VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Arvioinnin epävarmuus ja sen vaikutukset arvioinnin laatuun suomen kielen puhumisen taidon testiarvioinnissa

- Mari Honko, Soveltavan kielentutkimuksen keskus, Jyväskylän yliopisto,
- Reeta Neittaanmäki, Soveltavan kielentutkimuksen keskus, Jyväskylän yliopisto

Kirjoittajien yhteystiedot:
Mari Honko mari.h.honko@jyu.fi
Reeta Neittaanmäki reeta.neittaanmaki@jyu.fi

Tutkimuksessa tarkastellaan koulutettujen kielitaidon arvioijien arviointivarmuutta ja sen vaikutusta arvioinnin laatuun aikuisten suomenoppijoiden puhumisen taidon testiarvioinnissa. Aineisto koostuu 12 059 tehtäväkohtaisesta puhumisen taitotasoarviosta ja arvioinnin aikana tehdyistä muistiinpanoista sekä kyselytiedosta. Analyysimetodeina käytetään laadullista sisällönanalyysia ja sisällön erittelyä sekä tilastollisia menetelmiä. Tulosten perusteella arvioijat olivat pääsääntöisesti varmoja antamistaan puhumisen taitotasoarvioista mutta jäivät silti ainakin satunnaisesti epävarmaksi arvioistaan. Kun epävarmuutta esiintyi, se kohdistui tyypillisesti tietyn suorittajan tehtäväkokonaisuudessa vain yhteen tehtäväkohtaiseen arvioon. Arvioijat nimesivät epävarmuudelle useita syitä. Arviointikokemuksen pituus vuosina ei ollut yhteydessä arviointivarmuuteen. Epävarmuuden määrä vaihteli arvioinnin eri vaiheissa siten, että epävarmuus oli suurinta arvioinnin alussa ja lisääntyi myös toisen arvioijan aiemmin arvioitujen suoritusten saapuessa kaksoisarviointiin. Epävarmuuden kohdentuminen oli pitkälti yksilöllistä, sillä kaksoisarvioiduissa suorituksissa arvioijien epävarmuus kohdistui pääosin eri puhujien suorituksiin. Koettu epävarmuus ei vaikuttanut arvioinnin laatuun eli sen johdonmukaisuuteen tai ankaruuteen/lempeyteen. Arvioinnin laatutekijöistä ja riittävästä arviointivarmuudesta huolehtiminen on kuitenkin tärkeää, jotta epävarmuus ei kuormita arvioijia liikaa ja jotta arvioinnin luotettavuus säilyy.

Avainsanat: epävarmuus, kielitaidon arviointi, kielitesti, puhuminen

1 JOHDANTO

Arvioijana toimiminen kielitaidon testiarviointia toteuttavissa virallisissa järjestelmissä, kuten Suomessa Yleisissä kielitutkinnoissa, Valtionhallinnon kielitutkinnoissa ja Ylioppilastutkinnoissa, on vastuullinen tehtävä. Arviointitulosta voidaan käyttää riittävän kielitaidon osoittamiseen esimerkiksi opiskelu- ja työpaikan haun yhteydessä, ja suomen ja ruotsin kielen arvioinnista saatavaa todistusta myös Suomen kansalaisuutta haettaessa. Käyttötarkoitusten vuoksi arviointituloksen oikeudenmukaisuus ja oikeellisuus on tärkeää, sillä virheellinen arviointipäätös voi pahimmassa tapauksessa estää tavoitellun opiskelu- tai työpaikan saannin tai hidastaa kansalaisuuden hakemisen prosessia. Lisäksi arvioinnin luotettavuus on tulosten käyttötarkoituksen ja osallistujien oikeudenmukaisen kohtelun vuoksi yhteiskunnallisesti merkittävää, koska se on yhteydessä vieraskielisen työllistymiseen ja osallisuuteen Suomessa. Virallisissa testijärjestelmissä arviointiin liittyykin arviointitulosten käytön sääntelyn vuoksi erityistä vaikuttavuutta.

(Lynch, 2001; Suomessa esim. Tossavainen, 2016.)

Arvioinnin tarkoituksesta, vaikuttavuudesta ja toteuttamistavasta syntyy testiarviointiin kiinnostava jännite: Arvioinnin reiluuden ja laadukkuuden varmistamiseksi arvioijan täytyy olla pätevä ja luotettava asiantuntijaroolissaan. Koska arvioija on kokemuksineen ja tunteineen inhimillinen ihminen, hän toisaalta suorittaa tehtävää, jossa on aina myös subjektiivinen ulottuvuus (Fan & Yan, 2020; Tarnanen, 2002, 2014.) Jos arvioijan toiminnalla on vaikutusta testattavan elämään, herää kysymyksiä siitä, saako arvioija tuntea epävarmuutta arviointityössään, mitä mahdollisesta epävarmuudesta seuraa ja miten arvioijan epävarmuuteen pitäisi suhtautua.

Vaikka arviointitehtävän vastuullisuus korostuu testiarviointiin keskittyvissä erillisissä järjestelmissä, samat kysymykset koskevat kielitaidon arviointia opettajan työssä eri koulutusasteilla. Opettajan rooli esimerkiksi kokeiden tai muiden oppimistehtävien arvioinnissa sekä diagnostisessa arvioinnissa ja summatiivisten päättöarviointien yhteydessä muistuttaa arvioijan roolia testiarvioinnissa. Myös näissä opettajalla on vastuu arvioinnin oikeellisuudesta ja luotettavuudesta, ja opetuksen aikainen arviointi vaikuttaa mm. oppijan minäkäsitykseen ja opiskelumotivaatioon sekä tulevaisuuden mahdollisuuksiin (mm. Huhta & Ahola, 2019; Huhta & Hildén, 2013).

Sekä ammattimaisessa kielitaidon arvioinnissa että opetukseen liittyvässä arvioinnissa on tärkeää tunnistaa erityisesti tuottamistaitojen kuten puhumisen ja kirjoittamisen inhimilliseen arviointiin liittyvät erityispiirteet sekä arvioijien erilaiset tavat tehdä arviointia, sillä näissä taidoissa yksittäisen arvioijan tulkinnalla on tuloksen määrittymisen kannalta yleensä eniten merkitystä (Fan & Yan, 2020; Huhta & Hildén, 2013). Toisin kuin ymmärtämistaitojen arvioinnissa, tuottamisen arvioinnissa ei voida käyttää esimerkiksi tiettyihin oletusvastauksiin perustuvaa pisteytystä, vaan arvioijan on tehtävä jokaisesta tehtävästä arviointipäätös ennalta määrättyjen kriteereiden perusteella. Koska arviointipäätökset ovat alttiita arvioijien subjektiiviselle näkemykselle kielitaidosta, on pyrittävä varmistumaan siitä, että arviointiin liittyvät tunteet ja arvioijien erilainen arviointikokemus, kokemukset erilaisista oppijoista sekä erilaiset arvostukset tai arviointitavat eivät vaikuta arvioinnin lopputulokseen.

Yleisesti ajatellaan, että arviointikokemuksen karttuessa arviointivarmuus kasvaa ja että kokemuksen mukana arvioija saa myös lisää erilaisia keinoja käsitellä epävarmuutta (suomen kielen taidon arvioinnista Ahola, 2016, 2022; Tarnanen, 2002). Tarnasen (2002) kirjoittamisen arviointia käsittelevän väitöstutkimuksen haastatteluaineisto toi esille sen, että epävarmuuden kokemusten työstäminen on tärkeää, vaikkeivat ne poistuisikaan kokonaan. Kielitaidon arvioijien tunteisiin ja niiden vaikutuksiin keskittyvää tutkimusta on toistaiseksi kuitenkin tehty hyvin vähän sekä Suomessa että kansainvälisesti. Aiempi haastattelututkimus (Ahola, 2016, 2022) sekä suomen kielen kirjoitustaidon arviointiin kohdistuva tutkimuksemme (Honko, Neittaanmäki, Jarvis & Huhta, 2023[1]) osoittavat, että myös virallisessa kielitaidon testiarvioinnissa esiintyy epävarmuutta, joka syntyy eri tekijöiden vaikutuksesta ja joka koskee sekä vasta aloittaneita että pitkään arvioijana toimineita koulutuksen saaneita arvioijia. Arvioinnin yhteydessä tutkittuun arviointivarmuuteen liittyvää systemaattista empiiristä tutkimusta ei kuitenkaan ole aiemmin julkaistu suomen kielen puhumisen arvioinnista, eikä tiedossamme ole myöskään vastaavaa muiden kielten arviointia koskevaa tutkimusta.

