grin oma uuringus mainib.

Raske on ette teada, kas õpetajad suudavad teksti vastuvõtu oskuste kohta teha järjekindlaid otsuseid, mis korreleeruksid piisavalt hästi testiskooridega, et neid saaks kasutada. Võib juhtuda, et nad annavad hoopis üldhinnangu, mida mõjutab rohkem õpilase soorituse väline külg (tekstiloomeoskused), selle asemel et anda tõepärane hinnang teksti vastuvõtu oskuste kohta. Kui paluda neil iga testitavat kirjeldada, on kindel, et õpetajad vähemalt mõtlevad selle teema üle. Testi tulemustega kõige rohkem korreleeruvaid hinnanguid on sellisel juhul võimalik tõepoolest tasemepiiride määratlemisel kasutada.

Isegi kui hinnangut tekstiloomeoskuste kohta ei saa kasutada hinnangu andmiseks teksti vastuvõtu oskuste kohta, tundub olevat mõistlik neid andmeid niisuguse projekti raames koguda. Lõppude lõpuks on sedasorti projektide logistilised probleemid seotud rohkem nende mahuga (testitavate, õpetajate, läbiviidavate testide arv jne) ning osalejatele (testiküsimuste koostajatele, testide väljatöötajatele, hindajatest õpetajatele) vajaliku raamdokumendiga tutvumise ja standardimiskoolituse korraldamisega. Kui paluda igal õpetajal hinnata iga testitava puhul üht või kaht lisaaspekti – anda ülevaade neljast või viiest aspektist –, ei tekita see õpetajatele ega analüüsijatele korralduse mõttes suurt lisakoormust, kuid võib tasemepiiride määratlemisel end ära tasuda.

3.2.2.3.Leebus ja rangus

Hindajate rangus teatavasti erineb. Kuidas saab sel juhul olla kindel, et õpetajate liigne rangus või leebus ei põhjusta soovimatult suurt dispersiooni? Kui õpetajad hindavad üksnes oma õpilasi, ei ole seda võimalik mitte kuidagi kontrollida, ehkki loogiline oleks eeldada, et kõikide õpetajate peale kokku on see mõju keskmine. Kui õpetajatel saab lasta hinnata teatud hulka samu vastusenäidiseid või DVD-näidiseid, võib hindaja ranguse kindlaksmääramiseks lähtuda FACETSi programmis (Linacre 1989; 2008) kasutatud mitmemõõtmelisest Raschi mudelist, ja võtta seda seejärel arvesse testitavate võimekuse üle otsustamisel. See ei taga aga siiski õpetajate üldist täpsust. Ainus võimalus on kasutada analüüsi ankurdamiseks ühe või mitme autoriteetse eksperdi hinnanguid.

3.2.2.4.Kriteerium- ja normhindamine

Kui õpetajatel palutakse hinnata oma klassi õpilasi, kipuvad nad tugevate ja nõrkade erinevusi võimendama. Nad soovivad paremaid õpilasi esile tõsta ja neile võib olla vastumeelt panna kahele õpilasele sama hinne (mis võib aga rangelt kriteeriumeid järgides osutuda vajalikuks), kui nad näevad kahe õpilase saavutuses selget erinevust. Seda tõendasid Šveitsi õpetajad, kes osalesid raamdokumendi kirjelduskriteeriumite kalibreerimise projektis – märkimisväärselt sarnane tendents ilmnes kahe eri aasta erinevate andmekogumite puhul. Lahendusena otsustati igast klassist välja jätta kõige paremad ja kõige nõrgemad õpilased (kes teoreetiliselt peaksid asuma 25. ja 75. protsentiilil) ning lisaks veel kaks õpilast nende õpetajate klassidest, kellel täheldati hinnangute erinevust, mis oli keskmisest kaks standardhälvet suurem (North 2000 a: 215–216).

3.2.3. Keeleoskustasemete piiride määratlemine

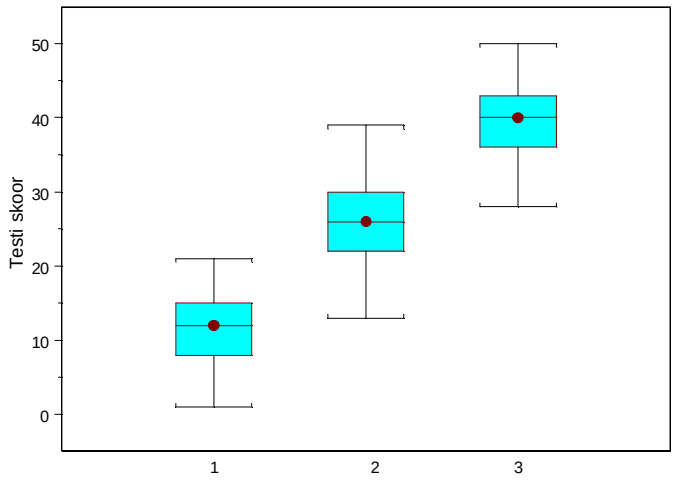
Järgmine küsimus on, kuidas saab selliseid üldhinnanguid testitava keeleoskustaseme kohta (üks tulemus iga testitava kohta, nt B1) siduda testiskooridega või punktidega IRT-l põhineval skaalal. Käsiraamatu osas 6.5 on käsitletud klassikalisi meetodeid, mille järgi seostatakse õpetajate hinnangud toorpunktidega – vastanduvate rühmade meetodit ja piiripealse rühma meetodit. Need meetodid on kõige kasulikumad ühe testi tulemuste sidumisel ühe standardiga, ent neid võib kohandada ka selleks, et siduda testi tulemusi järjestikuste standarditega, näiteks raamdokumendi keeleoskustasemetega.

Andmete graafiline esitus aitab meil hästi uurida, millist teavet need meetodid võivad anda. Järgmises näites kasutatakse testiskooride ja raamdokumendi eri keeleoskustaseme seose näitamiseks karpdiagramme. Joonise 2 aluseks on väljamõeldud andmekogum. See sisaldab 750 testitava skoore, mis on saadud 50 küsimusest koosneva testi põhjal. Iga testitava puhul on raamdokumendi

keeleoskustase saadud õpetaja hinnangust, milles lähtuti kolmest kõrvutiasetsevast keeleoskustasemest. Need keeleoskustasemed on joonise X-teljel tähistatud numbritega 1, 2 ja 3, kuid need võivad vastata ka keeleoskustasemetele A1, B1 ja B2. Y-teljel on testi skoorid.[2]

Joonisel on kujutatud kolme testitavate rühma, kelle õpetajad paigutasid 1., 2. ja 3. tasemele. Seest värvitud karp esindab skooride keskmist 50%. 1. tasemel jäävad 50 küsimusest koosneva testi keskmised skoorid vahemikku 8–15. Üle karbi ulatuv horisontaaljoon (mis läbib karbis olevat punkti) näitab mediaani. Karbi ülemine piir tähistab 75. protsentiili ja alumine 25. protsentiili. Karbist üles- ja allapoole jäävad nn sabad, mis väljendavad keskmisest väga erinevaid skoore. Siinses näites vastab ülemise saba ots suurimale skoorile ja alumise saba ots väikseimale. Vahel märgitakse ära ka 10. ja 90. protsentiil koos erindite täpse asukohaga.

