Author(s): Ahola, Sari; Neittaanmäki, Reeta; Hirvelä, Tuija

Title: Puheen ymmärtämisen tehtävien taitotasoille asettaminen Yleisissä kielitutkinnoissa

Year: 2016

Version:

Please cite the original version:

Ahola, S., Neittaanmäki, R., & Hirvelä, T. (2016). Puheen ymmärtämisen tehtävien taitotasoille asettaminen Yleisissä kielitutkinnoissa. In A. Huhta, & R. Hildén (Eds.), Kielitaidon arviointitutkimus 2000-luvun Suomessa (pp. 68-88). Suomen soveltavan kielitieteen yhdistys. AFinLA-e : soveltavan kielitieteen tutkimuksia, 9. http://journal.fi/afinla/article/view/60847

Huhta, A. & R. Hildén (toim.) 2016. Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016 / n:o 9. 68–88.

Sari Ahola, Reeta Neittaanmäki & Tuija Hirvelä

Jyväskylän yliopisto

Puheen ymmärtämisen tehtävien taitotasoille asettaminen Yleisissä kielitutkinnoissa

The aim of this study is to explore rater effects, such as consistency and severity of the judgements of 12 expert judges, and to report the extent of assessment criteria used by them. The data of the study are based on the three different phases of the standardization procedure which were analyzed using the MFRM model. In addition, feedback gathered from the expert judges was used to find out their opinions about the procedure and to validate the results of the standardization. The results of this study indicate that general discussion on and statistical information about the items influenced judges' consistency and severity and their adaptation of the assessment criteria. The judges themselves also found that these factors affected their judgements. In addition, the nature of listening and the criteria caused difficulties during the procedure.

Keywords: standard setting, language testing, expert judges, proficiency levels, test items

Asiasanat: taitotasoille asettaminen, kielitaidon arviointi, asiantuntijat, taitotasot, tehtäväosiot

1 Johdanto

Yleisten kielitutkintojen osallistujilla on erilaisia käyttötarkoituksia tutkintotodistukselle. (Ks. lisää tutkinnon käyttötarkoituksesta ja osallistujista: Neittaanmäki & Hirvelä 2014.) Osallistujalle todistus on usein varsin tärkeä dokumentti, sillä sen perusteella hänen kyvyistään viestiä ja ymmärtää kieltä tehdään päätelmiä, joilla saattaa olla hyvinkin kauaskantoisia vaikutuksia hänen tulevaisuudelleen. Toisaalta taas tutkintotodistusta edellyttävien tahojen, kuten viranomaisten, työnantajien ja koulutusta tarjoavien, on myös voitava luottaa siihen, että todistuksessa olevat arviot vastaavat osallistujan taitoa ja että sen perusteella tehty tulkinta osallistujan kyvystä vastaa todellisuutta. Jotta tutkinnon osoittamia tuloksia voidaan pitää luotettavina, edellytetään tutkinnon järjestäjältä toimenpiteitä. On pyrittävä erilaisin keinoin varmistamaan, että arvioinnissa toteutuvat laadukkaan arvioinnin periaatteet, ts. arviointi on objektiivisuuteen pyrkivää, oikeudenmukaista, läpinäkyvää ja luotettavaa sekä eettisesti kestävää. (Ks. lisää validoinnista: Messick 1989, 1994; Weir 2005.)

Määriteltäessä tutkintoihin osallistuvien kielitaitoa arvioinnin tueksi tarvitaan tehtäviä, joiden avulla osallistujat voivat osoittaa kielitaitonsa. Vaikka tutkintojärjestelmässä saadaan jokaisen testikerran jälkeen tilastollisiin analyyseihin pohjautuvaa tietoa tehtävien toimivuudesta ja vaikeustasosta, tarvitaan näiden tietojen tueksi vielä tutkintokielen asiantuntijoiden arvioita tehtävien vaativuustasoista ja testin pisterajoista.

Standard setting eli taitotasoille asettaminen tarkoittaa yleensä testin pisterajojen määrittämistä suhteessa kielitaidon kehittymistä kuvaavaan asteikkoon (Reckase 2009). Käytännössä tällä tarkoitetaan sitä, että kielen asiantuntijat määrittelevät tehtävissä vaadittavan osaamisen tason kielitaidon kehittymistä kuvaavien taitotasokuvainten avulla. Tämä prosessi on tärkeä vaihe arvioinnin ja koko tutkintojärjestelmän validoinnissa, sillä jos kielitaidosta halutaan tehdä luotettavia päätelmiä testin avulla, on varmistettava, että testin sisältö sopii käyttötarkoitukseen ja että testissä käytettävät tehtävät mittaavat oikean tasoista kielitaitoa (Bejar 2008). Bejar (2008) ja Cizek ja Bunch (2007) pitävätkin tärkeänä, että standard setting liitetään olennaiseksi osaksi testin kehittämistä ja suunnittelua.

Yleisissä kielitutkinnoissa on viimeisen neljän vuoden aikana järjestetty tilaisuuksia, joissa suomen kielen asiantuntijat ovat asettaneet tekstin ymmärtämisen tekstejä ja niihin liittyviä tehtäväosiota[1] taitotasoille, mutta puheen ymmärtämisen tehtävien standardointiprosessi on tutkintojärjestelmässä vasta aluillaan. Standardointiprosessit näissä ymmärtämisen taidoissa ovat varsin samankaltaiset, koska tekstin

1 Tehtäväosiolla viitataan tekstikohtaan liittyvään yksittäiseen kysymykseen, jolle annetaan erillinen pistemäärä. Tehtävä puolestaan sisältää sekä tekstin että siihen liittyvät osiot.

ja puheen ymmärtäminen ovat myös taitoina monilta osin samanlaisia. Molemmissa taidoissa tarvitaan mm. lingvistisiä taitoja, kontekstin tuntemusta, maailmantietoa ja päättelytaitoa.

Puheen ymmärtämisen testaukseen liittyy kuitenkin piirteitä, jotka on hyvä huomioida asetettaessa tehtäviä taitotasoille. Puheen ymmärtäminen on luonteeltaan reaaliajassa tapahtuvaa, ja se edellyttää osakokeen suorittajalta auditiivisten ja lingvististen taitojen lisäksi semanttisia ja kognitiivisia taitoja sekä hyvää työmuistia. Testiin osallistujan on hallittava nauhalta tulevan tekstin kuunteleminen, kuullun tiedon tunnistaminen ja sen prosessointi sekä tulkinta. Tämän lisäksi on ymmärrettävä kuunneltavaan tekstiin liittyvät tehtäväosiot ja vastattava niihin. Yksittäiset sanastolliset tai syntaktiset vaikeudet kuuntelun aikana saattavat aiheuttaa sen, että testiin osallistuja ymmärtää kuulemastaan vain vähän tai ei mitään, eikä hän välttämättä pysty enää korjaamaan tilannetta, koska puheen ymmärtämiseen ei voi palata samoin kuin kirjoitettuun tekstiin. Lisäksi puheen ymmärtämiseen liittyy myös monia ymmärrettävyyteen ja vaikeustasoon vaikuttavia ulkoisia tekijöitä, kuten äänitteellä kuuluvien puhujien puhenopeus ja artikulaatio. (Lisää puheen ymmärtämisestä: Buck 2001; Rost 1990.)

2 Tutkimuskysymykset ja tutkimuksen tarkoitus

Tutkimuksemme on tapaustutkimus (*case-study*), jonka avulla pyrimme saamaan yksityiskohtaisempaa tietoa eri standardointimenetelmien soveltuvuudesta testijärjestelmään. Tapaustutkimusta käytetäänkin usein pyrittäessä ymmärtämään tutkittavaa asiaa kokonaisvaltaisemmin, jotta tutkimuksen avulla saatuja tuloksia voidaan siirtää käytäntöön. (Tapaustutkimuksesta: Saarela-Kinnunen ja Eskola 2001.) Tutkimuksemme nivoutuukin tiiviisti yhteen Yleisten kielitutkintojen kehittämisen kanssa. Tavoitteena on ennen kaikkea luoda tutkintojärjestelmän omaan testauskontekstiin sopivat käytänteet ja vahvistaa näin entisestään arviointiprosessin luotettavuutta.