Tämä tutkimus kohdistuu puhumisen arvioijien kokemaan epävarmuuteen sekä sen syihin ja seurauksiin eurooppalaisen viitekehyksen taitotasoihin perustuvassa testiarvioinnissa Yleisissä kielitutkinnoissa. Selvitämme yhden tutkintokerran aikana kerätyn arviointi- ja kyselyaineiston avulla arvioijien kokemuksia omasta arviointivarmuudestaan, epävarmuuden ilmenemistä arvioinnin aikana sekä epävarmuuden mahdollisia vaikutuksia arvioinnin laatuun.

Tutkimuskysymyksemme ovat seuraavat:

1. Kuinka yleistä arvioinnin epävarmuus on puhumisen testiarvioinnissa ja miten se ilmenee?

1 Tutkimukset ovat osa samaa tutkimuskokonaisuutta. Kirjoittamisen osatutkimuksen tavoite ja menetelmät ovat pääosin samat kuin tässä puhumisen arvioijia koskevassa tutkimuksessa.

2. Onko epävarmuus yhteydessä arviointikokemukseen tai arvioinnin ajalliseen etenemiseen?

3. Mitä syitä arvioijat tyypillisimmin nimeävät epävarmuuden kokemuksilleen arvioinnin aikana ja toisaalta arviointitilanteen ulkopuolella?

4. Onko epävarmuus yhteydessä arvioinnin laatuun?

Ensimmäisen tutkimuskysymyksen osalta tarkastelemme aluksi sitä, kuinka varmoina arvioijat pitävät puhumisen taitotasoarviointia yleisesti. Toiseksi selvitämme, kuinka usein arvioijat jäävät antamastaan arviosta epävarmaksi puhumisen taitotasoarvioita tehdessään. Samalla selvitämme, kohdistuuko epävarmuus yksittäisiin tehtäviin vai tietyn puhujan koko suoritukseen sekä kasautuuko eri arvioijien epävarmuus samojen puhujien arviointiin. Arviointikokemusta lähestymme kolmella tavalla. Aluksi tarkastelemme arviointivarmuutta suhteessa arvioijan arviointivuosiin Yleisissä kielitutkinnoissa. Tämän jälkeen selvitämme, tapahtuuko arviointivarmuudessa muutoksia yhden arviointikerran aikana eri arviointirupeamien tai -arviointipäivien välillä. Epävarmuuden syiden tutkimisessa selvitämme, mitkä aiemmassa tutkimuksessa esille nousseista sekä arvioijien tämän tutkimuksen aikana nimeämistä tekijöistä ovat usein esiintyviä ja useiden arvioijien tunnistamia, mitkä taas näitä harvemmin esiintyviä tai huonommin tunnistettuja. Lopuksi tutkimme tilastollisten analyysien avulla, onko epävarmuuden kokemuksilla vaikutusta arvioinnin luotettavuuteen eli arvioijan arviointilinjan ankaruuteen ja johdonmukaisuuteen.

Tutkimuksen tulokset tuottavat tietoa erityisesti puhumisen osaamisperusteisesta taitotasoarvioinnista. Vaikka tutkimus keskittyy testiarviointiin, se tarjoaa myös muuhun kielitaidon arviointiin sovellettavaa tietoa esimerkiksi arviointiin liittyvän epävarmuuden syistä ja mahdollisista seurauksista. Tuloksia voidaan hyödyntää testiarvioinnin kehittämiseen testijärjestelmissä, ja lisäksi niitä voidaan hyödyntää kielikoulutukseen sisältyvän puhumisen arvioinnin tutkimisessa ja kehittämisessä.

2 ARVIOINTI PÄÄTÖKSENTEKONA

Epävarmuutta esiintyy kaikkialla arkisessa päätöksenteossa, mutta erityisesti tilanteissa, joissa on läsnä ihmisen omakohtaisia tulkintoja ja käsityksiä (Anderson ym., 2019). Näissä päätöksenteon subjektivisuus korostuu, eli päätöksen tekijän oman taustan, näkemysten ja päätäntävallan merkitys on suuri. Ihmisen suorittamaa kielitaidon arviointiakin voidaan tarkastella päätöksentekona eli tapahtumana tai prosessina, jossa arvioijan on tehtävä valinta useammasta eri vaihtoehdosta (Cumming ym., 2002; Tarnanen, 2002).

Arvioinnin subjektiivisen luonteen vuoksi tarvitaan erityisesti empiiriseen aineistoon perustuvaa arvioijakäyttäytymistä ja arvioinnin laatua koskevaa tutkimusta, jolla varmistetaan, että päätöksentekoon liittyvät kielenoppijan suorituksesta riippumattomat tekijät, kuten arvioijan tunteet, eivät vaikuta arvioinnin luotettavuuteen. Kansainväliseen aineistoon perustuva puhumisen arviointitutkimusta koskeva katsaus (Fan & Yan, 2020) osoittaa, että arvioijavaikutukset on yksi eniten tutkittuja aiheita puhumisen arvioinnissa puhumisen taidon ja siihen vaikuttavien tekijöiden ohella. Tutkimuksissa on ollut tyypillistä keskittyä esimerkiksi arvioijien liialliseen ankaruuteen/lempeyteen, sädekehävaikutukseen eli siihen, miten arvioijan vaikutelma arvioitavasta tai suorituksen tietyistä piirteistä ohjaavat arviointia, arviointiasteikon keskimmäisten arvioiden painottamiseen, samojen arvioiden painottamiseen tai arvioijien välisen yhdenmukaisuuden tai yksimielisyyden puutteeseen (ks. Ahola, 2022; Tarnanen, 2002, s. 68). Arvioijista riippumattomien syiden lisäksi arvioijavaikutuksista tiedostuminen itsessään voi aiheuttaa arvioijissa epävarmuutta (esim. Tarnanen, 2002).

Tarnasen (2002) mukaan arvioijan on tärkeä tiedostua arvioinnin inhimillisestä puolesta ja myös oppia reflektoimaan omaa arviointiaan, sillä liiallinen pelko päätöksenteon subjektiivisuudesta voi muodostua arvioinnin esteeksi ja liiallinen usko oman arvioinnin objektiivisuudesta puolestaan saattaa estää arvioinnin tarpeellisen reflektoinnin. Vaikka empiiristä tutkimusta on kielitieteen alalla toistaiseksi tehty vähän, myös kielitaidon arviointiin liittyvä epävarmuus on tunnistettu ilmiö, johon toisinaan viitataan sekä teoriakirjallisuudessa että empiirisen tutki-

muksen aineistoissa (esim. Ahola, 2016; Tarnanen, 2002, 2014; Youn, 2018). Tarnasen (2002) haastattelemat suomen kielen kirjoittamisen arvioijat kertoivat esimerkiksi arvioitavien tekstien ja puutteellisten arviointikriteerien aiheuttamista tunnereaktioista ja omaa arviointilinjaansa koskevasta epävarmuudesta. Itsereflektioon perustuvissa haastatteluissa osa Yleisten kielitutkintojen arvioijista puolestaan on todennut vaikeiden arviointien ainakin toisinaan lisäävän epävarmuutta ja kokemuksen puolestaan poistavan sitä (Ahola, 2016).

Muilla tieteenaloilla on esitetty, että asiantuntijan epävarmuus päätöksistään voi lisätä virheellisen päätöksenteon riskiä sekä aiheuttaa viivästyksiä ja resurssien tuhlaamista (lääketieteestä Bhise ym., 2018). Esille on noussut esimerkiksi huoli siitä, kuinka epävarmuus voi vaikuttaa kielteisesti arviointilinjan johdonmukaisuuteen (Youn, 2018). Arvioijan epävarmuuden vaikutuksia arviointipäätökseen on systemaattisesti kuitenkin käsitelty vain yhdessä tuntemassamme kieleen kohdistuvassa tutkimuksessa (Bosshardt ym., 2016). Tässä aihetta empiirisesti lähestyneessä tutkimuksessa tutkittiin alle kouluikäisten lasten puhumisen arviointia. Siinäkään ei tosin tutkittu kielitaidon arviointia vaan tarkasteltiin epävarmuuden ilmenemistä ja vaikutuksia änkyttämisen vaikeusasteen arvioimisen yhteydessä. Tulosten perusteella arvioijien arviointivarmuus 10-portaisella skaalalla arvioituna ei ollut yhteydessä arvioinnin laatuun.