Vaadeldavas näites laseb diagramm teha väga selge järelduse: test eristab kolme keeleoskustaset üsna hästi, kuigi märgata on eri tasemetele paigutatud testitavate skooride olulist kattumist. See ei ole siiski ebatavaline. Selleks et siduda üldhinnangud ja IRT teoorial põhinev skaala, mille aluseks on testiküsimuste andmebaasile tuginev(ad) test(id), on sobivam kasutada käsiraamatu osas 6.6.2 kirjeldatud logistilist regressiooni. Enamikul analüüsimeetoditel on sellise logistilise regressiooni väljendamiseks spetsiaalne funktsioon. Ülalkirjeldatud projektis *Asset Languages* kasutati õpetajate hindamisotsustest õpilaste võimekuse tuletamiseks IRT meetodeid. Nii saadi alus küsimuste kalibreerimiseks ja seejärel leiti enne kindlaks määratud keeleoskusskaala põhjal iga testi hinne.



Joonis 2. Karpdiagramm

3.3. Kirjelduskriteeriumid kui IRT kohased testiküsimused

Kui kasutatakse kontroll-lehtedel olevaid kirjelduskriteeriume, võib igaühte neist suhtuda ka kui eraldiseisvasse testiküsimusse, mida õpetaja või õpilane hakkab ise hindama, selle asemel et teisendada kontroll-lehtede skoorid üldtulemusteks, nagu kirjeldati punkti 3.2.1 alapunktis c. Kirjelduskriteeriumitega kontroll-leht hõlmab niisiis 30–50 testiküsimust, mida analüüsitakse ükshaaval.

RD kirjelduskriteeriumite kasutamisel selliseks andmete kogumiseks on järgmised eelised.

- **Andmepõhine tasemepiiride määratlemine.** Kirjelduskriteeriumite kasutamine võimaldab läbi viia ulatusliku testitavakeskse ja andmepõhise analüüsi. See lubab tasemepiiride määratlemisel aluseks võtta laiapõhise üksmeelse keeleoskustasemete tõlgenduse. Niisugune vaatenurk võib täiendada tasemepiiride määratlemise töörühma oma. Raamdokumendi

[2] Karpdiagramme käsitlev osa on võetud käsiraamatu tööversioonist ja selle autor on Norman Verhelst.

kirjelduskriteeriumeid võib pidada empiiriliselt valiidseks seepärast, et need peegeldavad suure hulga keeleõppega seotud isikute üksmeelset arvamust, mitte ühe komisjoni oma.

- **Keeleülene ankurdamine.** Selle abil saab eri keeled ja piirkonnad ühendada ühtsesse raamistikku isegi nii hästi, et võime olla kindlad nende sarnases tõlgendamises. Projektis *Asset Languages* hindasid kahes eri keeles testi sooritanud õppijad ise oma võimeid mõlemas keeles. See aitas hinnata, kui hästi saab kõnealuse meetodiga kinnitada objektiivtestide keeleülest kooskõla. Idee on huvitav, sest mitmekeelsete keeleõppijate hinnang omaenda võimetele peaks üsna hästi peegeldama nende suhtelist pädevust mõlemas keeles. Seega ei peaks probleeme tekitama ka asjaolu, et endale antud hinnangus pole keeleoskustase tavaliselt sama mis absoluutne keeleoskustase. Uuringus lähtuti meetodist, mida oli varem rakendatud standardite kontrollimiseks BULATSi testi puhul eri keeltes, kasutades iga keele kohta antud enesehinnanguid ALTE *Can Do*-nendingute abil.

- **Tsükliline tasemepiiride määratlemine.** Kui projekti igas etapis (katsetamine, eeltest, andmekogumine) kasutatakse õpetajapoolsel hindamisel ja enesehindamisel kirjelduskriteeriume, võib see innustada standardit kehtestama tsükliliselt, et vaadelda teemat mitme nurga alt (hinnangute või andmete põhjal). See tagab projekti edukuse ja õige suuna.

Puudus võib seisneda asjaolus, et korrelatsiooni ilmnemine iseenesest ei anna alust väita valiidsuse olemasolu. Testiskoori ja hinnangute selline korrelatsioon ei tõenda konstruktivaliidsust. See võib teoreetiliselt olla tingitud eeskätt konstruktide erinevusest (nt vanus). Seega peaks hüpoteetiliselt olema sel moel võimalik „siduda” 8–14aastaste matemaatikatestid raamdokumendi keeleoskustasemetega, kui kasutada suhtluspädevuse asemel konstruktina hoopis vanust. See tähendab, et enne kui asutakse kirjelduskriteeriumite abil tasemepiire määratlema, tuleb käsiraamatu 4. peatükis kirjeldatud meetodeid kasutades koostada täpne eristuskiri. Enesele antud hinnangutest võib aga ka konstrukti valideerimisel palju abi olla. Näiteks Ashton (2008) on kasutanud keeleoskustasemetel A1–B1 olevate keskkooliõpilaste saksa, jaapani ja urdu keele lugemisoskuse hindamiseks enesehindamist segameetodis, kus võrreldi, kuidas õpilased lugemisoskust kui konstrukti igal keeleoskustasemel tajusid. Analüüsi käigus saadi huvitavaid tõendeid sarnasustest ja erinevustest nii eri keelte kui ka keeleoskustasemete lõikes.

3.3.1. Hindamisskaala

Kui kirjelduskriteeriumeid kasutatakse testiküsimustena, saab neile skoori määrata jah-ei-vastuste või hindamisskaala abil (0–2, 0–3 või 0–4). Šveitsi RD uurimisprojektis kasutati näiteks sellist skaalat:

0	**1**	**2**	**3**	**4**
Tase, milleni testitav *ei küündi*	**Jah, soodsas olukorras**	**Jah, normaalses olukorras**	**Jah, isegi raskes olukorras**	**Ilmselgelt kõrgemal tasemel**

Rääkimisoskuse puhul kirjeldati skaalat juhistes järgmiselt.

⓪ Tase, milleni testitav kindlasti *ei küündi*. Testitavalt *ei* saa sellist sooritust eeldada.
① Testitavalt võib sellist sooritust eeldada siis, kui olukord on soodne, näiteks juhul, kui tal on veidi aega mõelda, mida öelda, või kui vestluspartner on leplik ja valmis abistama.
② Testitavalt võib sellist iseseisvat sooritust oodata normaalses olukorras.
③ Testitavalt võib sellist sooritust oodata isegi raskes olukorras, näiteks kui tuleb ette midagi ootamatut või kui vestluspartner on vähem koostööaldis.
④ Sooritus, mis on *selgelt madalam* testitava keeleoskustasemest. Testitav suudab ülesandega veelgi paremini toime tulla.