Taustoitamme tutkimustamme ensin kuvaamalla artikkelissa lyhyesti ensimmäisen suomen kielen puheen ymmärtämisen standardointiprosessin. Tämän jälkeen analysoimme prosessia sekä kvantitatiivisen että kvalitatiivisen aineiston avulla. Tuloksissa esittelemme, kuinka yhdenmukaisia asiantuntijoiden näkemykset tehtävien taitotasoista olivat ja mitä muutoksia ankaruuseroissa ja arviointilinjan johdonmukaisuudessa havaittiin standardointiprosessin eri vaiheissa. Lisäksi tarkastelemme asiantuntijoilta saatujen kirjallisten kommenttien avulla, mitkä tekijät vaikuttivat taitotason määrittelyyn ja mitä vaikeuksia taitotasolle määrittelyyn liittyi.

Yleisissä kielitutkinnoissa taitotasoille asettamisen tavoitteena on arvioida yksittäisten testitehtäväosioiden vaativuutta sen sijaan, että määritellään tiettyyn taitotasoon vaadittava pistemäärä kokonaiselle osakokeelle. Koska Yleisissä kielitutkinnoissa

tehtävät eivät ole kertaluonteisia, tutkintojärjestelmässä on käytössä osiopankki, joka koostuu hyvin toimivista tehtäväosioista. Jokaiselle tehtäväosiolle on määritelty Raschin mallin avulla vaikeustasoa kuvaava lukuarvo, jonka perusteella tehtäväosiot sijoittuvat vaikeustasoa kuvaavalle skaalalle (Raschin mallista: Wright & Stone 1979; Törmäkangas & Törmäkangas 2009). Yleisten kielitutkintojen standardoinnin tavoitteena onkin etsiä tälle vaikeustasoskaalalle taitotasojen katkaisukohdat, jotta voidaan varmistaa se, että testikerrasta riippumatta testikokonaisuudet ovat vertailukelpoisia ja samanlaatuisia keskenään. Tavoitteen saavuttamisen ehtona on kuitenkin se, että standardointia tekevät päätyvät arvioissaan mahdollisemman yhdenmukaiseen käsitykseen testitehtävien ja -osioiden vaatimasta osaamisesta eli taitotasosta.

3 Tutkimusaineisto

3.1 Aineisto

Puheen ymmärtämisen standardointiprosessiin osallistuneet asiantuntijat asettivat taitotasoasteikolle 35 suomen keskitason puheen ymmärtämisen tehtäväosiota, jotka edustivat sekä tehtävätyypeiltään että vaikeustasoiltaan kattavasti Yleisten kielitutkintojen osiopankkia. Valitut tehtäväosiot asetettiin taitotasoille standardointiprosessia varten koottujen taitotasokuvainten avulla. Kuvaimet koottiin eri dokumenteista: Eurooppalaisesta viitekehyksestä, Yleisten kielitutkintojen perusteista ja laadintaohjeista. Tällä pyrittiin varmistamaan se, että puheen ymmärtämisen taidosta saadaan mahdollisimman kattava kuva ja että kuvaus vastaa tutkintojärjestelmän dokumenteissa kuvattua taitoa.

Jotta tuloksia voidaan pitää luotettavina, myös prosessiin osallistuvien asiantuntijoiden on täytettävä tietyt vaatimukset. Cizekin ja Bunchin mukaan (2007) asiantuntijat ja heidän ominaisuutensa vaikuttavat standardoinnin tuloksiin vähintään yhtä paljon – tai jopa enemmän – kuin valittu *standard setting* -menetelmä. Asiantuntijoilta edellytetään ensisijaisesti vahvaa asiantuntemusta mitattavasta asiasta, mutta myös muut taustatekijät, kuten sukupuoli, ikä ja alueellisuus, on huomioitava valittaessa tarkoitukseen sopivaa asiantuntijaryhmää. Standardoinnin tavoite ja asiantuntijoille asetetut edellytykset määrittelevät myös pitkälti asiantuntijoiden lukumäärän. Riittävänä määränä luotettavien tulosten saamiseksi voidaan kuitenkin pitää vähintään 10 asiantuntijaa (ks. lisää: Kaftandjieva 2010).

Yleisten kielitutkintojen standardointiprosessiin valittiin 12 suomen kielen asiantuntijaa. He olivat eri-ikäisiä, pääosin S2-opettajina eri puolilla Suomea toimivia henkilöitä, jotka ovat toimineet vuosia Yleisissä kielitutkinnoissa joko arvioijina ja/tai tehtävän laatijoina. He tunsivat hyvin tutkintojärjestelmän, suomen kielen tutkinnon

ja taitotasojen asettamisessa käytettävät taitotasokuvaimet sekä heillä oli päivätyönsä kautta kokemusta testiin osallistujien kaltaisista kielenoppijoista. Lisäksi heillä kaikilla oli aikaisempaa kokemusta Yleisten kielitutkintojen tekstin ymmärtämisen tehtäväosioiden taitotasoille asettamisesta ja siihen liittyvästä prosessista.

Koska kyseessä on tutkintojärjestelmän kehittämiseen tähtäävä tutkimus, oli tärkeä myös huomioida asiantuntijoiden mielipiteet prosessista. Asiantuntijoilta kerättyä palautetta voidaan käyttää apuna kehitettäessä tutkintojärjestelmään puheen ymmärtämisen standardointikäytänteitä, mutta palautekyselyn vastauksia voidaan hyödyntää myös osana standardointiprosessin tulosten luotettavuuden tarkastelua. Tästä syystä käsittelemme myös artikkelimme loppuosassa asiantuntijoilta saatua palautetta, jonka olemme analysoineet sisällönanalyysimenetelmää käyttäen. (Sisällönanalyysista: Tuomi & Sarajärvi 2009.)

3.2 Standardointiprosessin kuvaus

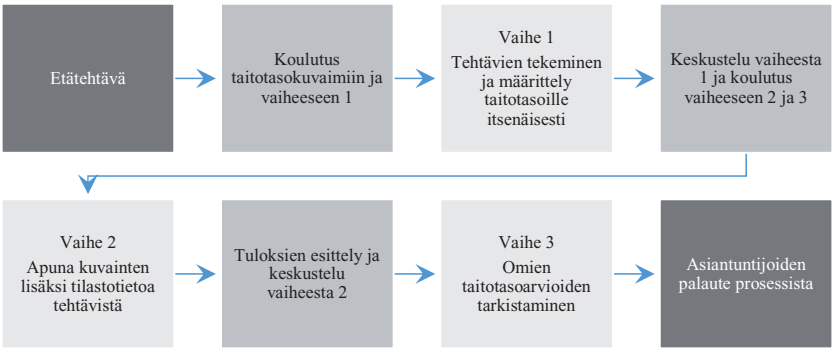
Erilaisia menetelmiä taitotasojen asettamiseksi on kymmeniä. Menetelmän valinta on tehtävä huolellisesti, sillä se vaikuttaa tuloksiin, ja eri menetelmillä on myös erilaiset vaatimukset aineistolle. (ks. menetelmistä: Cizek & Bunch 2007; Kaftandjieva 2010.) Reckasen (2009) mielestä ongelmallista standardoinnissa onkin juuri se, että yhtenäistä *standard setting* -teoriaa ei ole olemassa. Eri menetelmät tuottavat usein toisistaan poikkeavia tuloksia, ja tulokset ovat riippuvaisia monista prosessiin ja sen toteutukseen liittyvistä yksityiskohdista.

Yleisissä kielitutkinnoissa standardoinnissa käytettävän menetelmän valinta perustui standardoinnin tavoitteiden lisäksi olemassa olevaan IRT (*item response theory*) -pohjaiseen osiopankkiin, käytössä oleviin tehtävätyyppeihin ja taitotasoasteikkoon sekä aika-, kustannus- ja henkilöstöresursseihin. Lisäksi puheen ymmärtämisen menetelmä haluttiin pitää mahdollisimman samankaltaisena kuin tekstin ymmärtämisen, jotta voitiin hyödyntää standardointiin osallistuvien asiantuntijoiden aiempaa kokemusta taitotasojen asettamisesta.