3 PUHUMISEN TESTIARVIOINNIN ERITYISPIIRTEITÄ

Arjessa yksilön puhe kietoutuu yleensä tiiviiksi osaksi muiden ihmisten kanssa käytävää ja usein monimodaalista vuorovaikutusta. Puhumisen testiarvioinnissa tarkoitus on kuitenkin tuottaa tietoa nimenomaan yksilön puhumisen taidosta, ja siksi siinä joudutaan tasapainoilemaan toisaalta arviointiin valittujen tilanteiden ja tehtävien autenttisuuden sekä toisaalta taidon erotettavuuden ja yksilöitävyyden välillä. (Carter & McCarthy, 2017; Fan & Yan, 2020.) Puhuttu kieli on vahvasti sidoksissa puhehetkeen ja kontekstiin, jossa se on tuotettu, ja esimerkiksi testitilanteessa tuotettuun kirjoitukseen nähden arvioitava puhe on yleensä spontaanimpaa ja strukturoimattomampaa. Yleisissä kielitutkinnoissa puhuminen on arkista vuorovaikutusta kontrolloidumpaa ja puhetta ohjataan joko haastattelun tai simuloitujen tehtävien avulla. Puheeseen liittyy silti lähes aina dialogisuuden ulottuvuus (keskustelukumppanit ja heidän toimintansa), ja siihen kytkeytyy kielenulkoista viestintää sekä paralingvistisiä vihjeitä (esimerkiksi äänen voimakkuus, korkeus ja puhenopeus). Lisäksi puhe koostuu monesta eri piirteestä, kuten ääntämisestä ja fonologisesta hallinnasta, tarkkuudesta, ilmaisun laajuudesta jne. (esim. Luoma, 2004.) Nämä seikat tekevät puhumisesta arvioinnin näkökulmasta varsin moniulotteisin ja haastavan taidon. Vaikka tehtävissä pyritään simuloimaan todellisia kielenkäyttötilanteita ja tehtävien määrällä pyritään saamaan kielitaidosta mahdollisimman kattava kuva, testiarvioinnissa päätelmät kielitaidosta myös tehdään aina hyvin rajallisen näytteen perusteella.

Oman haasteensa arviointiin tuo se, että puhumisen arviointi voi henkilöityä esimerkiksi kirjoittamisen arviointia voimakkaammin, sillä puhujan ääni, puherytmi ja aksentti sekä oletettu ensikieli herättävät arvioijassa väistämättä sekä tietoisia että tiedostamattomia mielikuvia, tunteita ja tulkintoja, jotka voivat vaikuttaa arviointiin (Ahola, 2022; Halonen ym., 2020). Arvioijan osaamat kielet ja kokemus erikielisistä oppijoista sekä eritaustaisten oppijoiden arvioinnista ja sitä kautta esimerkiksi erilaisten aksenttien tuttuus voivat myös vaikuttaa siihen, miten ankara/lempeä hän on eri kieliryhmien puhumisen arvioinnissa (Ahola, 2022; Fan & Yan, 2020; Winke & Gass, 2011). Etenkin tuottamisen arvioinnissa myös puhujasta syntyvällä ensivaikutelmalla on merkitystä arviointitilanteessa (Ahola & Halonen, 2022).

Arvioijilla on erilaisten taustojensa, kokemustensa, taitojensa ja henkilökohtaisten ominaisuuksiensa vuoksi erilaisia tapoja toimia arviointitilanteessa, mikä voi ilmetä esimerkiksi eroina kielikäsityksissä ja edelleen arviointikriteerien erilaisena painottamisena (Ahola, 2022; Duijm, Schoonen & Hulstijn, 2017), tietynkielisten puhujien suosimisena arvioinnissa (Winke & Gass, 2011) sekä arvioijien välisinä ankaruuseroina (Eckes, 2005; Wind, Jones & Bergin, 2021). Samanlaisetkaan taustatekijät eivät silti välttämättä vaikuta samalla tavalla yksittäisten arvi-

oijien toimintaan, ja esimerkiksi arviointikokemuksen vaikutuksesta arvioijien ankaruuteen on saatu osin ristiriitaisia tuloksia (Davis, 2016; Fan & Yan, 2020).

Kriteerien ulkopuolisten tekijöiden vaikutukset eivät välttämättä ole lopullisen arvion kannalta yksittäiselle kielenoppijalle suuria, mutta ne tulee silti huomioida testiarvioinnin kehittämisessä ja arvioijien koulutuksessa (Fan & Yan, 2020). Arvioinnin laatua pyritäänkin testijärjestelmissä varmistamaan monin tavoin. Koska kielitaidon arvioijilla on arviointitilanteessa paljon päätäntävaltaa, monet laadun varmistamisen keinot kohdistuvat suoraan tai epäsuorasti juuri arvioijiin (esim. McNamara ym., 2019). Tyypillisesti arvioijat valikoidaan huolellisesti ja arviointia ohjataan kaikille arvioijille yhteisellä arviointiprosessilla sekä yhteisillä arviointikriteereillä ja -ohjeilla. Arvioijien kouluttamisella ja arviointiharjoituksilla puolestaan pyritään vahvistamaan kriteerien ja ohjeiden yhdenmukaista tulkintaa ja lisäämään arviointien yhdenmukaisuutta (Euroopan neuvosto, 2011; Huhta & Hildén, 2013).

Koska arvioitavan aineiston laatu ja käytössä oleva arviointitapa voivat vaikuttaa arvioinnin luotettavuuteen, ne on yhden testijärjestelmän sisällä ja yhdessä tutkimuksessa pidettävä mahdollisimman samankaltaisina kaikilla arvioijilla. Laadukkaassa testiarvioinnissa on lisäksi tarpeen seurata arviointikäyttäytymistä, jotta mahdolliset vääristymät arvioinnissa tulevat esiin. Arvioijien välistä yhdenmukaisuutta on tärkeää seurata esimerkiksi pyytämällä useampaa arvioijaa arvioimaan samat suoritukset. Tutkintojärjestelmän on myös huolehdittava henkilökohtaisen palautteen antamisesta ja palautteen vaikutusten seurannasta, jotta arvioijat voivat tarvittaessa korjata arviointilinjaansa. (Esim. Euroopan neuvosto, 2011.)

4 AINEISTO JA METODIT

4.1 Aineiston yleisesittely

Tutkimusaineisto on kerätty Yleisten kielitutkintojen arvioijilta suomen kielen keskitason tutkinnon yhden arviointikerran yhteydessä syksyllä 2021 osana tavanomaista arviointitoimintaa. Tutkimukseen osallistuminen oli arvioijille vapaaehtoista, ja tutkintokerran arvioijista 27 (73 %) osallistui tutkimukseen. Tutkimusaineisto muodostuu ennen arviointia, arviointirupeamien välissä sekä arviointien jälkeen kerätystä kyselytiedosta, arvioinnin aikana tehdyistä muistiinpanoista sekä suoritusten tehtäväkohtaisista taitotasoarvioinneista. Arvioinnit kohdistuivat yhteensä 971 henkilön puhumisen suorituksiin. Näistä 39,1 % (n = 380) kaksoisarvioitiin, mikä mahdollisti eri arvioijien arviointitulosten vertailun. Aineiston osat yhdistettiin toisiinsa arvioijakohtaisten tunnistetietojen avulla.