3.3.2. Õpetajapoolne hindamine

Sellise käsitlusviisi puhul on oluline mitte nõuda õpetajatelt rohkem, kui nad teha suudavad. Tõenäoliselt ei ole realistlik paluda õpetajal anda 30 õpilasega klassis igale õpilasele 50 punktist koosnevas küsimustikus iga *Can Do*-kirjelduskriteeriumi eest neljaastmelisel skaalal hinnang. See annaks vaid puudulikke tulemusi.

Projektis Asset Languages paluti õpetajatel täita küsimustik oma klassi „heade", „keskmiste" ja „nõrkade" õpilaste kohta ning järjestada õpilased selle rühmitusviisi kohaselt. Umbes samamoodi tehti Šveitsi projektis – õpetajad järjestasid kõik kahe klassi õpilased ja valisid seejärel mõlemast klassist viis õpilast: mediaanile paigutuva õpilase, 25. ja 75. protsentiilil asuvad õpilased ning mediaani ja nende protsentiilide vahelisel keskpunktil asuvad õpilased.

3.3.3. Enesehindamine

Enesehindamist kirjelduskriteeriumitega kontroll-lehtede abil on varem kasutatud testide sidumiseks keeleoskusskaaladega (TOEIC/TOEFL: Boldt jt 1992; Wilson 1989, 1999, 2001; ALTE: Jones 2002). Kontroll-lehtede koostamiseks võib kasutada nii raamdokumendi kui ka Euroopa keelemapi kirjelduskriteeriume. Raamdokumendi kirjelduskriteeriumite eelis on see, et need põhinevad valdavalt kirjelduskriteeriumitest testiküsimustel, mis on juba kalibreeritud. Šveitsi projekti raames kalibreeritud kirjelduskriteeriumite skaala on esitatud Northi (2000 a) teose lisas.

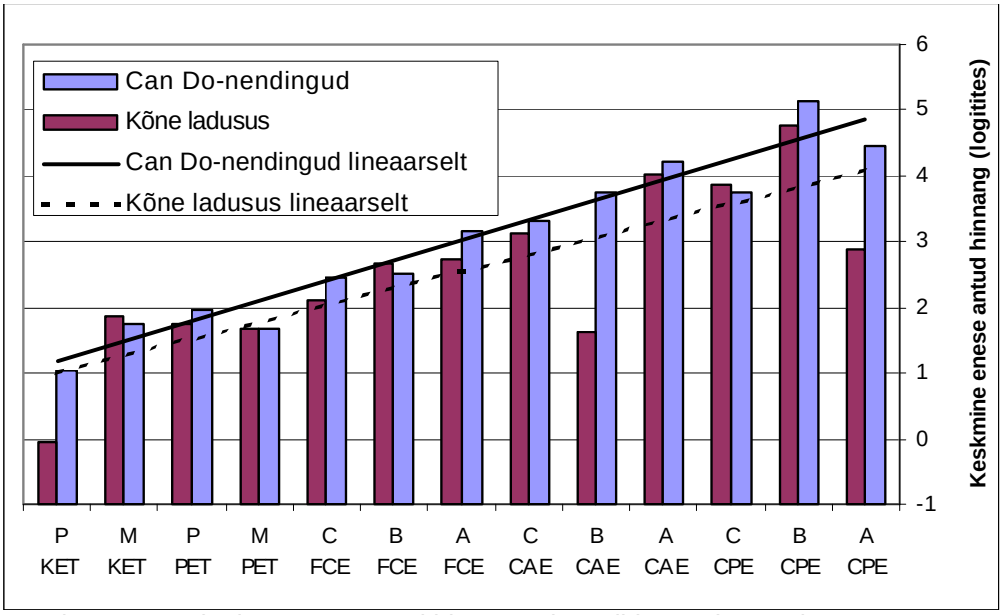
Arvamused selle kohta, kui usaldusväärne on enesehinnangute kasutamine tasemepiiride määratlemisel, on erinevad. Paljudes projektides on sel moel enesehindamist siiski kasutatud. Kui üldse probleeme tekib, siis mitte kirjelduskriteeriumite kalibreerimisega, sest testitavad kalduvad tajuma suhtelist keerukust üldjoontes sarnaselt. Õpetajate hinnangutel ja enesehinnangutel põhinevate kirjelduskriteeriumite raskusparameetrites on näha väga tugevat korrelatsiooni. Näiteks raamdokumendi uurimisprojektis oli eraldi õpetajate hinnangutest ja enese antud hinnangutest saadud kirjelduskriteeriumite skaalaväärtuste korrelatsioon 0,98. Vähem usaldusväärne on just testitavate endi hinnang oma keeleoskusele: nad kipuvad oma keeleoskust kas üle- või alahindama.

Üsna tihti tuleb ette, et madala keeleoskustaseme õpilased kalduvad oma võimeid üle hindama, kõrgema keeleoskustaseme õpilased aga peavad oma võimeid väiksemateks. Sellel on loomupärane seletus: madala keeleoskustasemega õppijad on oma teadmiste üle uhked ega mõista, kui palju neil veel õppida on, kõrgemal keeleoskustasemel ollakse seevastu aru saanud, et õppida on veel väga palju. Tegelikult ei ole nende tähelepanekute selgitamiseks üldse vaja teooriast tuge otsida. Enesehindamine kaldub oma iseloomult olema ekslik ja kummagi rühma õppijad eksivad tavaliselt samas suunas – madalama keeleoskustaseme õppijad ülespoole ja kõrgema keeleoskustaseme õppijad allapoole. See on komplitseeriv tegur, mida nende andmete tõlgendamisel võib ette näha.

Vaatamata ülalkirjeldatud põhjustest tulenenud individuaalsetele erinevustele enese antud hinnangutes leidis Jones (2002) ALTE *Can Do* projektis andmete rühmitamisel eksamite kaupa tugeva korrelatsiooni enese antud hinnangu ja Cambridge'i ESOLi eksami tulemusena määratud keeleoskustaseme vahel. Ta sidus 478 enese antud hinnangut ALTE *Can Do*-nendingute kohta ja raamdokumendi eriti stabiilsete kõne ladususe aspekte hõlmavate kirjelduskriteeriumite kohta ESOLi eksamitel saadud hinnetega viiel keeleoskustasemel.

Joonisel 3 on esitatud testitavate keskmine enesehinnang nende saadud eksamihinnete kaupa. Selliseid andmeid võimaldavad üsna üksikasjalikku tõlgendust, näiteks viitavad need sellele, et hinne A eksamil First Certificate of English (FCE) vastab ligikaudu hindele C järgmise keeleoskustaseme eksamil Certificate in Advanced English (CAE) (seda suhet kinnitavad teisedki tõendusmaterjalid).