Standardoinnin tuloksiin vaikuttavat keskeisesti menetelmästä riippumatta siihen osallistuvat asiantuntijat ja se, kuinka heidät on perehdytetty koko prosessiin (Cizek & Bunch 2007; Council of Europe 2009). Perehdytyksen tavoitteena on, että asiantuntijat ymmärtävät mm. standardoinnin tavoitteet, keskeiset määritelmät ja taitotasokuvaimet ja että he osaavat käyttää valittua menetelmää prosessin eri vaiheissa. Yleisissä kielitutkinnoissa asiantuntijoiden perehdytys taitotasojen asettamiseen alkoi puheen ymmärtämisen 6-portaiseen taitotasokuvaimiin liittyvällä etätehtävällä jo ennen varsinaista standardointitilaisuutta. Etätehtävän avulla saatiin tietoa asiantuntijoiden taitotasokuvaimiin liittyvistä näkemys- ja tulkintaeroista, joita voitiin hyödyntää varsinaisena standardointipäivänä annettavassa koulutuksessa. Koska asiantuntijat

tunsivat jo entuudestaan tutkintojärjestelmän, perehdytystä tutkintojärjestelmään ei tarvittu, vaan koulutuksessa voitiin keskittyä taitotasokuvainten sisältöihin. Koulutuksen tavoitteena oli, että asiantuntijoille syntyisi mahdollisimman yhdenmukainen kuva eri taitotasoilla vaadittavasta osaamisesta ja että he tulkitsisivat taitotasokuvaimia samankaltaisesti.

Taitotasoille asettaminen on monivaiheinen prosessi, joka koostuu useista eri vaiheista (ks. kuvio 1). Taitotasokuvaimien koulutuksen jälkeen alkoi varsinainen työskentely tehtävien parissa (kuvio 1, vaihe 1), joka aloitettiin suorittamalla tehtävät samoin kuin osallistujat testitilanteessa. Tällä pyrittiin siihen, että asiantuntijat pystyivät ottamaan huomioon tallenteelta kuuluvien puhujien vaikutuksen tehtävien vaikeustason määrittelyssä. Tämä myös auttoi asiantuntijoita asettautumaan testattavan rooliin, ja samalla he tutustuivat tehtäviin. Tehdessään tehtäviä he samanaikaisesti lisäksi arvioivat itsenäisesti ensimmäisen kerran tehtäväosioissa vaadittavan osaamisen tason. Nämä arviot perustuivat asiantuntijoiden omiin käsityksiin ja tulkintoihin taitotasoista.



KUVIO 1. Yleisten kielitutkintojen standardointiprosessin vaiheet.

Standardoitavien tehtäväosioiden toimivuudesta ja vaikeustasosta on tutkintojärjestelmässä saatavilla laajasti tilastotietoa, joka perustuu tehtäväosioista riippuen tuhansien osallistujien vastauksiin (vrt. Yleisten kielitutkintojen osiopankki, s. 72). Tämä tieto haluttiin tarjota päätöksenteon tueksi, ja vaiheessa 2 (kuvio 1, vaihe 2) asiantuntijoita pyydettiinkin huomioimaan taitotason määrittelyssä kuvainten lisäksi osallistujien suorituksiin perustuva tilastotieto tehtäväosioiden vaikeudesta. Asiantuntijoille jaettiin tehtävät uudestaan siten, että tehtäväosiot oli laitettu vihkoon niiden vaikeustason mukaiseen järjestykseen helpoimmasta vaikeimpaan. Asiantuntijoiden tehtävänä oli nyt uudelleen tarkastella tehtäväosioita ja arvioida, missä eri taitotasojen rajat ovat. Käytännössä asiantuntijat päättivät ensin, minkä taitotason osaaja pystyy mahdollisesti ratkaisemaan vihkon ensimmäisen eli helpoimman osion. Tämän jälkeen he etenivät osio osiolta saman taitotason puitteissa niin pitkälle, kunnes vastaan tuli tehtäväosio,

jonka ratkaisemiseksi heidän mielestään tarvittiin jo osaamista seuraavalta taitotasolta. Näin he etenivät vihkon viimeiseen eli vaikeimpaan tehtäväosioon saakka. Koko vaiheen ajan asiantuntijoilla oli mahdollisuus palata kuuntelemaan tehtäviä uudelleen, jos he kokivat sen taitotason määrittelyn kannalta tarpeelliseksi.

Ennen vaiheeseen 3 siirtymistä asiantuntijoille esitettiin vaiheen 2 tulokset. Vaiheesta 2 saatuja tuloksia pidettiin prosessin kannalta merkityksellisinä, koska asiantuntijoilla oli tässä vaiheessa käytössään tehtäväosioista paljon enemmän tietoa kuin vaiheessa 1. Lisäksi pyrittäessä yhdenmukaisuuteen asiantuntijoiden oli hyvä nähdä, miten muut olivat arvioineet ja kuinka kukin itse oli arvioinut tehtäväosion vaativuuden suhteessa muihin asiantuntijoihin. Tulosten esittämistä voidaan myös perustella sillä, että vaiheessa 3 asiantuntijaryhmän on jo päästävä koko prosessin onnistumisen kannalta hyvin samankaltaisiin tulkintoihin osioiden vaikeustasosta. Jotta tämä tavoite saavutetaan, keskityttiin ennen vaihetta 3 käydyssä keskustelussa ennen kaikkea niihin tehtäväosioihin, joissa asiantuntijoiden käsitykset tehtäväosion taitotasosta poikkesivat huomattavasti toisistaan tai tehtäväosiot olivat ratkaisevassa roolissa taitotasorajoja määriteltäessä. Tämän keskustelun pohjalta vaiheessa 3 asiantuntijat saattoivat vielä muuttaa arviointejaan lähemmäksi toisiaan (kuviossa 1 vaihe 3).

Edellä kuvatut vaiheet perustuvat pääosin Bookmark-menetelmään (ks. menetelmästä lisää: Cizek & Bunch 2007), vaikkakin joiltain osin menetelmää on muokattu Yleisten kielitutkintojen tavoitteisiin sopivammaksi. Teknisesti menetelmä ei aiheuttanut asiantuntijoille vaikeuksia, sillä se oli tekstin ymmärtämisen standardointitilaisuuksista tuttu, mutta menetelmään sopivan oppijan määrittely oli koulutuksen ja myös asiantuntijoiden näkökulmasta vaikeaa. Koulutuksessa asiantuntijoita ohjeistettiin taitotasoille määritellessään (vaiheet 1, 2 ja 3) miettimään oppijaa, joka todennäköisesti juuri ja juuri kykenee vastaamaan tehtäväosioon oikein. Koska tuloksiin vaikuttaa se, miten yhdenmukainen käsitys ja tulkinta asiantuntijoilla on tällaisen kahden taitotason rajalla olevan oppijan osaamisesta, aiheesta käytiin keskusteluja prosessin eri vaiheissa.

Asiantuntijat saivat keskustella myös tehtävistä, taitotasoista ja standardointiprosessiin liittyvistä tunnelmistaan aina jokaisen vaiheen jälkeen ennen seuraavaan vaiheeseen siirtymistä. Avoin ja salliva keskustelu on tärkeä osa standardointiprosessia, mikä edellyttää standardointiprosessin vetäjältä taitoa ohjata keskustelua. Prosessin vetäjän keskeinen tehtävä on luoda asiantuntijoiden välille luottamuksellinen ja kannustava ilmapiiri, jossa voi vapaasti ilmaista omia mielipiteitään. Vetäjän on myös huolehdittava siitä, että ryhmän sisäinen dynamiikka toimii ja että keskustelun fokus säilyy tehtävien taitotasoissa ja taitotasokuvaimissa. Keskustelulla tuetaan asiantuntijoiden arviointityötä ja pyritään siihen, että heidän käsityksensä tehtäväosioiden vaatimuksista saadaan mahdollisimman yhdenmukaiseksi.

4 Tulokset

4.1 Tilastolliset tulokset

Standardointiprosessin päätulokset ovat tehtävien osiokohtaiset taitotasot, mutta tässä artikkelissa emme raportoi niitä, vaan keskitymme tarkastelemaan asiantuntijoiden toimintaa, kuten ankaruutta, taitotasoasteikon käytön johdonmukaisuutta ja laajuutta sekä näihin liittyviä muutoksia prosessin eri vaiheissa.