Arviointi Yleisissä kielitutkinnoissa perustuu Yleiseurooppalaisen viitekehyksen (Euroopan neuvosto, 2001) toiminnallisen kielitaidon taitotasoista johdettuihin osaamisperusteisiin YKI-arviointikriteereihin (OPH 2003). Puhumisen testissä arvioitavia osasuorituksia on useita, ja ne pohjautuvat keskenään erilaista kielellistä toimintaa edellyttäviin tehtävänantoihin. Tällä pyritään varmistamaan puhumisen näytteen riittävyys ja edustavuus sekä vähentämään yksittäisen tehtävän ja tilanteen painoarvoa arvioinnissa. Arvioijalla on arviointitulosten muodostumiseen keskeinen rooli: vertaamalla puhumisen suorituksia taitotasokuvauksiin (keskitasolla kriteerit taitotasoille 3 ja 4) hän arvioi, mitä taitotasoa kielenoppijan puhumisen taito vastaa. Käytössä on tehtäväkohtaiset arviointiohjeet sekä taitotasoa kuvaava yleiskriteeri ja kuusi analyyttista kriteeriä kuvauksineen (*sujuvuus*, *joustavuus*, *koherenssi/sidoksisuus*, *ilmaisun tarkkuus/laajuus/idiomaattisuus, ääntäminen / fonologian hallinta, rakenteiden tarkkuus*).

Arvioijana toimiminen on säädetty asetuksella (A 1163/2004), jonka mukaan arvioijana toimivalla täytyy olla tutkintokielen opinnot ja/tai yliopiston ja ammattikorkeakoulun opettajalta vaadittu kelpoisuus tutkintokielestä. Lisäksi arvioijalta edellytetään kokemusta arvioitavan kielen opettamisesta ja kielitaidon arvioinnista ja hänen tulee suorittaa Opetushallituksen hyväksymä arvioijakoulutus. Arvioijan edellytetään myös osallistuvan yhteisiin arviointikoulutuksiin ennen jokaista arviointitilaisuutta. Arvioijakäyttäytymisen jatkuvalla seurannalla varmistetaan, että arvioijat toimivat johdonmukaisesti ja että heidän lempeys–ankaruus-linjansa ei poikkea merkittävästi muiden arvioijien arviointilinjasta.

4.2 Osa-aineistot

Kyselytieto kerättiin Webropol-kyselyohjelmalla kolmessa eri vaiheessa. Kyselyt koostuivat sekä monivalinnoista (Likert-tyyppinen asteikko) että avoimista kysymyksistä ja näiden yhdistelmistä. Ennen arviointien aloittamista kerätyn alkukyselyn avulla muodostettiin yleiskuva arviointiin liittyvistä kokemuksista, kuten arvioinnin helppoudesta/vaikeudesta sekä yleisestä arviointivarmuudesta. Lisäksi alkukyselyssä selvitettiin, kuinka varmalla mielellä arvioijat lähtivät aloittamaan arviointia kerralla, johon tutkimuskeruu sisältyi. Arvioinnin aikana arvioijien oli mahdollista jaksottaa arviointityötään eri mittaisiksi arviointirupeamiksi, joiden pituuteen ja määrään vaikutti esimerkiksi työn jakautuminen eri päiville sekä pidempien taukojen pitäminen. Kunkin yhtenäisen arviointijakson jälkeen arvioijat täyttivät kyselyn (nk. sessiokysely), jolla kartoitettiin tilanteista vaihtelua arviointiolosuhteissa ja arviointivarmuudessa sekä epävarmuuden vähentämiseen käytettyjä keinoja. Arviointien päätyttyä loppukyselyllä selvitettiin vielä mm. arvioijakoulutuksen vaikutusta arviointilinjaan ja kokemuksia tutkimukseen osallistumisesta.

Arvioinnin aikana arvioijat tekivät Excel-taulukkopohjaan suorittaja- ja tehtäväkohtaisesti muistiinpanot niistä arvioinneista, joista he jäivät epävarmoiksi taitotasoarvion annettuaan[2]. Arvioijia pyydettiin myös pohtimaan valmiiden vaihtoehtojen ja vapaan tekstikentän avulla, mitkä syyt heidän mielestään aiheuttivat epävarmuutta (*rajatapaus*, *epätasainen suoritus*, *tehtävänannon täyttyminen*, *suppea näyte*, *muu syy - mikä*) sekä miten he pyrkivät ratkaisemaan arvioinnin. Lisäksi heitä pyydettiin erittelemään epävarmaksi merkitsemiensä suoritusten arviointiin kulunutta aikaa (*Arviointi sujui: nopeasti tai melko nopeasti / ei nopeasti eikä hitaasti / melko hitaasti / hyvin hitaasti*).

Arvioijat antoivat kullekin suoritukselle yhteensä yhdeksän erillistä taitotasoarviota kaikkiaan neljästä eri tehtävästä tai niiden osioista. Keskitason tutkinnossa mahdolliset arviot ovat alle 3, 3 ja 4 (≈ alle B1, B1 ja B2). Arviointi perustui äänitettyihin puhumisen suorituksiin, joita arvioija saattoi kuunnella tarvittaessa useita kertoja joko kokonaan tai osittain. Arviointi toteutettiin etäarviointina.

4.3 Analyysimenetelmät

Määrällistä aineistoa tarkasteltiin tilastotieteen perusmenetelmien avulla (mm. frekvenssijakaumat, ristiintaulukointi, Khiin neliö -testit, korrelaatiokertoimet). Havaintojen kokonaismäärä vaihtelee kysymyksittäin, koska kaikkiin kysymyksiin vastaaminen ei ollut pakollista ja osa kysymyksistä salli useamman vastausvaihtoehdon valitsemisen. Arvioijien lempeyttä/ankaruutta sekä johdonmukaisuutta tutkittiin kielitaidon arvioinnin tutkimuksessa vakiintuneen menetelmän (ks. Aryadoust, Ng & Sayama, 2021), Many-Facets Rasch Measurement (MFRM) -mallin avulla (Linacre, 1994) Facets-ohjelmalla (versio 3.85.1, Linacre, 2023). Lisäksi muuttujien yhdysvaikutuksia tarkasteltiin Bias-analyysien avulla (Eckes 2011).

Kyselyiden avovastaukset sekä arvioinnin aikana tehdyt sanalliset muistiinpanot analysoitiin laadullisen temaattisen sisällönanalyysin keinoin etsimällä vastauksissa esiintyneet teemat, pelkistämällä teemat ja tekemällä sen jälkeen aineistolle teemoihin pohjautuva sisällönerittely (Braun & Clarke, 2006). Sisällönerittelyn avulla saatiin määrällistä tietoa tiettyjen teemojen esiintyvyydestä ja keskinäisistä suhteista aineistossa.

5 TULOKSET

5.1 Arviointivarmuus

5.1.1 Yleinen arviointivarmuus

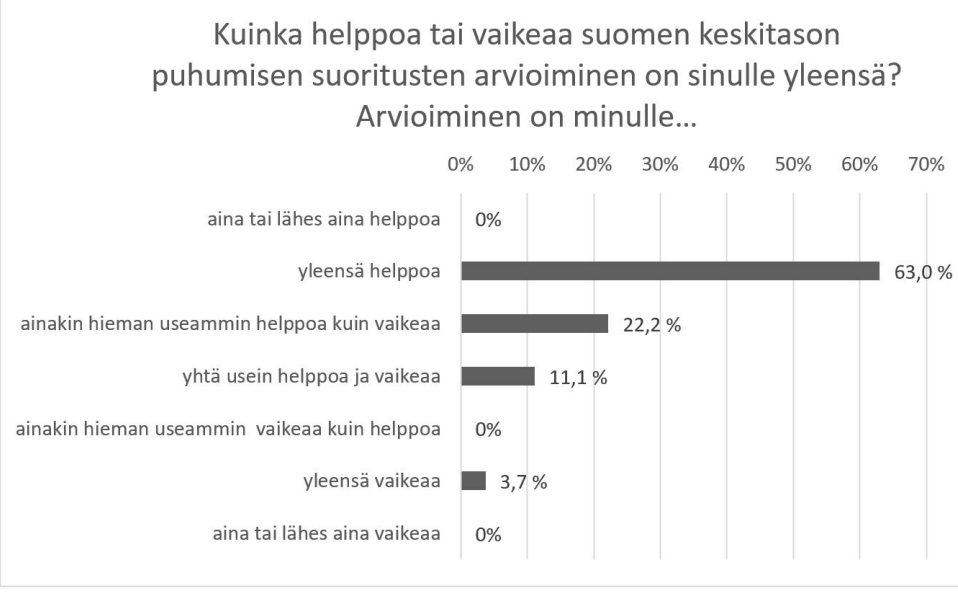
Alkukyselyn perusteella puhumisen testiarvioijien arviointivarmuus on yleisesti hyvä eikä arviointi tunnu kohtuuttoman vaikealta. Suurin osa tutkimukseen osallistuneista arvioijista, 21 arvioijaa (77,8 %), koki olevansa *yleensä varmoja* antamistaan puhumisen arvioista ja loput 6 arvioijaa (22,2 %) *ainakin hieman useammin varmoja kuin epävarmoja*. Kukaan arvioijista ei toisin sanoen kokenut epävarmuutta arviointiaan hallitsevaksi tunteeksi. Toisaalta kukaan ei myöskään ollut välttynyt epävarmuudelta.