Kui enesehindamisel saadud andmete absoluutses keeleoskustasemes on kahtlusi, on neid kõige parem ankurdada võrdluses õpetajate hinnangutega.



Joonis 3. Keskmine enese antud hinnang eksamihinnete kaupa (ALTE *Can Do* projektis)

3.3.4. Keeleoskustasemete piiride määratlemine

Peale selle, et tulemused saab koondada ühtseks resultaadiks ja raamdokumendi keeleoskustaseme kohta anda üldhinnangu, nagu kirjeldati eespool punkti 3.2.1 alapunktis c, on veel vähemalt kolm võimalust kasutada õpetajate hinnanguid ja enese antud hinnanguid raamdokumendi *Can Do*-nendingutest koosnevate kontroll-lehtede kohta selleks, et analüüsida IRT põhjal andmeid, mis on saadud nii objektiivselt hinnatavatest testiküsimustest kui ka subjektiivselt hinnatavatest kirjelduskriteeriumitest, et kehtestada standardid (ehk keeleoskustasemete alampiirid) analüüsi käigus saadud teetaskaalal.

a) Raamdokumendi eri keeleoskustasemete kirjelduskriteeriumite asukohta skaalal saab kasutada selleks, et otsustada, kus üks keeleoskustase lõpeb ja teine algab. See käsitlusviis põhineb eeldusel, et raamdokumendi kirjelduskriteeriumi keeleoskustase on hõlpsasti nähtav. Selleks oleks mõistlik kirjelduskriteeriumid analüüsi jaoks niimoodi ka kodeerida. Näiteks võib A2-taseme kirjelduskriteeriumile „Oskab *küsida lihtsat infot reisi kohta*" panna koodi „*A2-Reisiinfo*". Kodeeritud kirjelduskriteeriumid väljendavadki nüüd visuaalselt seda, millise osa iga raamdokumendi keeleoskustaseme kirjelduskriteeriumid skaalal enda alla võtavad.

   Analüüsija või töörühma käsitlusviis on käsiraamatu osas 6.8 kirjeldatud järjestusmeetodi üks juhistega variant, kus lihtsalt antakse põhjalikke suuniseid järjehoidja asukoha määramiseks. Eri keeleoskustasemete kirjelduskriteeriumid võivad aga mõnevõrra kattuda. Seetõttu on vaja otsustada tasemepiiri täpne asukoht ja selleks tuleb tasakaalustatult arvestada punktis 2.4 kirjeldatud tegureid.

b) Punkte, kuhu raamdokumendi eri keeleoskustasemete kirjelduskriteeriumid skaalal paigutatakse, saab seejärel kasutada lisateabena – omamoodi välisvalideerimise vahendina – tasemepiiride määratlemise töörühmas.

Seda lähenemisviisi kirjeldatakse käsiraamatu osas 7.5.4.2 testivälise valideerimise teema all, tuues näiteks järjestusmeetodi Hollandi riikliku eksamikeskuse CITO variandi.

c) Raamdokumendi uurimisprojekti kirjelduskriteeriumeid saab kasutada selleks, et rakendada otse tasemepiire mõõtmisskaalal, mis on raamdokumendi kirjelduskriteeriumite skaalatabelite aluseks. Nii seotakse tasemepiirid, mitte testiküsimused.

Seda käsitlusviisi kirjeldatakse 4. osas.

4. Raamdokumendi kirjelduskriteeriumite skaala vahetu rakendamine

Kui raamdokumendi kirjelduskriteeriumeid kasutatakse IRT-l põhinevas analüüsis eraldiseisvate küsimustena, võib uue uuringu käigus saadud õpetaja hinnangud või enese antud hinnangud teisendada raskusteks samal skaalal kui algse raamdokumendi uuringu puhul.

Sellise meetodi üks eelis on see, et hindamiseks määratud kirjelduskriteeriumite hulgast võib välja valida (vähemalt üldjoontes) suvalise väiksema komplekti. Seoses sellega tasub märkida, et North (2000 a: 268–270) on jaganud raamdokumendi uurimisprojekti skaala kirjelduskriteeriumid kolme rühma: a) normaalsed kirjelduskriteeriumid; b) mõnel määral varieeruvad kirjelduskriteeriumid (küsimuse toime); c) mudelile hästi vastavad ja kontekstist tingitud varieeruvuseta kirjelduskriteeriumid (sihtkeel, kasutatav keel, haridussektor, keelepiirkond). Viimast rühma hakati nimetama ka „suurepärasteks kirjelduskriteeriumiteks" ja need võivad olla sobilikud ankurküsimused tulevastes projektides. Need „suurepärased kirjelduskriteeriumid" kirjeldavad eelkõige ladusa suhtlusoskuse aspekte. Just seda rühma kasutas Jones (2002) ALTE *Can Do* projektis, mida kirjeldati eespool punktis 3.3.1.

Kui andmekogum hõlmab samade testitavate kohta nii raamdokumendi kirjelduskriteeriumeid kui ka uuritava eksami küsimusi, saab testiküsimuste prognoositava raskuse siduda kirjelduskriteeriumite raskuse kaudu ühe ja sama, s.t raamdokumendil põhineva skaalaga.

Uue skaala saab algses raamdokumendi uuringus kasutatud skaalaga ankurdada mitmel moel.

- Lähtuvalt kirjelduskriteeriumi raskusest:
  - ankurdades kirjelduskriteeriumitest testiküsimused nende raskusega algses uuringus (North 2000 a: 358–415);
  - viies läbi sõltumatu, ankurdamata IRT-l põhineva analüüsi, ent võrdsustades seejärel uue skaala ja raamdokumendi algse uuringu skaala, tuginedes raamdokumendi kirjelduskriteeriumite suhtelisele asukohale mõlemal skaalal.
- Lähtuvalt keeleoskustaseme piiridest:
  - ankurdades skaalaastmed vahetult algse uuringu tasemepiiridega (North 2000 a: 274), nagu on näidatud tabelis 1.

David (2007) on toonud näite nii raamdokumendi kirjelduskriteeriumite väga stabiilse allrühma („suurepärased kirjelduskriteeriumid") kui ka raamdokumendi uurimisprojekti skaala tasemepiiride kasutamisest. Ta kirjeldab, kuidas õpetajate hinnanguid võrreldi raamdokumendi kirjelduskriteeriumitega, et leida keeleoskustasemete B1, B2 ja C1 tasemepiirid kohalike keeleõppijate testi logitiskaalal. David sai raamdokumendi kirjelduskriteeriumite asukohta oma uuel logitiskaalal

kasutada selleks, et teisendada raamdokumendi uurimisprojekti skaala logitite tasemepiirid testiküsimuste ja kirjelduskriteeriumite kohaliku analüüsi põhjal loodud skaala logititeks. Seejärel ankurdati skaalaastmed nende uute, raamdokumendi kohaliku projekti tasemepiiridega ja saadi testiskaala, mis põhines raamdokumendi keeleoskustasemetel. Sellise meetodi korral ei seota raamdokumendi keeleoskustasemetega mitte testiküsimusi ega ülesandeid, vaid hoopis tasemepiirid.