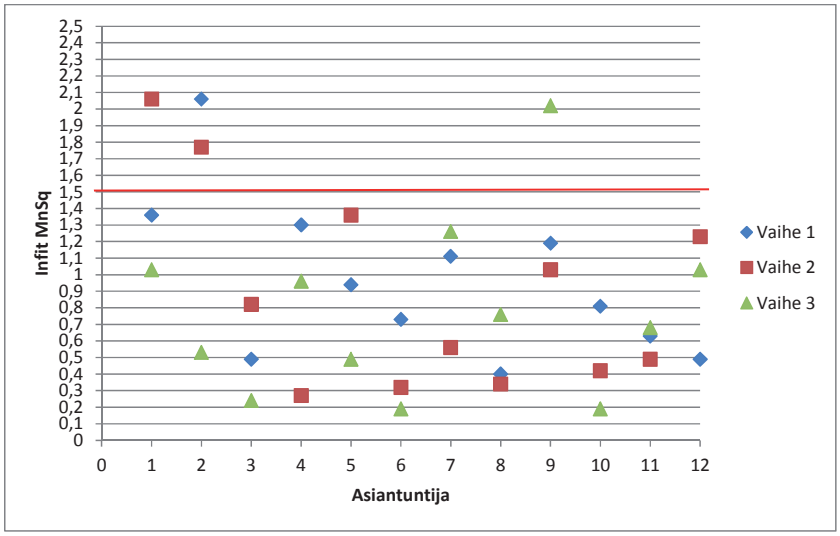
Asiantuntijoita tarkasteltaessa on kuitenkin hyvä muistaa, että tehtäväosioiden taitotasoille arviointi on aina subjektiivista toimintaa, johon vaikuttavat myös useat ulkopuoliset tekijät, kuten asiantuntijan henkilökohtaiset kokemukset ja ominaisuudet. Näiden tekijöiden vaikutus saattaa ilmetä ja vaikuttaa asiantuntijoiden näkemykseen eri tavoin, esimerkiksi ankaruutena/lempeytenä. Tästä syystä standardointiprosessin aikana asiantuntijoita johdateltiin eri tavoin kohti yhdenmukaisempaa käsitystä tehtäväosioiden vaativuudesta. Tällaisessa toiminnassa on aina vaarana se, että asiantuntijat menettävät kosketuksensa omaan arviointilinjaansa yrittäessään mukautua asiantuntijaryhmän yleiseen mielipiteeseen.

Standardointiprosessin eri vaiheista saatu aineisto analysoitiin Facets-ohjelmalla, joka perustuu IRT -menetelmiin pohjautuvaan Raschin mallin laajennukseen, MFRM (*Many-Facets Rasch Measurement*) -malliin (ks. Linacre 1989; Eckes 2011). Ensin aineisto analysoitiin 2-facetsin (ulottuvuudet: asiantuntijat ja tehtäväosiot) *rating scale* -mallilla, jonka tuloksissa kiinnitettiin erityisesti huomiota asiantuntijoiden taitotasoasteikon käytön johdonmukaisuuteen sekä siihen, oliko joukossa asiantuntijoita, jotka systemaattisesti määrittelivät tehtäväosioita korkeammalle tai matalammalle taitotasolle verrattuna muihin. Asiantuntijan toiminnan johdonmukaisuudesta tehtiin päätelmiä lähipainotetun keskineliöpoikkeaman (*Infit Mnsq*) avulla, jonka raja-arvot olivat tässä tutkimuksessa 0,5–1,5 (ks. Linacre 2002a, 2003). Asiantuntijoiden ankaruudesta kertovat arvot määriteltiin mallissa siten, että mitä suurempi arvo oli, sitä korkeammille taitotasoille asiantuntija arvioi tehtäväosioita. Koska 2-facetsin mallilla ei vielä kyetty selvittämään, vaihteliko yksittäisten asiantuntijoiden ankaruus tilastollisesti merkittävästi eri vaiheiden välillä, malli laajennettiin 3-facetsin malliksi (ulottuvuudet: asiantuntijat, tehtäväosiot ja standardointiprosessin vaiheet). Tämän pohjalta tehtiin bias-analyysi asiantuntijoiden ja prosessin eri vaiheiden välillä.

4.1.1 Asiantuntijoiden arviointilinjan johdonmukaisuus ja ankaruus

Kuviossa 2 kuvataan asiantuntijoiden arviointilinjan, eli tehtäväosioiden taitotasoille määrittelyn johdonmukaisuutta standardointiprosessin eri vaiheissa. Jokaiselle 12 asiantuntijalle on laskettu kuvioon kolme erillistä *Infit Mnsq*-arvoa. Näiden arvojen olles-

sa punaisen viivan alapuolella (eli *Infit Mnsq* <1.5) on asiantuntijan arviointilinja kaikissa prosessin eri vaiheissa johdonmukainen. Kuten kuviosta 2 on havaittavissa, asiantuntijoiden työskentely oli prosessin eri vaiheissa johdonmukaista muutamaa poikkeusta lukuun ottamatta. Kahden asiantuntijan (1 ja 2) arviointilinja ei ollut johdonmukainen vaiheessa 2 (*Infit Mnsq* >1.5), jossa heidän oli ensimmäisen kerran hyödynnettävä tilastollista tietoa tehtäväosioiden vaikeudesta taitotasokuvainten lisäksi. Asiantuntijoiden arviointilinja kuitenkin korjaantui vaiheessa 3. Lisäksi yhdellä asiantuntijalla (9) arviointilinja pysyi johdonmukaisena kahdessa ensimmäisessä vaiheessa, mutta viimeisessä vaiheessa arviointilinja katosi.



KUVIO 2. Asiantuntijoiden arviointilinjan johdonmukaisuus standardointiprosessin eri vaiheissa.

Tulokset arviointilinjasta olivat positiivisia: asiantuntijat pystyivät pitämään tehtäväosioiden taitotasoille määrittelyssä johdonmukaisuuden, vaikkakin koko prosessin ajan heidän käsityksiään ja tulkintojaan pyrittiin muokkaamaan lähemmäksi toisiaan. Kuviossa 2 esiintyvät matalat arvot (*Infit Mnsq* <0.5), erityisesti vaiheissa 2 ja 3, eivät olleet huolestuttavia, koska ne johtuivat osittain siitä, että asiantuntijat yhä enenevässä määrin antoivat yhdenmukaisia taitotasoarvioita.

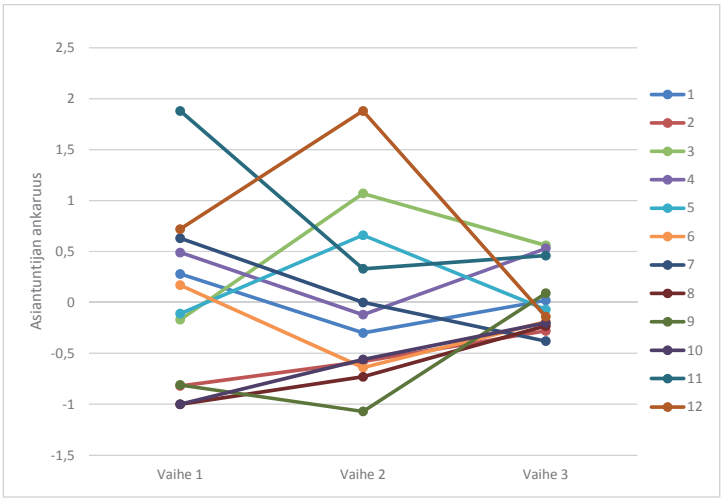
Taulukkoon 1 valittiin Facets-analyyseistä kolme havainnollistavaa tunnuslukua kuvaamaan asiantuntijaryhmän ankaruutta standardointiprosessin eri vaiheissa. Korkeat reliabiliteetin arvot eri vaiheissa kertoivat siitä, että asiantuntijat erosivat ankaruudeltaan toisistaan. Suurimmat erot ankaruusparametreissa olivat vaiheessa 2 (*R=0.95*), jolloin myös *stratan* arvo oli suurin 5.95. Sen mukaan asiantuntijat voitiin jakaa lähes kuuteen ankaruudeltaan toisistaan tilastollisesti eroavaan ryhmään. Edelleen vaiheessa

3 asiantuntijat voitiin jakaa yli kolmeen toisistaan tilastollisesti eroavaan ankaruusryhmään. Ankaruuden suhteen homogeenisen lopputuloksen saavuttamiseen ei lähtökohtaisesti uskottu, mutta keskustelujen toivottiin kuitenkin selvästi pienentävän eroja, ja näin tapahtuikin. Keskustelujen vaikutus näkyi myös siinä, että asiantuntijat olivat taitotasoille määrittelyissään vaihe vaiheelta yhdenmukaisempia, sillä yhdenmukaisuus kasvoi alun 35 %:sta 85 %:in (ks. taulukko 1, *exact agreements*).