2 ”Jäitkö epävarmaksi siitä, mille tasolle arvioimasi tehtävä sijoittuu? Tee taulukkoon muistiinpanot epävarmuuden syistä sekä tavoista, joilla pyrit varmistumaan arvioinnista. ”

Arviointivarmuus on arvioijahaastattelujen perusteella yhteydessä arviointitehtävän koettuun vaikeuteen siten, että erityisesti vaikealta tuntuva arviointi voi herättää epävarmuutta (Ahola, 2016). Myös omassa aineistossamme puhumisen arvioijien kokemus yleisestä arviointivarmuudestaan vastasi yksilötasolla pitkälti kokemusta arvioinnin helppoudesta/vaikeudesta: Spearmanin korrelaatiokertoimen arvo 0,73 on kohtalaisen korkea ja tilastollisesti erittäin merkitsevä ($p < 0{,}001$). Ne, jotka pitivät itseään yleensä varmoina puhumisen arvioijina, pitivät arviointia yleensä helppona. Arvioinnistaan useammin epävarmuutta kokevat puolestaan kokivat arvioinnin useammin myös vaikeaksi.

Arviointivarmuutta ja arvioinnin vaikeutta koskevien vastausten vertailu kuitenkin antaa viitteitä siitä, että kokemus arvioinnin haasteellisuudesta ei aina synnytä epävarmuutta tai johda tunteen säilymiseen. Arvioijilla on toisin sanoen keinoja selvittää vaikealtakin tuntuvia arviointeja niin, että he eivät lopulta jää epävarmoiksi antamistaan arvioista. Arvioinnin kokeminen vaikeaksi ei sekään kuitenkaan ole hallitseva tunne, vaan kolme viidestä tutkimukseen osallistuneesta arvioijasta (n = 17, 63,0 %) piti arviointia yleensä helppona, ja useimmat muutkin useammin helppona kuin vaikeana (kuvio 1).

KUVIO 1.
Arvioinnin vaativuus alkukyselyn mukaan



Alkukyselyn avovastausten perusteella[3] arvioijat pitivät arviointivarmuutta yleisesti lisäävinä tekijöinä erityisesti kokemusta (n = 15) sekä arvioijakoulutusta (n = 12) ja sitä, jos koulutuksen aikana syntyy yhteisymmärrys koulutusnäytteiden arvioinnista (erikseen mainitsi 6 edellisistä). Tähän liittyi tärkeänä osana riittävä yksimielisyys rajatapauksista sekä tehtävänantoihin liittyvien vaatimusten linjaaminen. Tärkeäksi koettiin myös arviointikriteerit, kuten niiden kertaaminen tai sujuva käyttö (n = 13), sekä oma ammatillinen osaaminen ja kokemus (n = 6). Muina arviointivarmuutta lisäävinä tekijöinä mainittiin itselle sopiva arviointitapa (n = 4), riittävä arvioitavien suoritusten määrä tai näytteen laajuus (n = 3), hyvä arviointivire (n = 2) sekä mahdollisuus arviointia koskeviin keskusteluihin kollegoiden kanssa, toimivat tehtävät ja riittävä

3 Maininnat alkukyselyn kysymykseen: ”Millaiset seikat tuovat arviointiisi varmuutta? Luettele tai kerro vapaamuotoisesti.”

arviointiaika (1 maininta kustakin).

Alkukyselyn lopussa arvioijilta kysyttiin, millä mielellä he lähtevät tekemään arviointeja, joista tutkimusaineisto kerättiin. Enemmistö arvioijista (n = 20, 74,1 %) koki olevansa arvioimaan lähtiessään *varmalla mielellä* (”arviointi on tuttua ja koulutus hyvässä muistissa”) tai *melko varmalla mielellä* (”edellisestä arvioinnista tai koulutuksesta on esimerkiksi ehkä jo kulunut aikaa tai koulutusnäytteistä oli pientä epävarmuutta, mutta mikään erityinen ei silti paina mieltä”, n = 5). Kaksi arvioijaa kuitenkin tunsi olevansa *vähän epävarmalla mielellä* (”edellisestä arvioinnista tai koulutuksesta on esimerkiksi ehkä jo kulunut aikaa, koulutusnäytteistä oli epävarmuutta tai oma vireystila epäilyttää, mikä jonkin verran painaa mieltä”).

5.1.2 Arviointivarmuus arvioinnin aikana

Arvioinnin aikana tehtyjen muistiinpanojen avulla syntyvä kuva arviointivarmuudesta on samansuuntainen kuin oman arvioinnin reflektointi ennen arviointia osoitti (ks. edellistä alalukua). Arvioinnin aikana suurin osa (n = 1020[4], 75,4 %) suorituksista koettiin kokonaisuudessaan varmoiksi arvioida; niissä ei ollut yhtään tehtäväkohtaista arviota, jonka antamisesta arvioija olisi jäänyt epävarmaksi. Jokainen tutkimukseen osallistunut puhumisen arvioija jäi kuitenkin vähintään satunnaisesti epävarmaksi ainakin yksittäisestä tietyn puhujan tehtäväkokonaisuuteen liittyvästä arviosta. Kaikkea epävarmuutta ei siis pystytty poistamaan esimerkiksi arviointikriteerien kertaamisen tai puhumisen näytteen uudelleen kuuntelemisen avulla. Silloin, kun arvioija tunsi epävarmuutta annetusta arviosta, se kuitenkin kohdistui tyypillisimmin vain yhteen (n = 121, 36,4 %) tai kahteen (n = 69, 20,8 %) tietyn puhujan tehtäväkohtaiseen arvioon. Sen sijaan aineistossa oli harvinaista, että tietyn puhujan suoritus olisi koettu kokonaisuudessaan (n = 12, 3,6 %) tai edes suurimmaksi osaksi epävarmaksi arvioida.

Epävarmuuden kokemuksen ja annetun taitotasoarvion yhteyttä tarkasteltiin tehtäväkohtaisesti ristiintaulukoinnin ja Khiin neliö -testin avulla. Khiin neliö -testin tuloksesta (χ^2 = 59,3, df = 2, $p < 0{,}001$) voidaan päätellä, että epävarmuuden kokemuksella ja taitotasoarviolla on yhteys. Jäännösarvoja (sovitettu standardoitu jäännös, SSJ) tarkastelemalla havaittiin, että epävarmuus yhdistyi useammin taitotasoarvioon alle 3 ja varmuus vastaavasti taitotasoarvioon 4. Arvioijat toisin sanoen kokivat enemmän epävarmuutta, kun he joutuivat pohtimaan, riittääkö suoritus keskitasolle eli täyttyvätkö taitotason 3 kriteerit, kuin pohtiessaan taitotasojen 3 ja 4 välistä rajaa.

Tutkitussa aineistossa epävarmuus oli yksilöllistä. Arvioinnin aikana ilmoitetussa epävarmuuden määrässä ensinnäkin oli suuria eroja eri arvioijien välillä ja lisäksi epävarmuus kohdistui pitkälti eri suorituksiin. Yhdistävä tekijä oli, että kukaan arvioijista ei ollut varma kaikista antamistaan arvioista. Jokaisella arvioijalla oli ainakin yksi epävarmaksi koettu arvio, ja vain kolmella arvioijalla tällaisia suorituksia oli vähemmän kuin kymmenesosa kaikista arvioiduista suorituksista.