Siinkohal tuleb tähelepanu juhtida probleemidele, mis võivad ühe skaala teiseks teisendamisel ette tulla. Logitiväärtused ***ei*** ole mõõtühikud, mida saab automaatselt ühest kontekstist teise tõsta. Need on näitajad, mis saadakse konkreetse andmete kogumise ja skaala moodustamise meetodiga. Ei ole mingit põhjust eeldada, et logitiskaalal, mille koostamiseks ankurdati osahinnetest koosneva neljapunktise skaala abil õpetaja hinnangute andmed kirjelduskriteeriumite kontroll-lehtedega, oleksid samasugused omadused (nt proportsionaalne vahe keeleoskustasemete vahel, üldine ulatus) kui skaalal, mille koostamiseks kasutati erinevate vertikaalselt seotud ja objektiivselt hinnatud kuulamistestide analüüsi IRT järgi. Suhe ei pruugi olla lineaarne. Raamdokumendi uurimisprojekti skaalal (nagu tabelis 1), on tasemed C1 ja C2 teistest keeleoskustasemetest poole laiemad, kui neid arvestatakse ka koos pluss-tasemetega. Sama täheldati ka ALTE *Can Do*-skaala kalibreerimisel ning see võib olla seotud asjaoluga, et C-tasemete jaoks on *Can Do*-nendinguid keeruline kirjutada. Sama ei pruugi aga ilmneda skaalade puhul, mille aluseks on objektiivselt hinnatavate, teksti vastuvõtu oskust mõõtvate testidega kogutud andmed. Cambridge'i ESOLi eksamite ühtne skaala, mis on koostatud empiiriliselt objektiivsete vastuseandmete põhjal, näitab tegelikult, et sedamööda, kuidas keeleoskustasemed tõusevad, nende ulatus väheneb. See annab tunnistust, et õppimise kasutegur kahaneb aja jooksul proportsionaalselt – mida kauem õpid, seda vähem see asja muudab.

Tabel 1. Raamdokumendi keeleoskustasemete ja pluss-tasemete piirid logitiskaalal

Keeleoskustasemed		**Tasemepiiri väärtus**	**Vahemik logitiskaalal**
C2		3,90	
C1		2,80	1,10
	B2+	1,74	1,06
B2		0,72	1,02
	B1+	–0,26	0,98
B1		–1,23	0,97
	A2+	–2,21	0,98
A2		–3,23	1,02
A1		–4,29	1,06
	Turist	–5,39	1,10

4.1. Standardimine FACETSi meetodil

Veel üks viis, kuidas kasutada raamdokumendi uurimisprojekti kirjelduskriteeriumeid ja skaala tasemepiire, seisneb sooritusnäidiste võrdlemises, mitte niivõrd objektiivhinnatavate testiküsimuste jaoks tasemepiiride määratlemisel.

Nagu eespool märgitud, annab FACETSi programmi (Linacre 2008) analüüs võimaluse võtta testitava õiglaseks hindamiseks sobivusstatistika ja standardvigade kõrval arvesse ka hindajate ranguse erinevust (Linacre 1989). FACETSi analüüsi jaoks saadakse vajalikud andmed üldjuhul hindamisseminaril, kus eri asjatundjad annavad erinevatele kriteeriumitele üldhinnanguid ja võib-olla ka analüütilisi hinnanguid. Käsiraamatu osas 5.7 soovitatakse sellistel seminaridel raamdokumendi suuliste ja kirjalike sooritusnäidiste keeleoskustaseme hindamisel kasutada võrdlusalusena vastavalt käsiraamatu tabelit C2 (RD tabelit 3) või käsiraamatu tabelit C4 (kirjutamise puhul). Samuti aitavad sellised seminarid erinevaid juhtumeid arutades üksmeelele jõuda. Analüüsi seisukohast on aga väga

oluline, et *enne* igasugust arutelu oleksid olemas sõltumatud hindamisotsused. Need annavad parema ülevaate hindajate tegevusest ja kriteeriumite kohaldamisest. Kui sooritusnäidised on FACETSi abil sõltumatute hindamisotsuste kaudu raamdokumendi keeleoskustasemetele kalibreeritud, saab tulemust võrrelda üksmeelse otsusega, milleni töörühm pärast arutelu jõudis. Nüüd saab heade sooritusnäidistena välja valida need, mis vastavad mudelile (s.t ei tekita segadust), mille suhtes saavutati suur üksmeel ja mille puhul on raamdokumendi keeleoskustase nii FACETSi analüüsi kui ka arutelujärgse üksmeele põhjal väga lähedane.

Kui hindamiskriteeriumitena kasutatakse käsiraamatu tabeleid C2 või C4 või ka teisi raamdokumendi kirjelduskriteeriumitele tuginevaid tabeleid, saab skaalavahemike (nn skaalaastmete) künnised ankurdada tasemepiiridega logitiskaalal, mis tugineb tabelis 1 esitatud raamdokumendi uurimisprojekti kirjelduskriteeriumite skaalale. Seejärel saab IRT-l põhinevad hinnangud testitava võimete kohta arvutada välja sama logitiskaala abil nagu raamdokumendi kirjelduskriteeriumite puhul. Sellisel puhul tuleb aga ühe skaala teiseks teisendamisel olla ikka ettevaatlik. Skaalaastmed näivad pärast raamdokumendi uurimisprojekti skaalale ankurdamist väga ühtlased. Tähtis on kontrollida, kas ankurdamata skaalal on määratlused sama selged. Kehvemini määratletud skaala põhjuseks võib olla hindajate otsuste vähene kokkulangevus, pluss-tasemete erinev kasutusviis või ka lihtsalt see, kuidas näidised hindamiseks välja jagatakse. Sellist mõju tuleks uurida enne ankurdamist, sest pärast seda muutub see mõttetuks.

Sellist FACETSi lähenemisviisi katsetati 1997. aastal Fribourgis toimunud konverentsil Šveitsi raamdokumendi / Euroopa keelemapi uurimisprojekti lõpus, seda rakendati esimesel rahvusvahelisel raamdokumendi sooritusnäidiste võrdlusseminaril (Jones 2005; Lepage ja North 2005 a, 2005 b) ning hiljem on seda korduvalt kasutatud samalaadsetel eri keeli käsitlenud sündmustel, samuti 2008. aastal Sèvres'is toimunud eri keelte võrdlusseminaril (Breton, Jones, Laplannes, Lepage ja North; ilmub peagi).