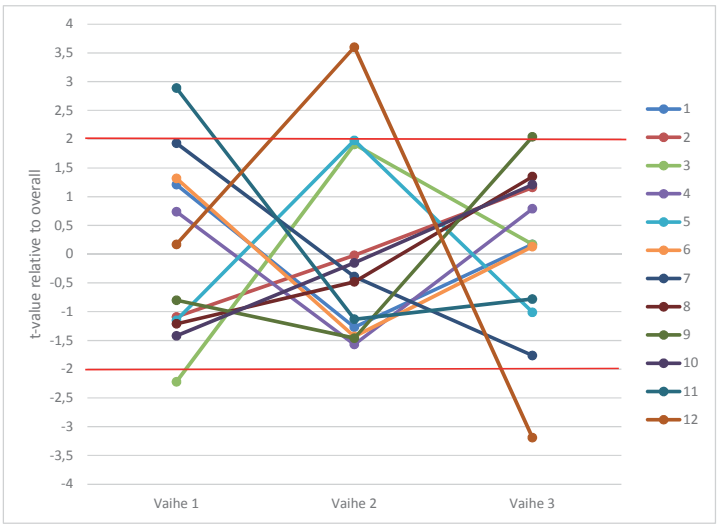
TAULUKKO 1. Asiantuntijoiden (n=12) erottelua kuvaavia yhteenvetolukuja standardointiprosessin eri vaiheissa.

	Vaihe 1	Vaihe 2	Vaihe 3
Strata	4.61	5.95	3.29
Rater separation reliability	0.91	0.95	0.83
Exact agreements	35 %	57 %	85 %

Kuvioissa 3 ja 4 asiantuntijoiden ankaruutta tarkastellaan bias-analyysin tuottamien tulosten avulla. Myös kuviosta 3 nähdään, kuinka erot asiantuntijoiden henkilökohtaisessa ankaruudessa olivat suurimmillaan vaiheessa 2 ja kaventuivat selvästi prosessin edetessä vaiheeseen 3. Kuvio 4 puolestaan havainnollistaa sitä, vaihteleeko asiantuntijoiden ankaruus tilastollisesti merkitsevästi prosessin eri vaiheissa. Kuvioon 4 on merkitty tulkintaa helpottamaan ylä- ja alarajat, joiden ulkopuolella olevat arvot ovat tilastollisesti merkitseviä merkitsevyystasolla 0.05. Kuvio 4 paljastaa, kuinka kolmen asiantuntijan ankaruus vaihtelee prosessin eri vaiheissa tilastollisesti merkitsevästi. Kahdella asiantuntijalla (3 ja 11) vaiheen 1 ankaruus eroaa vaiheista 2 ja 3 siten, että toinen heistä (3) on ollut alussa ankarampi ja toinen (11) lempeämpi verrattuna seuraaviin vaiheisiin. Yksi asiantuntija (12) on muuttanut linjaansa lempeästä ankaraksi vaiheiden 2 ja 3 aikana.



KUVIO 3. Asiantuntijoiden ankaruuserot standardointiprosessin eri vaiheissa.



KUVIO 4. Bias-kuvio asiantuntijoiden ja standardointiprosessin eri vaiheiden välillä.

Tuloksien tulkinnassa on huomioitava se, että tulosten laskennassa käytetty malli (*MFRM*) olettaa asiantuntijoiden toimivan toisistaan riippumattomasti. Vaiheessa 1 ja osittain vaiheessa 2 tämä toteutuukin, mutta ei enää vaiheessa 3. Tämä voi aiheuttaa ongelmia malliin (ks. Linacre 2002b), ja siksi varsinkin vaiheen 3 tuloksiin on suhtauduttava varoen. Tarkastelut osoittivat kuitenkin sen, että kaiken kaikkiaan asiantuntijat

määrittelivät tehtäväosioita taitotasoille johdonmukaisesti, ja standardointiprosessin edetessä asiantuntijoiden erot ankaruudessa kaventuivat.

4.1.2 Arviointiasteikon käytön laajuus

Asiantuntijat käyttivät vaiheessa 1 ja 2 taitotasoasteikkoa hyvin erilaisella laajuudella. Osa asiantuntijoista käytti arvioinneissaan koko arviointiskaalaa (1–6), kun taas toiset hyödynsivät vain kolmea taitotasoa (2–4). Yleisimmin asiantuntijat määrittelivät tehtäväosiot kuitenkin taitotasoille 2–5. Taitotasoasteikon käytön laajuus vaihteli myös prosessin edetessä. Taulukossa 2 esitellään sitä, kuinka laajasti asiantuntijat standardointiprosessin eri vaiheissa käyttivät puheen ymmärtämisen 6-portaista taitotasoasteikkoa määritellessään tehtäväosioita taitotasoille. Taulukossa 2 käytetyt prosentit viittaavat taitotasoarvioiden kokonaismäärään (12 asiantuntijaa * 35 tehtäväosiota = 420). Taulukosta nähdään, kuinka ennen viimeistä vaihetta (vaihe 3) käyty keskustelu ja tilastollinen palaute sekä omien arviointien vertailu muiden arviointeihin vaikuttivat taitotasoasteikon käyttöön siten, että asiantuntijat luopuivat asteikon ääripäiden tasoista (1 ja 6).

Taulukossa esitetyt taitotasojakaumat näyttävät varsin samankaltaisilta prosessin eri vaiheissa, mutta todellisuudessa asiantuntijat vaihtoivat käsityksiään tehtäväosioiden vaikeustasosta, etenkin vaiheiden 1 ja 2 välillä. Tämä tarkoittaa sitä, että esimerkiksi asiantuntija, joka oli määritellyt vaiheessa 1 tietyn tehtäväosion taitotasolle 3, määritteli sen vaiheessa 2 taitotasolle 4, ja vastaavasti toinen asiantuntija toimi juuri päinvastoin.

TAULUKKO 2. Asiantuntijoiden taitotasoasteikon käytön laajuus.

Taitotaso	Vaihe 1 (n=420)	Vaihe 2 (n=420)	Vaihe 3 (n=420)
1	2 %	2 %	0 %
2	11 %	15 %	19 %
3	40 %	37 %	34 %
4	33 %	35 %	32 %
5	12 %	9 %	15 %
6	2 %	2 %	0 %

4.1.3 Tehtäväosiokohtaiset tulokset

Standardointiprosessin päämääränä oli määritellä tehtäväosioille mahdollisimman luotettavana pidettävät taitotasot ja löytää näin Yleisten kielitutkintojen osiopankkiin taitotasorajat. Tehtäväosioille määritellyt taitotasot analysoitiin vaiheittain sekä käytetyn *standard setting*-menetelmän mukaisesti että *MRFM*-malleilla, jonka jälkeen tuloksia

vertailtiin keskenään. Tässä artikkelissa näitä tuloksia ei esitellä, vaan kuten jo aiemmin mainitsimme, artikkelin painopiste on asiantuntijoiden toiminta standardointiprosessin aikana. Asiantuntijoiden onnistumista standardoinnissa tarkastellaan kuitenkin sen kautta, miten yhdenmukaisia tehtäväosioiden tasoarviot lopulta olivat, joten näitä tuloksia on myös tarpeellista kuvata lyhyesti.

Ensimmäisessä vaiheessa asiantuntijoiden arviot tehtäväosioiden vaativuudesta jakautuivat lähes poikkeuksetta kolmelle tai neljälle peräkkäiselle taitotasolle. Toisessa vaiheessa jakaumat kaventuivat kahteen tai kolmeen peräkkäiseen taitotasoon, mutta ne keskittyivät jo selkeästi tietyille taitotasoille. Joukossa oli lisäksi tehtäväosioita (28 %), joiden vaikeustasosta asiantuntijoiden mielipiteet jakautuivat varsin tasaisesti kahdelle eri taitotasolle. Vaiheessa 3, viimeisen keskustelun jälkeen, asiantuntijat olivat yksimielisiä 19 osion (54 %) taitotasosta, ja loput 16 tehtäväosiota (46 %) sijoitettiin edelleen kahdelle peräkkäiselle taitotasolle. Näissäkin tapauksissa – kahta osiota lukuunottamatta – selkeä enemmistö asiantuntijoista sijoitti tehtäväosiot kuitenkin samalle taitotasolle.

4.2 Asiantuntijoiden kommentteja prosessista

Kun varsinainen tehtäväosioiden standardointi oli ohi, asiantuntijoilta pyydettiin vielä kirjallista palautetta standardointiprosessista. Koska suurimmalla osalla asiantuntijoista oli kokemusta ennestään tekstin ymmärtämisen taitotasoille asettamisesta, he varsin usein vastauksissaan vertasivat näiden kahden taidon standardointiprosessia keskenään.