Puhumisen suorituksista 380 (39,1 %) kaksoisarvioitiin. Kaksoisarvioiduista suorituksista yli puolet (n = 202, 53,2 %) sisälsi vähintään yhden epävarmaksi koetun arvion. Näistä vain noin viidenneksessä (n = 44, 21,8 %) arvioijien kokemus epävarmuudesta kohdistui saman puhujan suoritukseen, tosin ei tällöinkään välttämättä samaan tehtäväkohtaiseen arvioon. Loput (n = 158, 78,2 %) suorituksista olivat sellaisia, joissa vain toinen arvioijista merkitsi yhden tai useamman tehtäväkohtaisen arvion epävarmaksi. Tulos tarkoittaa, että epävarmaksi koetut arviot eivät kasaantuneet samoille suorittajille.

Epävarmaksi merkittyjen arvioiden hajoaminen eri puhujien suorituksiin osoittaa, että arvioijien kokema epävarmuus ei selity pelkästään suorituksiin tai puhujiin liittyvillä tekijöillä, vaan myös muilla seikoilla, kuten arvioijakohtaisilla eroilla sekä arviointikontekstiin liittyvillä seikoilla, on merkitystä. Seuraavissa alaluvuissa erittelemme näitä tekijöitä niin pitkälti kuin se keräämämme aineiston valossa on mahdollista. Käsittelemme aluksi arviointikokemuksen sekä arvioinnin ajallisen etenemisen mahdollista yhteyttä epävarmuuden kokemiseen. Esittelemme sen jälkeen lyhyesti arvioijien itse erittelemiä

4 Sisältää kaksoisarvioinnit.

epävarmuuden syitä. Hyödynnämme tässä sekä kyselyvastauksia että arvioinnin aikana tehtyjä muistiinpanoja. Lopuksi tarkastelemme sitä, onko koetulla epävarmuudella tutkitussa aineistossa yhteyttä arvioinnin laatuun.

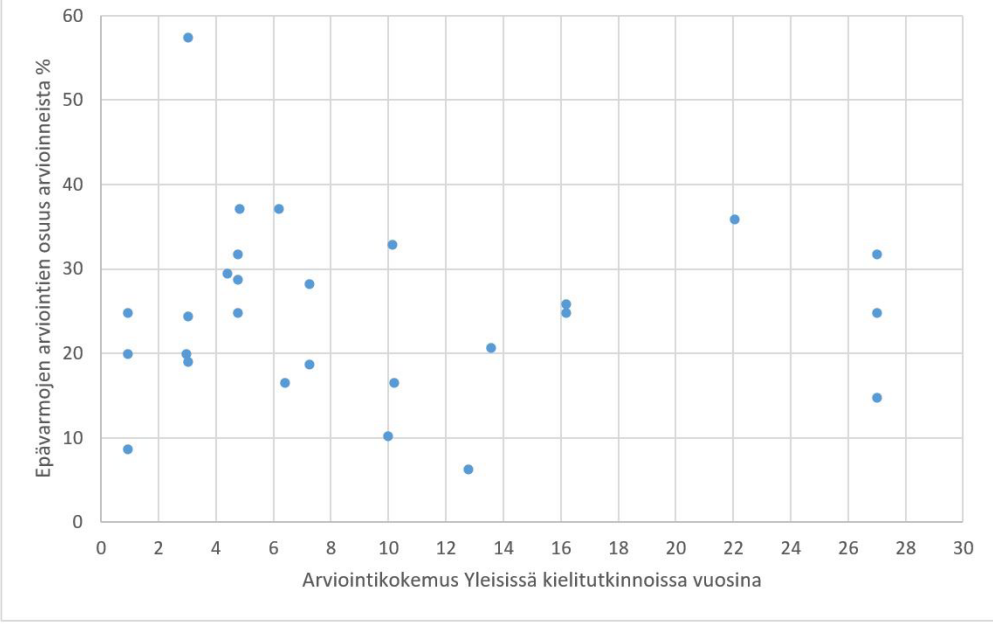
5.2 Kokemuksen vaikutus epävarmuuden määrään

5.2.1 Arviointikokemus vuosina

Arvioijien arviointikokemus testijärjestelmässä vaihteli noin vuodesta reiluun 27 vuoteen keskiarvon ollessa vajaa 10 vuotta. Kaikki arvioijat olivat osallistuneet aktiivisesti koulutuksiin ja puhumisen arviointiin. Arviointikokemuksen yhteyttä arvioinnin epävarmuuteen selvitettiin Spearmanin järjestyskorrelaatiokertoimen ja Pearsonin korrelaatiokertoimen avulla. Epävarmuutta mitattiin neljällä eri tavalla: arvioijien kokemuksella yleisestä arviointivarmuudestaan sekä varmuudestaan tutkintokerran arviointiin valmistautuessa, sessiokohtaisen arviointivarmuuden keskiarvolla (alku- ja sessiokyselyiden kysymykset sekä muistiinpanopohja, ks. Honko, Neittaanmäki, Huhta & Jarvis, 2023) sekä epävarman arvion sisältämien suoritusten suhteellisella määrällä.

Kokemuksella ja eri tavoin mitatulla epävarmuudella ei ollut tilastollisesti merkitsevää yhteyttä. Spearmanin sekä Pearsonin korrelaatiokertoimet olivat alhaisia (r_s = -0,258–0,311; r_p = -0,294–0,172), eivätkä ne eronneet tilastollisesti merkitsevästi nollasta ($p > 0,05$). Kuvio 2 havainnollistaa, kuinka epävarmojen arviointien osuus arvioinneista ei riipu arvioijan kokemusvuosista.

KUVIO 2.
Arviointivuosien yhteys epävarmaksi jääneisiin arviointeihin



5.2.2 Arviointikerran eteneminen

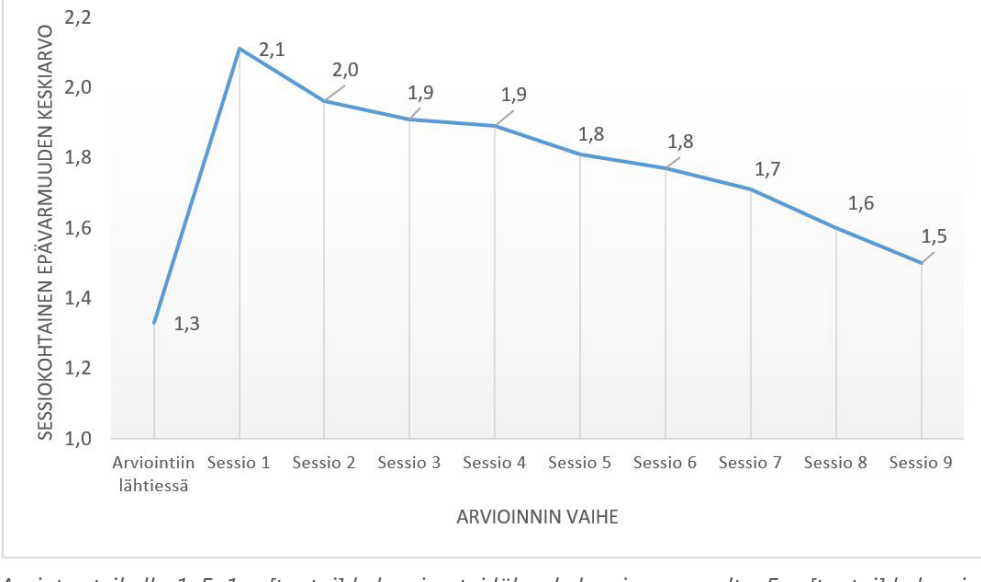
Tässä osiossa tarkastelemme yhden tutkintokerran arvioinneista muodostuvaa ajallista kokonaisuutta arviointikokemuksen karttumisen kannalta ja selvitämme, muuttuiko puhumisen arvioijien arviointivarmuus arvioinnin edetessä. Tätä tarkoitusta varten otimme kyselyaineistosta useita tarkastelupisteitä (kuvio 3). Ensimmäinen tarkastelupiste kertoo, kuinka varmalla mielellä arvioijat olivat alkukyselyn perusteella ennen arvioinnin aloittamista. Muut tarkastelupisteet kertovat arviointivarmuudesta kunkin yhtäjak-

soisen arviointirupeaman (session) aikana, mikä mahdollistaa eri arviointirupeamien keskinäisen vertailun.