5. Järjestuse kasutamine keeleülesel tasemepiiride määratlemisel

Need, kes on kasutanud käsiraamatu tööversiooni ja raamdokumendi näitlikke materjale eri keelte jaoks, on vahel seadnud kahtluse alla, kas väidetavalt samal keeleoskustasemel olevaid eri keelte näidiseid saab võrrelda. On mitu põhjust, mis annavad alust väita, et näidised, mis peaks olema samal keeleoskustasemel, ei näi võrreldavad:

- testitavad võivad olla samal üldtasemel, ehkki nende oskuste üksikasjalik kirjeldus on väga erinev;
- testitavate puhul võivad rolli mängida eri konstruktide aspektid (täiskasvanu/teismeline; võõrkeele õppija / sisserändaja jne);
- taseme määramise aluseks on sooritusnäidis, mis kajastab keeleoskust puudulikult; kui sooritatavad ülesanded on väga erinevad, on testitavate vahetu võrdlemine keeruline.

Teisalt tekib küsimus, kas näidiseid esitavatel organisatsioonidel on õigupoolest ühtne arusaam raamdokumendist. Kas nad kasutavad sama raamistikku või järgivad selle kohalikku tõlgendust? Olenemata sellest, kui põhjalikult on raamdokumendiga tutvutud, standardimiskoolitus läbi viidud või järjepidevuse ja kokkulangevuse indeksid prognoositud, ei saa kuidagi tõendada, et mingi keele oskuse tasemele hinnangut andvat konkreetset asjatundjate rühma ei mõjuta nende kultuuritaustast tingitud tõlgendus ülesandest. Aga miks peaksimegi eeldama, et see neid ei mõjuta?

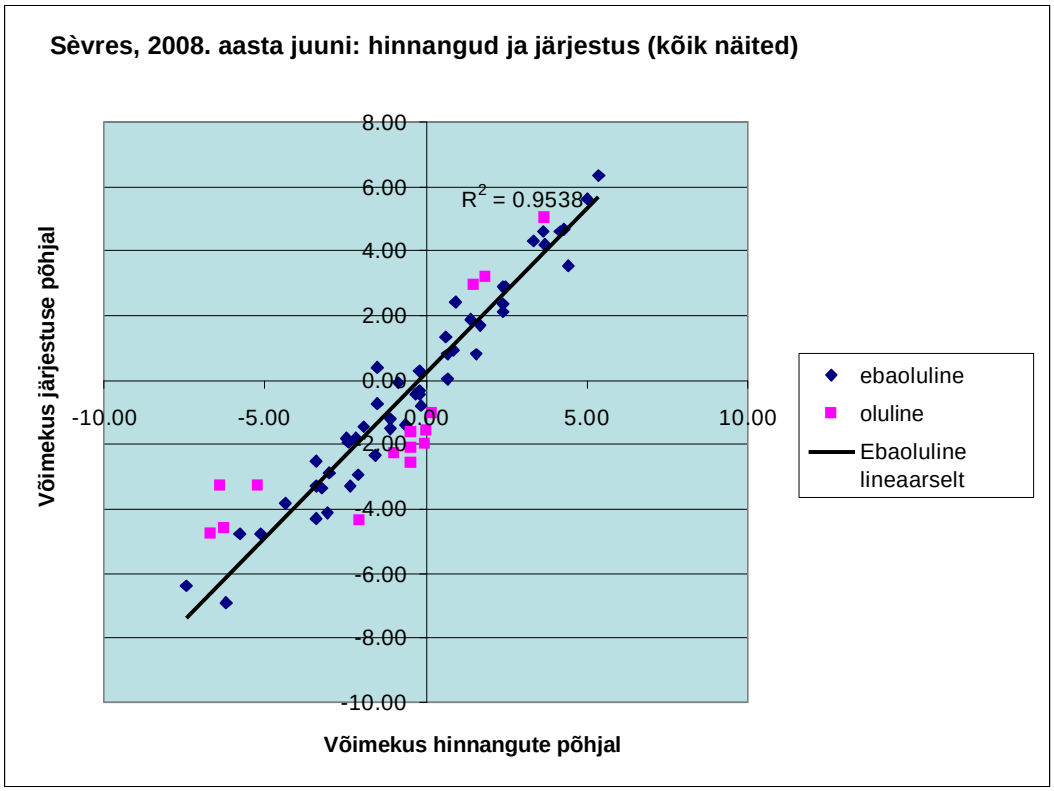
Üks võimalus selle probleemiga toime tulla on kasutada mitmekeelseid hindajaid ja paluda neil hinnata kaht või enamat keelt, nii nagu käsiraamatu osas 6.10.3 on kirjeldatud. Seda tehti ka 2008. aasta juunis Sèvres'is toimunud eri keelte võrdlusseminaril, kus mitmekeelsed hindajad hindasid kõigepealt inglise ja prantsuse keele näidiseid ning seejärel väiksematesse rühmadesse jaotatuna

esmalt inglise ja prantsuse keele ankurnäidiseid ja seejärel muid kolmanda keele (vastavalt saksa, hispaania ja itaalia keele) näidiseid.
Testitavakesksete meetodite kasutamist koos raamdokumendi kirjelduskriteeriumitel põhineva õpetajapoolse hindamisega ning seejärel nende hinnangute põhjal keelte samale skaalale ankurdamist käsitleti punktides 3.2 ja 4.

Alternatiiviks on järjestamine, mille aluseks on üldtuntud tõde, et igasugune hindamine hõlmab ka võrdlemist. Iga standardi aluseks on norm ja standardi kehtestamine hõlmab soorituse või testiküsimuse võrdlemist mingi omaksvõetud normiga. Siin esitatud käsitlusviisis pannakse rõhku võrdlustele, kus vahendajaks on raamdokumendi kirjelduskriteeriumid. Kuid iga katse siduda keelesooritust või testi raamdokumendiga on ikkagi kaudne võrdlus teiste keelesoorituste ja testidega.

Paarisvõrdluse meetod (Thurstone 1927) põhineb ideel, et mida kaugemal kaks võrreldavat objekti varitunnuste skaalal on, seda suurem on tõenäosus, et üks neist osutub võrdluses võitjaks. Selle käsituse puuduseks on kordamine ja vajalike paarisvõrdluste tohutu arv. Bramley (2005) pakub atraktiivse ja praktilise alternatiivina välja järjestuse, kus võrreldakse rohkem kui kaht tunnust. Kui hindajarühm järjestab 10 näidist selliselt, et igale näidisele määratakse sõltuvalt asukohast skoor ühest kümneni, saab neid andmeid kasutada peale õiges järjestuses kokkuleppimise ka selleks, et määrata nende suhteline asukoht Raschi mudelil põhineval mõõtmisskaalal.