(1) Kuullun ymmärtämisen osiossa oli vaikeampi asettua osallistujan nahkoihin. Tekstin ymmärtämisessä – ehkä harjoittelun vuoksi – jo vähän helpompaa. Mutta myös taitoina erilaiset. (Elli)

(2) Luetun ymmärtäminen tuntui helpommalta kuin kuullun ymmärtäminen. (Aliisa)

Puheen ymmärtämisen taitotasoille asettamista pidettiin vaikeana sen vuoksi, että sekä itse puheen ymmärtäminen että standardointi ovat prosesseina monitahoisia luonteeltaan. Puheen ymmärtäminen taitona vaati asiantuntijoista monen asian hallintaa yhtäaikaisesti, mutta myös taitotasoille asettaminen edellytti usean eri asian ja dokumentin hallintaa.

(3) Vaikeaa. Liikkuvia palikoita on liian paljon. (Sirkka)

(4) On niin monta asiaa mitkä pitää ottaa huomioon: kuvaimet, testin taitotaso, kysymysten taitotaso ja arviointikriteerit. (Erja)

(5) Hankaluutta tietää, mihin oikeastaan tulisi nojata: taitotasokuvaimiin, prosentteihin/tuloksiin vai kokonaisvaltaisempaan tulkintaan tekstistä ja tehtävistä. En oikein osannut ajatella, mikä on tärkein kriteeri määrittelyssä. Minusta ymmärtäminen on niin monitahoinen taito, että tällä tavalla tuntui vaikealta päästä sen ytimeen. (Jaana)

Jotta tehtäväosiot voidaan asettaa taitotasoille, täytyy asiantuntijoiden tuntea asettamisen apuna käytettävät kuvaimet hyvin. Vaikka Yleisten kielitutkintojen asiantuntijoilla oli paljon kokemusta opetustyönsä ja Yleisten kielitutkintojen kautta erilaisten kuvainten käytöstä, he kokivat puheen ymmärtämisen kuvainten käytön vaikeaksi taitotasojen asettamisessa. Vaikeudet liittyivät siihen, että heistä kuvaimissa puheen ymmärtämisen taito kuvattiin varsin yleisellä tasolla, ja toisaalta taas kuunneltavan tekstin ja siihen kohdistuvien osioiden vaikeustasot usein olivat varsin erilaiset, mikä vaikeutti niiden taitotason määrittelyä tietylle tasolle.

(6) Meni aika intuitiivisesti, monessa kohdin asteikosta ei paljon apua. (Jaana)

(7) Epävarmuutta aiheuttaa se, että tekstin ja kysymyksen vaikeustaso voivat olla hyvin erilaiset. (Kaisa)

Vaikka määrittelyä ohjaavat kuvaimet, taitotasoille asettamisen taustalla vaikuttivat usein myös asiantuntijan henkilökohtaiset ominaisuudet ja kokemukset. Ennen kaikkea omia kokemuksia oppijoista käytettiin hyväksi prosessin aikana, sillä pääosa asiantuntijoista työskentelee päivätyössään testiin osallistujien kaltaisten oppijoiden kanssa.

(8) Vaikea päättää, mitä linjaa noudattaa: millaisen testattavan kuvittelee tehtävää tekemään. (Aulikki)

(9) Monta muuttujaa – teksti, tehtävä, oma kokemus. (Minttu)

Kokemus kielenoppijoista oli myös muokannut asiantuntijoiden käsitystä siitä, mitä tietyllä taitotasolla oleva suomen kielen oppija osaa ja mitä hän ei osaa. Tämä tuli konkreettisesti esiin vaiheen 2 aikana (ks. kuvio 1), jolloin asiantuntijat saivat tietää tehtäväosioiden vaikeusjärjestyksen, joka oli määritelty testiin osallistuneiden kielenoppijoiden vastausten perusteella. Muutamien tehtäväosioiden vaikeustaso yllätti asiantuntijat, sillä he olivat arvioineet tehtävän sen tilastolliseen tietoon verrattuna helpommaksi tai vaikeammaksi. Tilastollinen tieto osioiden vaikeustasosta sai asiantuntijat pohtimaan omaa näkemystään tehtävän vaikeustasosta ja sitä, miksi tehtävä oli ollut vastoin heidän odotuksiaan helppo/vaikea osallistujille.

(10) Opiskelijoiden tulokset yllättivät. (Kaisa)

(11) Vaikeaa on löytää katkaisukohdat, koska tehtävät ovat omien arviointien mukaan ”väärässä” järjestyksessä. (Elli)

(12) Vaikka tuntui helpolta, lopputulos oli usein erilainen kuin mitä tilastotieto kertoo. Olo oli jonkin verran ristiriitainen, mutta tuli myös oivalluksia, miksi joku tietty osio olikin lopuksi helpompi tai vaikeampi kuin kuvittelin. (Aliisa)

(13) Edellistä epävarmemmalta. Onnistumisprosentit sekoittivat pään lopullisesti. (Sirkka)

Vaikka tilastotiedot tehtäväosioista eivät välttämättä aina vastanneet asiantuntijan omaa käsitystä osion tilastollisesta vaikeustasosta, osioista saatu tilastotieto kuitenkin lähes kaikkien asiantuntijoiden mielestä helpotti taitotason määrittelyä.

(14) Vaikutti paljonkin. Muutti käsitystä joistakin tehtävistä. (Jaana)

(15) Ohjasi epävarmoissa kohdissa. Antoi ymmärrystä, mikä osio on osoittautunut vaikeaksi itse testitilanteessa. (Ronja)

(16) Helpotti ehkä hieman päätösten tekemistä. (Tea)

(17) Helpottavasti, oli helppo tehdä päätös. (Veera)

(18) Vaikutti yhtenäistävästi ja pohdintaa tarkentavasti. (Aulikki)

Muutamasta asiantuntijasta tilastotieto kuitenkin ohjasi ehkä liikaakin toimintaa, sillä oli vaikea asettua tilastotietoja vastaan, jos heidän oma käsityksensä tehtäväosion vaikeustasosta oli alun perin ollut toisenlainen.

(19) Jos kerran 4000 testattavaa on jotain mieltä, pakkohan sitä on uskoa. (Aliisa)

(20) Asettamisesta tuli ”mekaanisempaa” sikäli, että katsoin vaikeustasoarvioita ja vastausprosentteja enkä pohtinut vain tehtävää ja sen helppoutta/vaikeutta. (Elli)

(21) Oli helpompi siinä mielessä, että järjestys oli tiedossa, tarvitsi vain miettiä, kummalle puolelle rajatapaukset menevät. (Erja)

Taitotasojen asettamisessa keskustelu sekä kuvaimista että tehtäväosioiden vaikeustasoista kuuluvat osioiden tilastollisen tiedon lisäksi keskeisesti standardointiprosessiin. Keskustelujen avulla pyritään siihen, että asiantuntijoilla on samankaltainen käsitys taidon kehittymisestä ja että he tulkitsevat käytettävissä olevat taitotasokuvaimet samalla tapaa. Asiantuntijoista keskustelut olivat hyödyllisiä ja ne myös helpottivat osioiden asettamista taitotasoille.

(22) Keskustelu antoi näkökulmia, jotka ohjasivat arvioimaan. Hyvä, että on tilaisuus pohdiskella rajoja eri taitotasojen välillä ja muuttaa perustellun keskustelun jälkeen. (Ronja)

(23) Kompromisseja joutui tekemään sekä muiden vastausten että keskustelun perusteella. (Erja)

(24) Toisten analyyttinen argumentointi vakuutti. (Minttu)

(25) On hyvä, että pyritään yhdenmukaiseen lopputulokseen ja vaihdetaan näkemyksiä. (Kaisa)

Keskustelu yhtenäisti asiantuntijoiden näkemyksiä osioiden taitotasosta, mutta heitä hämmästyttivät vaikeustasoihin liittyvät näkemyserot. Suurimmassa osassa tehtäväosioita asiantuntijoiden mielipiteet tehtävien vaikeustasoista olivat samankaltaisia, mutta muutamissa osioissa taitotasojakauma oli varsin laaja.

(26) Hajonta kummastutti. (Veera)

(27) Kuinka paljon variaatiota arvioijien välillä sallitaan? (Aulikki)

Vaiheessa 3 (ks. kuvio 1) vaikeustason mukaan järjestetyille osioille täytyy määritellä katkaisukohdat, joissa raja taitotasojen välillä sijaitsee. Osalle yhteisen päätöksen hyväksyminen taitotasorajoista oli helppoa, eikä heillä ollut keskustelun jälkeen vaikeuksia taipua toisten kanssa yhteiseen päätökseen.