Puhumisen arvioijat täyttivät arviointirupeamien jälkeen yhteensä 137 sessiokyselyä eli keskimäärin 5,1 arvioijaa kohti. Arviointirupeamien määrässä (1–9) oli vaihtelua eri arvioijien välillä, mitä selittää ennen kaikkea etäarvioinnin mahdollistama arviointityön melko vapaa rytmittäminen. Tarkastelupisteiden vertailu osoittaa, että arviointivarmuus oli suurimmillaan yhteisen koulutuksen jälkeen ennen arvioinnin aloittamista ja pienimmillään ensimmäisen arviointirupeaman aikana. Myös arviointisessioiden välillä oli hienoisia eroja siten, että arviointivarmuus hieman lisääntyi arviointien edetessä, vaikkei aivan saavuttanut lähtötilannetta. Arvioijat kokivat, että noin kolmanneksessa (f = 42, 30,7 %) rupeamista arviointi oli tuntunut *koko ajan tai lähes koko ajan varmalta* ja hieman yli puolessa (f = 71, 51,8 %) rupeamista *enemmän varmalta kuin varmalta*. Koko aineistossa vain muutamissa yksittäisissä rupeamissa (f = 5, 3,6 %) arvioijan tunne kallistui *epävarman* puolelle, ja lopuissa (n = 19, 13,9 %) rupeamista arviointi ei tuntunut *erityisen varmalta tai epävarmalta (tai oli hyvin vaihtelevaa)*. Arviointivarmuus oli kaikissa tarkastelupisteissä kuitenkin hyvä ja erot tarkastelupisteiden välillä pieniä.

KUVIO 3.

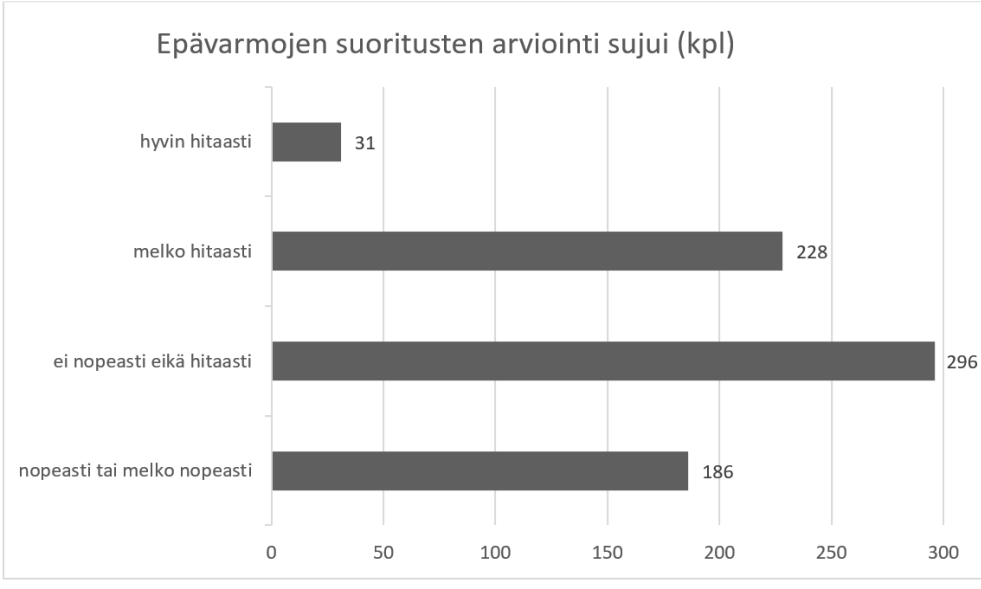
Epävarmuus arviointien edetessä arvioijien muistiinpanojen perusteella keskimäärin



Arviot asteikolla 1–5: 1 = [tuntui] koko ajan tai lähes koko ajan varmalta, 5 = [tuntui] koko ajan tai lähes koko ajan epävarmalta.

Jos epävarmaksi koetut arvioinnit olivat erityisen työläitä, myös tällä saattoi olla vaikutusta arviointirupeamien kestoon ja välillisesti myös niiden määrään. Siksi tarkistimme kyselyaineistosta, kokivatko arvioijat epävarmojen arviointien hidastaneen arviointia. Arvioijat arvioivat arviointinopeuttaan yhteensä 741 epävarmaksi arvioimansa tehtävän kohdalla. Vain harvoin (f = 31, 4,2 % tapauksista) he ilmoittivat epävarmaksi jääneen arvioinnin sujuneen hyvin hitaasti ja vain noin kolmasosassa tapauksista (f = 228, 30,8 %) melko hitaasti (kuvio 4).

KUVIO 4.
Epävarmoiksi koettujen arviointien arviointinopeus (yht. 741 kpl)



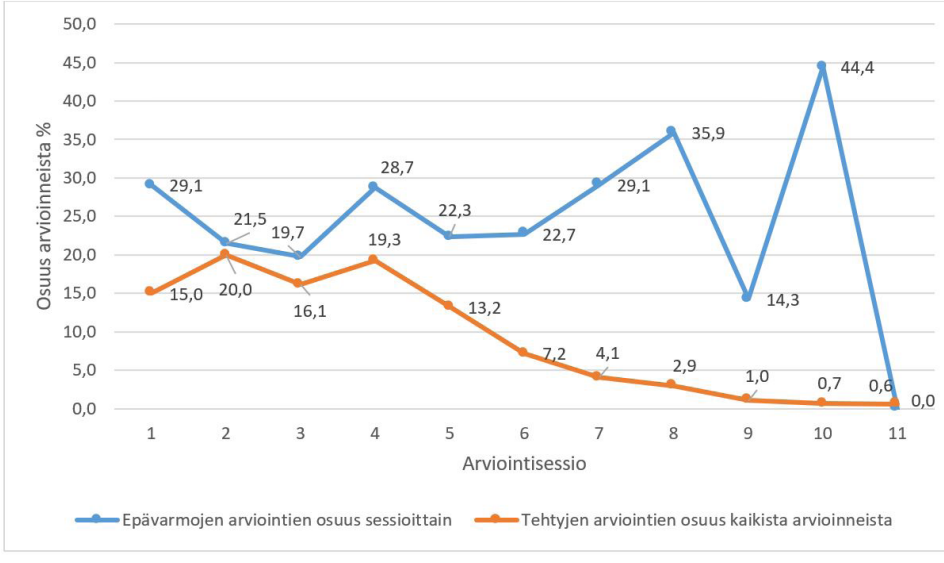
Usein epävarmuuden aiheuttamat syyt vaikuttavat toisin sanoen olleen sellaisia, että arvioijat eivät pysähtyneet suorituksen arviointiin erityisen pitkäksi aikaa. Tämä tarkoittaa, että epävarmat arvioinnit eivät voimakkaasti vaikuttaneet arviointirytmiin ainakaan sitä systemaattisesti hidastaen.

Edellä esitellyt tulokset perustuvat arvioinnin reflektointiin arviointitilanteen jälkeen. Arviointitietoon liitetty aikaleimatieto mahdollisti arviointivarmuuden vertailun arviointirupeamien välillä myös reaaliaikaisesti. Tähän käytimme tietoa siitä, kuinka paljon koulutuksesta oli kulunut aikaa arvioijien tehdessä arviointityötään sekä tietoa yhtenäisten, korkeintaan alle kahden tunnin taukoja sisältävien arviointisessioiden määrästä.

Aikaleimojen perusteella suurin osa arvioinneista tehtiin melko tasaisesti 2–4 päivän aikana arviointikoulutuksesta ja lähes kaikki arvioijat tekivät arviointinsa loppuun viikon kuluessa arviointikoulutuksesta. Kolmasosa arvioijista (n = 9, 33,3 %) aloitti työskentelemisen heti koulutuspäivänä ja noin puolet (n = 14, 51,9 %) seuraavana päivänä. Viimeistään neljäntenä päivänä koulutuksesta viimeisetkin arvioijat aloittivat arvioinnin. Yli 70 % (n = 952) arvioinneista tehtiin neljän ensimmäisen ja noin 84 % (n = 1131) viiden ensimmäisen arviointisession aikana (Kuvio 5).