Esimest korda kasutati järjestamist keeleüleses võrdluses 2008. aasta juunis Sèvres'is toimunud eri keelte võrdlusseminaril. Enne seminari paluti osalejatel järjestada suurem hulk näidiseid, kui seminaril oleks reaalselt olnud võimalik analüüsida. Selleks kasutati spetsiaalset veebiplatvormi, kus osalejatel oli võimalik näidiseid vaadata ja neid nimekirjas ümber paigutades järjestada. Näidiste jagamisel järgiti põhimõtet, et iga hindaja hindaks kaht keelt ja et andmed oleksid näidiste ja keelte lõikes seotud.



Joonis 4. Järjestuse ja hinnangute võrdlus (Sèvres, juuni 2008)

Joonisel 4 on võrreldud järjestamiseks ja hindamiseks esitatud näidiste põhjal võimekust, mis on tuletatud järjestusest (andmed kogutud veebilehel enne seminari) ja hinnangutest (kogutud seminaril).

Heledamate ruudukestega on väljendatud tulemused, mis lahknevad palju rohkem, kui mõõtmisviga võimaldaks. Tulemustes on ilmselgelt olulisi erinevusi, kuid arvestades, et järjestamine toimus enne seminari veebis individuaalselt, ilma juhiste, arutelude või meetodi tutvustamiseta ja et juhised võisid põhjustasid mõningast segadust, ei ole see üllatav. Mõlema andmekomplekti vaheline korrelatsioon on kõigil juhtudel ikkagi suur – 0,94. Esimese uuringu tulemuste järgi on järjestuspõhimõte igatahes paljutõotav.

Järjestamine võib osutuda äärmiselt kasulikuks meetodiks, sest selle tulemused ei sõltu arusaamast raamdokumendi keeleoskustasemetest. Sellegipoolest on ülimalt tähtis, et hindajad mõistaksid, millistel alustel ja milliste kriteeriumite põhjal nad sooritusi hindama peavad. Tänu sellele suudavad nad luua seose raamdokumendiga ilma tasemepiiride määratlemise seminari läbimata.

Samuti pakub see võimalust laiendada sagedamini õpetatavate keelte (inglise, prantsuse, saksa, hispaania, itaalia keele) sooritusnäidiste komplekti, mida uuriti Sèvres'i seminaril, vähem õpetatavatele keeltele, säilitades keeltevahelise võrreldavuse. Kujutlegem, et sellise mitut keelt hõlmava järjestamisseminari tulemusel jõutakse laialdaselt õpitavate keelte puhul usaldusväärse näidisekogumini. Seejärel saab üht või mitut nendest keeltest kasutada võrdlusalusena järjestamisel ka mõnes teises keeles. Nende võrdlusaluste tasemepiire saab kohaldada otse selle keele suhtes ja nende tõlgendus peab olema samasugune. Sellise sidumisprotsessi standardviga on võimalik prognoosida (ehkki töö selles valdkonnas veel käib) ja nii saame kindlalt öelda, kui täpne on tulemus – seda ei ole võimalik teha siis, kui tasemepiiride määratlemisel keskendutakse ühele keelele. Keeleülene võrdlus ei ole seega pelgalt üks valdkond, mis pakuks huvi konkreetselt eri keelte testimisega tegelevatele asutustele. See on paljulubav ja laialt rakendatav võimalus iga sidumisprojekti jaoks.

Nagu kõigi teiste võrdlus- ja tasemepiiride määratlemise meetodite puhul, tunneme ka siin end kindlamalt väljendusoskustega, milles hinnatakse pigem jälgitavat käitumist kui testiküsimusi. See, kas võrdlusmeetodit saab ka sel puhul edukalt kasutada, ei ole veel selge. On õige, et hindajad ei ole üldjuhul suutnud väga hästi testiküsimuste suhtelist raskust prognoosida ning seetõttu ongi mõningate tasemepiiride määratlemise meetodite puhul tavaks hindajaid testiküsimuste raskusest teavitada. Kuid välja võiks töötada korra, mille kohaselt järjestaksid hindajad testiküsimusi nii ühes kui ka kahes keeles. Keeleülest võrdlust saab teha mitmel moel – esitada teabe testiküsimuste suhtelise raskuse kohta kas ühe või mõlema testiküsimuste komplekti kohta või mitte kummagi kohta. Tulemused saaks omakorda panna korreleeruma empiiriliste andmete põhjal kalibreeritud testiküsimustega, et luua ühe keele jaoks tõenäolise täpsuse indeks ja laiendada seda mitmekeelsele juhtumile. See valdkond vajab aga veel uurimist.

6. Kokkuvõte

Käesolevas dokumendis, kus keskendutakse skaleerimismeetodite kasutamisele ja testiküsimuste andmebaasi loomisele, rõhutatakse tasemepiiride määratlemise ja standardi järjepidevuse tagamise osatähtsust testi korraldamise tsüklis. Kui olemas on stabiilsed mõõtmisskaalad, hõlbustab see tasemepiiride määratlemist ja täpset kohandamist, samuti võimaldab nende pikaajalisel kohaldamisel saavutada järjepidevuse ja põhimõtteühtsuse. Kindlasti ei ole see lühike tee, sest andmepõhise ja testitavakeskse käsitlusviisi juures ei piisa väiksearvulise töörühma liikmete koolitamisest – koolitada on vaja paljusid õpetajaid. Kindel on aga, et see võimaldab tulemuslikumalt eesmärgi poole liikuda. Keeletestijate jaoks on põhiülesanne heade testiküsimuste koostamine. Oskuslikuks testiküsimuste koostajaks saadakse aastatega, kuid selleks tehtud pingutused tasuvad end kindlasti ära. Samuti tasub end ära toimiva testiküsimuste andmebaasi meetodi rakendamine, sest see vähendab tublisti vajadust hindamiste järele standardi edaspidisel täiustamisel. Selline lähenemisviis eeldab aga testide väljatöötamise töörühmalt käsiraamatu 3., 4. ja 5. peatükis kirjeldatud raamdokumendiga tutvumise, eristuskirja koostamise ja standardimise koolituse tõhusat läbimist.

Käesolevas dokumendis püüdsime näidata ka seda, kuidas saaks sidumisprojekti kõikidesse etappidesse lõimida testivälise kriteeriumi. Seos väliskriteeriumiga on iga sidumisprojekti tuum, nii et mida paremini suudetakse selline kriteerium projekti kaasata, seda suurem on tõenäosus saavutada hea tulemus.