(28) Oli helppoa mukautua yhteisiin taitotasorajoihin, ei mitään ongelmia. (Aliisa)

(29) ”Niska” taipuu jo tässä vaiheessa. (Sirkka)

(30) Olin jo edellisillä kerroilla oppinut, ettei muutos ole pahasta☺. (Tea)

Osalle asiantuntijoista taas oman taitotasorajan muuttaminen yhtäläiseksi muiden kanssa herätti ristiriitaisia tuntemuksia. He eivät joko hyväksyneet yhteistä päätöstä tai he eivät kaikilta osin olleet rajasta samaa mieltä, mutta muuttivat sen kuitenkin yleisen keskustelun perusteella.

(31) Osa muutoksista oli vähän vastentahtoisia, joistakin olin samaa mieltä. Kaikkia en kuitenkaan muuttanut. (Elli)

(32) Jokin raja-arvo pysyi aika hyvin kohdillaan, toista rajaa muutin toisten mielipiteiden vaikutuksesta. Jäi lopulta kuitenkin hieman epäselväksi, miten raja-arvoja on mielekästä asettaa. (Jaana)

Asiantuntijoiden antamat palautteet vahvistavat useita prosessin aikana saatuja tuloksia, mutta arvioitaessa tehtäviä ja niihin liittyviä osioita ei voida koskaan poissulkea tehtävien ja osioiden sisältöihin liittyviä näkemyksiä eikä niiden vaikutusta vaikeustason määrittelyyn. Vaikka puheen ymmärtämisen tehtävävalikoima koostui tilastollisesti hyvin toimivista ja osallistujia hyvin erottelevista osiosta, asiantuntijat saattoivat nähdä niissä ongelmia, jotka myös vaikuttivat vaikeustason määrittelyyn. Keskustelujen ja palautteen perusteella oli huomattavissa, että asiantuntijat tulkitsevat sisällöllisestikin osiot varsin monella tapaa. Vaikka keskusteluissa tulee keskittyä taitotasoihin, ei tehtävien sisältöön liittyviä kysymyksiä voi täysin ohittaa, koska niillä on myös vaikutusta asiantuntijan arvioihin.

5 Pohdinta

Vaikka taitotasolle määrittelyä ohjaavat taitotasokuvaimet, prosessin taustalla vaikuttivat tutkimuksemme mukaan myös asiantuntijan henkilökohtaiset ominaisuudet ja kokemukset. Henkilökohtaiset erot ja erilaiset tulkinnat käytettävistä arviointikuvaimista näkyivät ennen kaikkea vaiheessa 1, jossa asiantuntijoiden näkemykset tehtäväosioiden vaativuudesta vaihtelivat välillä voimakkaastikin. Tämä näyttäytyi myös asiantuntijoiden välisinä eroina ankaruudessa. Tästä syystä tämän vaiheen tulosten perusteella ei tehtäväosiolle voitu määritellä luotettavia taitotasoarvioita. Syynä suureen vaihteluun asiantuntijoiden välillä oli todennäköisesti se, että määrittelyä ohjasivat heidän omat tulkintansa taitotasoista ja että he ajattelivat vaikeustasoa määritellessään erilaisia kielenoppijoita (*borderline examinee*). Vaiheen merkitystä voidaan kuitenkin perustella sillä, että vaihe konkretisoi hyvin taitotasoihin liittyviä näkemyseroja ja että keskustelu kuvaimista ja osioista on tarpeellinen, jotta voidaan päästä lähemmäksi samankaltaista tulkintaa osioiden vaativuustasoista.

Vaiheessa 2 asiantuntijoiden tulkinnat tehtäväosioiden vaikeustasoista vaihtelivat vielä suuresti, vaikka asiantuntijoilla oli käytössään osioihin liittyvät tilastotiedot. Eroja voidaan selittää sillä, että tilastotietojen perusteella useat asiantuntijat vaihtoivat käsitystään tehtäväosioiden vaikeustasosta. Lisäksi käytössä ollut menetelmä ohjasi heitä etsimään selkeät katkaisukohdat, joissa peräkkäisten tehtäväosioiden vaikeustaso oli ratkaisevasti erilainen. Tämä aiheutti muutamille asiantuntijoille ongelmia, koska heistä osa tehtäväosioista jäi menetelmän vuoksi väärälle taitotasolle katkaisukohdan takia.

Vaiheessa 3 asiantuntijoiden mielipiteissä tehtäväosioiden vaativuudesta oli vielä eroja, mutta he kuitenkin olivat lähentyneet toisiaan vaihe vaiheelta. Keskusteluilla ja tilastollisella palautteella vaiheesta 2 oli selkeästi merkitystä erojen pienenemiseen, sillä ne herättivät asiantuntijoita pohtimaan omaa näkemystään ja vertaamaan

sitä toisten asiantuntijoiden näkemyksiin. Skorupksi (2012) ja Reckase ja Chen (2012) ovat myös havainneet, että mitä enemmän asiantuntijat keskustelevat, sitä lähemmäksi heidän näkemyksensä taitotasoista tulevat. Keskustelujen merkitystä arvioitaessa on kuitenkin hyvä muistaa, että keskusteluihin liittyvät aina ryhmädynamiikkaan kuuluvat piirteet. Aktiivisilla ja voimakkailla persoonilla on vaikutusta standardointiprosessissa (esim. Skorupksi 2012; McGinty 2005), ja tämän havaitsivat myös tutkimukseemme osallistuneet asiantuntijat. Osa asiantuntijoista koki taipuvansa keskustelun perusteella muuttamaan näkemystään, mutta toki joukossa oli myös niitä, jotka halusivat pitää oman käsityksensä vaikeustasosta keskustelujen jälkeenkin.

Keskustelua ratkaisevampi merkitys asiantuntijoiden ankaruuserojen kaventumiseen oli osioihin liittyvällä tilastollisella tiedolla (ks. myös Reckase & Chen 2012). Tilastolliseen tietoon suhtauduttiin kuitenkin kaksijakoisesti: se helpotti päätöksentekoa, mutta se myös aiheutti ristiriitaisia tuntemuksia, koska oma näkemys vaikeustasosta ei aina vastannut tilastollisen tiedon antamaa informaatiota. Tulosten näkökulmasta tilastollinen tieto näyttäytyi positiivisena tekijänä, mutta palautteen ja keskustelujen perusteella sen arvo oli ajateltua suurempi. Tilastotiedon saamisen jälkeen osa arvioijista teki selkeästi päätökset pelkästään sen perusteella. Tämä todennäköisesti vaikutti siihen, että taitotasokuvaimet jäivät taitotason määrittelyssä sivurooliin. Tilastotietoon nojautuminen tuntui asiantuntijoista turvallisemmalta, sillä useat heistä kokivat taitotasokuvainten käytön vaikeana, koska taidon kuvaus niissä oli hyvin yleisellä tasolla.

Tarkasteltaessa asiantuntijoiden johdonmukaisuutta ja ankaruutta voidaan todeta standardointiprosessin onnistuneen, vaikka asiantuntijoiden välille jäi ankaruuseroja vielä viimeisenkin vaiheen jälkeen. Ankaruus on kuitenkin pysyvä ominaisuus, ja tästä syystä täydellistä yhdenmukaisuutta ei voida saavuttaa. Hälyttävää sen sijaan oli se, että muutaman asiantuntijan ankaruus vaihteli voimakkaasti prosessin aikana. Todennäköisesti taustalla olivat epävarmuus ja taipumus muuttaa helposti omaa näkemystään eri syiden vuoksi. Vaihtelu toi kuitenkin mukanaan sen kysymyksen, miten vastaisuudessa voimme saada paremmin tietoa siitä, mitkä asiat vaikuttavat asiantuntijan päätöksentekoon kussakin vaiheessa ja miten.

Tutkintojärjestelmän näkökulmasta puheen ymmärtämisen standardointikokeilu oli hyvä alku tulevia käytänteitä ajatellen. Itse prosessi oli selkeä, ja tulokset paranivat vaihe vaiheelta. Muutamat asiat kuitenkin vaativat vielä kehittämistä ennen seuraavaa standardointikertaa.

Ratkaistavat asiat liittyvät ensinnäkin kuvainten yksityiskohtaisuuteen. Tutkimuksemme asiantuntijoiden näkemyksiä kuvainten puutteista ja ongelmista vahvistavat myös aikaisemmat tutkimukset (esim. Papageorgiou 2009; Brunfaut & Harding 2014; Moe 2009; Figueras, Kaftandjieva & Takala 2013). Näissä kritiikki kohdistui mm. siihen, että viitekehykseen pohjautuvat kuvaimet sijoittuvat jokapäiväisen elämän kontekstiin, mutta testitilanne ei ole normaali jokapäiväisen elämän tilanne, ja tästä syystä kuvai-

met eivät välttämättä ole kaikilta osin käyttökelpoisia testitilanteessa. Lisäksi Papageorgioun (2009) tutkimukseen osallistuneet asiantuntijat näkivät kuvaimissa muitakin puutteita, jotka liittyivät kuvainten sanavalintoihin ja sisältöihin. Standardointiprosessin onnistumisen kannalta kuvainten on kuitenkin tärkeä olla yksityiskohtaisuudeltaan sellaiset, että asiantuntijat voivat niihin tukeutua myös tilanteissa, joissa tekstin ja osioiden vaikeustasot eroavat toisistaan paljon tai joissa osion vaikeustaso sijoittuu heistä kahden taitotason rajalle (*borderline case*).

Toinen kehittämistä edellyttävä asia on kahden taitotason välillä olevan ns. borderline-suorittajan (*borderline examinee*) taidon tarkempi määrittely. Ongelmat tässä liittyivät ennen kaikkea siihen, millaista oppijaa asiantuntijat ajattelivat, ja sitä kautta luonnollisesti siihen, kuinka he hahmottivat eri taitotasot. Kuvainten selkeyttäminen ja niiden avaaminen keskusteluissa ovat myös tämän ongelman ratkaisemiseksi keskeinen kehittämiskohde.

Suurimmat kysymykset kuitenkin liittyvät eri vaiheiden tulosten käyttöarvoon ja menetelmän valintaan. Aikaisemmat tutkimukset osoittavat, että menetelmän valinnalla on vaikutusta tuloksiin, ja tässä standardointikokeilussa tämä myös tuli selkeästi ilmi. Vaiheen 1 ja 3 tulokset eivät tukeneet toisiaan vaan erosivat toisistaan merkittävästi. Vaiheessa 1 käytetyn menetelmän tuloksissa oli niin paljon hajontaa, että niiden pohjalta ratkaisuja oli mahdotonta tehdä, joten vaihtoehdoksi jäi määritellä taitotasorajat osiopankin vaikeustasoskaalalle vain vaiheen 3 tulosten pohjalta. Pohdittavaksi jääkin, mikä merkitys vaiheella 1 on tulevaisuudessa. Tarvitaanko vaihetta vain konkretisoimaan asiantuntijoille heidän välisiä näkemyserojaan ja tutustuttaamaan heidät tehtäviin ja taitotasokuvaimiin, vai voidaanko vaiheesta muokata vertailukelpoinen menetelmä vaiheiden 2 ja 3 rinnalle? Jatkossa on myös mietittävä, miten varmistetaan että asiantuntijat ovat ymmärtäneet standardointiprosessin tehtävänannon.

Standardointikokeilu herätti lisäksi kysymyksiä siitä, parantaako vai heikentääkö uuden menetelmän lisääminen prosessiin tulosten luotettavuutta. Useiden eri menetelmien käyttö standardointiprosessin aikana voi tehdä asiantuntijoiden näkökulmasta prosessista liian raskaan ja monimutkaisen. Koska tässä kokeilussa oli myös paljon hyviäkin puolia, kokeilussa käytettyjä menetelmiä voidaan mahdollisesti vielä soveltaa ja validoida siten, että asiantuntijat jaetaan kahteen eri ryhmään ja näistä kahdesta ryhmästä saatuja tuloksia verrataan toisiinsa. Brunfaut ja Harding (2014) kokeilivat vastaavanlaista lähestymistapaa kahdella eri asiantuntijaryhmällä linkittäessään GEPT:n (General English Proficiency Test) puheen ymmärtämisen testiä Eurooppalaiseen viitekehykseen, ja he saivat kokeilusta rohkaisevia tuloksia.

Kirjallisuus

Bejar, I. 2008. Standard setting: What is it? Why is it important? *R & D Connections 7*. Princeton: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RD_Connections7.pdf.

Brunfaut, T. & L. Harding 2014. Linking the GEPT listening test to the Common European Framework of Reference. *LTTC-GEPT Research Reports RG-05*. Taipei: The Language Training and Testing Center. https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG05.pdf.

Buck, G. 2001. *Assessing Listening*. Cambridge: Cambridge University Press.

Cizek, G. & M. Bunch 2007. *Standard Setting. A guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, Cal: Sage.

Council of Europe 2009. *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR). A Manual.* Strasbourg: Language Policy Division. http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf.

Eckes, T. 2011. *Introduction to Many-Facet Rasch Measurement. Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Figueras, N., F. Kaftandjieva & S. Takala 2013. Relating a reading comprehension test to the CEFR levels: a case of standard setting in practice with focus on judges and items. *The Canadian Modern Language Review, 69* (4), 359–385.

Kaftandjieva, F. 2010. *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL*. EALTA. http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf.

Linacre, J. 1989. *Many-facet Rasch measurement*. Chicago, Ill: MESA Press.

Linacre, J. 2002a. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16* (2), 878. https://www.rasch.org/rmt/rmt162f.htm.

Linacre, J. 2002b. Judge ratings with forced agreement. *Rasch Measurement Transactions, 16* (1), 857–8. http://www.rasch.org/rmt/rmt161.pdf.

Linacre, J. 2003. Size vs. significance: infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions, 17* (1), 918. https://www.rasch.org/rmt/rmt171n.htm

McGinty, D. 2005. Illuminating the "black box" of standard setting: an exploratory qualitative study. *Applied Measurement in Education, 18* (3) 269–287.

Messick, S. 1989. Validity. Teoksessa R. Linn (toim.) *Educational measurement* (3. painos). Phoenix, Az: Oryx, 13–104.

Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), 13–23. http://journals.sagepub.com/doi/pdf/10.3102/0013189X023002013.

Moe, E. 2009. Jack of more trades? Could standard setting serve several functions? Teoksessa N. Figueras & J. Noijons (toim.) *Linking to the CEFR levels: research perspectives*. Cito, Institute for Educational Measurement, Council of Europe, European Association for Language Testing and Assessment (EALTA), 131–137.

Neittaanmäki, R. & T. Hirvelä 2014. Yleisten kielitutkintojen osallistujat taustatietojen valossa. Teoksessa T. Leblay, T. Lammervo & M. Tarnanen (toim.) *Yleiset kielitutkinnot 20 vuotta.* Raportit ja selvitykset 2014: 16. Helsinki: Opetushallitus, 46–60.

Papageorgiou, S. 2009. Analyzing the decision-making process of standard setting participants. Teoksessa N. Figueras & J. Noijons (toim.) *Linking to the CEFR levels: research perspectives.*

Cito, Institute for Educational Measurement, Council of Europe, European Association for Language Testing and Assessment (EALTA), 75–78.

Reckase, M. 2009. Standard setting theory and practice: issues and difficulties. Teoksessa N. Figueras & J. Noijons (toim.) *Linking to the CEFR levels: research perspectives*. Cito, Institute for Educational Measurement, Council of Europe, European Association for Language Testing and Assessment (EALTA), 13–20.

Reckase, M. & J. Chen 2012. The role, format, and impact of feedback of standard setting panelists. Teoksessa G. Cizek (toim.) *Setting performance standards: foundations, methods, and innovations* (2. painos). New York: Routledge, 149–164.

Rost, M. 1990. *Listening in language learning, applied linguistics and language study*. Singapore: Longman.

Saarela-Kinnunen, M. & J. Eskola 2001. Tapaus ja tutkimus = Tapaustutkimus. Teoksessa J. Aaltola & R. Valli (toim.) *Ikkunoita tutkimusmetodeihin I*. Jyväskylä: Gummerus, 158–169

Skorupski, W. 2012. Understanding the cognitive processes of standard setting panelists. Teoksessa G. Cizek (toim.) *Setting performance standards: foundations, methods, and innovations* (2. painos). New York: Routledge, 135–147.

Tuomi, J. & A. Sarajärvi 2009. *Laadullinen tutkimus ja sisällönanalyysi.* Helsinki: Tammi.

Törmäkangas, K. & T. Törmäkangas 2009. *Osioanalyysi testien arvioinnissa.* Jyväskylä: Jyväskylän yliopisto, Koulutuksen tutkimuslaitos.

Weir, C. 2005. *Language testing and validation. An evidence-based approach*. Lontoo: Palgrave Macmillan.

Wright, B. & M. Stone 1979. *Best test design*. Chicago, Ill: MESA Press.