Selle materjali kasutajad võiksid mõelda:

- *kuidas kasutada eeltestimise ajal raamdokumendi sooritusnäidiseid vastava(te) osaoskus(t)e kohta;*
- *kas neid huvitab pigem ühe konkreetse ja tulemusega „arvestatud/mittearvestatud" väljendatud standardi sidumine (töörühmapõhine käsitlusviis) või eri keeleoskustasemete jaoks mõeldud testiseeria standardite sidumine (skaalapõhine käsitlusviis);*
- *kuidas nad lähenevad keeleülese tasemepiiride määratlemise küsimusele;*
- *kas kaasata oma projekti õpetajaid olenemata sellest, millist tasemepiiride määratlemise meetodit kasutatakse;*
- *kas neil on olemas vajalikud teadmised, et viia läbi andmepõhine ja testitavakeskne IRT-le tuginev tasemepiiride määratlemise uuring, või kas neil on võimalik hankida selliseid teadmisi mujalt;*
- *kas uuritava(te) testi(de) sisu on tänu käsiraamatu 4. peatükis kirjeldatud eristuskirja koostamisele piisavalt selgesti raamdokumendiga seotud, et sellises uuringus tasemepiiride määratlemisel oleks piisav valiidsus;*
- *kas on piisavalt palju õpetajaid, kellel on küllaldased teadmised raamdokumendist, ja kas on võimalik pakkuda neile põhjalikumat koolitust (raamdokumendiga tutvumine, standardimiskoolitus), et nad saaksid anda raamdokumendile tuginevaid valiidseid hinnanguid, mida saaks ka uuringus kasutada;*
- *kas Euroopa keelemapi kontroll-lehti saab sellises uuringus asjaomases kontekstis kasutada (s.t kas kõnealune keelemapp on valideeritud; kas keelemapi kõiki kirjelduskriteeriumeid saab tõesti siduda raamdokumendi algsete kirjelduskriteeriumitega või kas tuleks otsida alternatiive andmebaasist aadressil www.coe.int/portfolio);*
- *kas raamdokumendi tabelist 3 (käsiraamatu tabelitest C2–C4) peaks asjakohaseid keeleoskustasemeid ja kategooriaid valides tuletama teisi hindamistabeleid.*

Kasutatud kirjandus

Ashton, K. (2008): *Comparing proficiency levels in an assessment context: the construct of reading for secondary school learners of German, Japanese and Urdu. Cambridge Esol Manuscript.*

Baker, R. (1997): *Classical Test Theory and Item Response Theory in Test Analysis. Extracts from: An Investigation of the Rasch Model in Its Application to Foreign Language Proficiency Testing.* Language Testing Update Special Report No 2.

Boldt, R. F., Larsen-Freeman, D., Reed, M. S., Courtney, R. G. (1992): *Distributions of ACTFL Ratings by TOEFL Score Ranges.* Research Report RR-92-59. Princeton, New Jersey: Educational Testing Services.

Bramley, T. (2005): A Rank-Ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement,* 6 (2), 202–223.

Breton, Jones, Laplannes, Lepage, North, (ilmub peagi): *Seminaire interlangues / Cross language benchmarking seminar, CIEP Sevres, 23-25 June 2008: Report.* Strasbourg: Council of Europe.

Council of Europe (2001): *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe (2007): „The *Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities."* Intergovernmental Language Policy Forum, Strasbourg, 6-8 February 2007, Report.

David, G. (2007): *Building a Case for Euro Examinations: a case study.* Paper given at the Seminar for a joint reflection on the use of the preliminary pilot version of the Manual for „Relating

Language Examinations to the CEFR" 2004–2007: Insights from Case Studies, Pilots and other projects. Cambridge, United Kingdom, 6-7 December 2007.

Grin, F. (1999): *Competences linguistiques en Suisse: Beneficesprives, benefices sociaux et depenses, rapport de valorisation.* Berne/Aarau, PNR33/CSRE.

Grin, F. (2000): *Fremdsprachenkompetenzen in der Schweiz: Privater Nutzen, gesellschaftlicher Nutzen und Kosten, Umsetzungsbericht.* Bern/Aarau, NFP33/SKBF.

Jones, N. (2002): Relating the ALTE Framework to the Common European Framework of Reference. In Alderson, J. C. (ed.) (2002): *Case studies in the use of the Common European Framework.* Strasbourg: Council of Europe, ISBN 92-871-4983-6: 167–183.

Jones, N. (2005): Seminar to calibrate examples of spoken performance, CIEP Sevres, 02.–04.12.2004. Report on analysis of rating data. Final version. March 1st 2005. Cambridge ESOL internal report.

Jones, N., Ashton, K., Walker, T. (2007): *Asset Languages: A case study piloting the Manual.* Paper given at the Seminar for a joint reflection on the use of the preliminary pilot version of the Manual for „Relating Language Examinations to the CEFR" 2004–2007: Insights from Case Studies, Pilots and other projects. Cambridge, United Kingdom, 6-7 December 2007.

Lepage S., North, B. (2005 a): *Seminaire pour le calibrage des productions orales par rapport aux echelles du Cadre europeen commun de reference pour les langues, CIEP, Sevres, 2-4 decembre 2004* : Rapport. Strasbourg: Council of Europe DGIV/EDU/LANG (2005) 1.

Lepage, S., North, B. (2005 b): *Guide for the organisation of a seminar to calibrate examples of spoken performance in line with the scales of the Common European Framework of Reference for Languages.* Strasbourg: Council of Europe DGIV/EDU/LANG (2005) 4.

Linacre, J. M. (1989): *Multi-faceted Measurement.* Chicago: MESA Press.

Linacre, J. M. (2008): *A User's Guide to FACETS. Rasch Model Computer Program.* ISBN 0-941938-03-4. www.winsteps.com.

North, B. (2000 a): *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.

North, B. (2000 b): Linking Language Assessments: an example in a low-stakes context. System_28: 555–577.

Szabo, G. (2007): *Potential Problems concerning the Empirical Validation of Linking Examinations to the CEFR.* Paper given at the Seminar for a joint reflection on the use of the preliminary pilot version of the Manual for „Relating Language Examinations to the CEFR" 2004–2007: Insights from Case Studies, Pilots and other projects. Cambridge, United Kingdom, 6-7 December 2007.

Thurstone, L. (1927): A Law of Comparative Judgement. *Psychological Review*, 3, 273–286.

Wilson, K. M. (1989): *Enhancing the Interpretation of a Norm-referenced Second Language Test through Criterion Referencing: A research assessment of experience in the TOEIC testing context.* TOEIC Research Report No 1. RR-89-39. Princeton, New Jersey: Educational Testing Services.

Wilson, K. M. (1999): *Validating a Test designed to assess ESL Proficiency at Lower Developmental Levels.* Research Report RR-99-23. Princeton, New Jersey: Educational Testing Services.

Wilson, K. M. (2001): *Overestimation of LPI Ratings for Native Korean Speakers in the TOEIC Testing Context: Search for explanation.* Research Report RR-01-15. Princeton, New Jersey: Educational Testing Services.

Wilson, K. M., Lindsey, R. (1999): *Validity of Global Self-ratings of ESL Speaking Proficiency based on an FSI/ILR-referenced Scale.* Research Report RR-99-13. Princeton, New Jersey: Educational Testing Services.