Aikaleima-aineistosta lasketuilla arviointisessioilla ja epävarmuuden määrällä havaittiin yhteys (χ^2 = 158,1, df = 10, p<0,001). Ensimmäisessä arviointisessiossa epävarmuutta oli odotettua enemmän (SSJ = 3,7) kuten myös neljännessä (SSJ = 3,9), kahdeksannessa (SSJ = 6,6) ja kymmenennessä (SSJ = 6,5) sessiossa. Epävarmuutta oli odotettua vähemmän vain kolmannessa (SSJ = 4,0) ja viimeisessä sessiossa (SSJ = -2,5).

KUVIO 5.
Epävarmuus arviointien edetessä aikaleimatiedon perusteella



Koulutuksesta kuluneiden päivien määrällä ja koetulla epävarmuudella havaittiin myös yhteys (χ^2 = 83,3, df = 8, p <0,001). Päivien määrä rajattiin tarkasteluissa arviointiohjeen suosituksen mukaiseen enimmäismäärään, yhdeksään ensimmäiseen päivään. Epävarmoja arviointeja oli odotettua enemmän koulutuspäivänä (SSJ = 2,0) ja sitä seuraavana päivänä (SSJ = 2,3). Neljäntenä päivänä koulutuksesta epävarmojen arviointien määrä jäi selvästi odotettua alhaisemmaksi (SSJ = -5,2). Myös viidentenä päivänä epävarmuutta esiintyi vähemmän (SSJ = -2,1). Kuudentena päivänä koulutuksesta epävarmoja arviointeja annettiin taas odotettua enemmän (SSJ = 5,8). Seitsemäntenä päivänä epävarmuus väheni (SSJ = -3,1) ja kahdeksantena päivänä koulutuksesta epävarmuutta esiintyi jälleen odotettua enemmän (SSJ = 3,0).

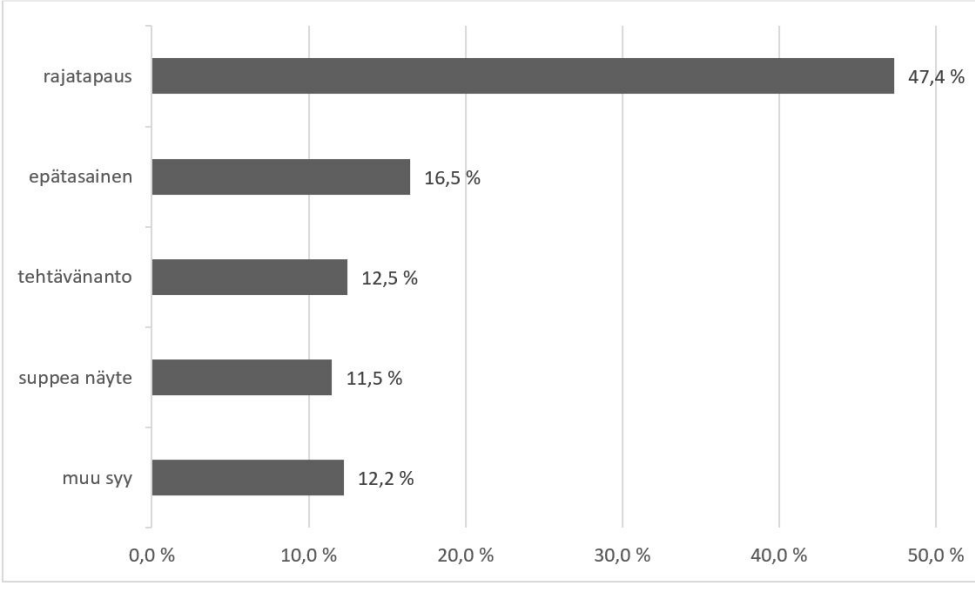
Sekä arvioinnin etenemisen että koulutuksesta kuluneen ajan yhteys epävarmuuteen osoittautui aaltoilevaksi. Tätä voivat osaltaan selittää puhumisen etäarviointiin liittyvät arviointikäytänteet, kuten kaksoisarviointien aikataulu, joihin palaamme myöhemmin.

5.3 Epävarmuuden syyt

5.3.1 Arvioinnin aikana

Taitotasojen rajalle sijoittuvat tapaukset ja epätasaiset suoritukset koetaan yleisesti vaikeiksi arvioida (esim. Ahola 2016). Nämä kaksi tekijää nimettiin myös tutkimassamme puhumisen testiaineistossa keskeisimmiksi epävarmuuden syiksi arvioinnin aikana. Arvioijien tekemien muistiinpanojen perusteella rajatapaukset (f = 581) kattoivat 47,4 % kaikista ilmoitetuista epävarmuuden syistä, mikä on enemmän kuin kolmen muun muistiinpanopohjassa valmiiksi nimetyn syyn yhteenlaskettu osuus. (Kuvio 6.) Suorituksen epätasaisuuden koettiin tuottavan epävarmuutta aina silloin tällöin (f = 202, 16,5 % valinnoista) ja hieman useammin kuin sen, täyttikö suoritus riittävästi tehtävänantoa (153, 12,5 %) tai oliko siinä riittävästi arvioitavaa (f = 141, 11,5 %).

KUVIO 6.
Syyt suorituksen merkintään epävarmaksi (muistiinpanot arvioinnin aikana)



On kuitenkin huomioitava, että kaikki edellä mainitut vastausvalikkoon valmiiksi nimetyt syyt liittyivät suorituksen ominaisuuksiin. Kohdassa ”muu syy” (f = 150, 12,2 % valinnoista) arvioijat ilmoittivat lisäksi muita suoritukseen liittyviä sekä tilanteisia ja yksilöllisempiä suoritukseen liittyviä syitä epävarmuudelle. Näistä kuvauksista (yht. 130) lähes neljännes (f = 31, 23,8 %) tosin toisti kuviossa 6 kuvattuja valmiisiin vastausvaihtoehtoihin sisältyviä suoritukseen liittyviä syitä, lähinnä epävarmuutta suorituksen epätasaisuudesta (esimerkki 1) tai tehtävänannon täyttymisestä.

(1) Uskomattoman epätasainen puhuja! Ehkä vaikein tapaus tähän mennessä. toistoa paljon

Jäljelle jäävistä kuvauksista (f = 104) yli puolet (59 mainintaa, 56,7 %) liittyi niin ikään suorituksen ominaisuuksiin. Eniten niissä käsiteltiin jollain tapaa epäselvää tai katkonaista puhetapaa (18 mainintaa, esimerkki 2) tai äidinkielen – yleensä viron – siirtovaikutusta (11 mainintaa, esimerkki 3). Lisäksi useita kertoja mainittiin huomion kiinnittyminen puutteelliseen rakenteiden hallintaan (8 mainintaa, esimerkki 4) sekä suorituksen sisällöllinen epäselvyys tai epäloogisuus (8 mainintaa, esimerkki 5).

(2) sanojen välissä aaaaa-epäröinti häiritsee todella paljon ja vaikeuttaa osittain ymmärtämistä (jännitys?); ääntäminen vaikeuttaa ajoittain ymmärtämistä

(3) Vieras korostus niin selvä, että kuunteleminen vaatii tarkkuutta. Mielestäni kokonaisuus kuitenkin selviää, joten hyväksyin.

(4) ”Tämä on näitä tapauksia, että kuinka paljon oikeita rakenteita painotetaan. Puhuu aiheista, selviää tehtävistä, mutta rakenteissa puutteita (ehkä fossiloituneita jo?): ”Hän on asuu”, jne.; rakenteellisesti perustasolla siis vielä, mutta ilmaisee itseään suht. helposti ja koherentisti. Mielipiteessä perustelut melko suppeat. Mietin itse häntä [oppilaitos] ja ajattelen, että pärjäisi siellä suullisesti, siksi päädyin kolmoseen; mutta tämä on todella kysymys, joka kiinnostaa minua (tuntuu, että monet on tosi tiukkoja rakenteissa ja itsekin olen opetustilanteessa)!”

(5) En tiedä, kummasta aiheesta puhuu.

Lisäksi epävarman arvion syyksi kuvattiin aiheen käsittelyn jääminen arvioitavaan taitotasoon ja tehtävään nähden liian hajanaiseksi, keskeneräiseksi, hankalaksi ymmärtää (”Tuotosta

on paljon, mutta pitää miettiä, onko se ymmärrettävää") tai omakohtaisten kokemusten varaan ("onko aiheen käsittely riittävän yleisellä tasolla") (n = 2 kussakin), sekä muita, yksittäisiä syitä, kuten puheen listamaisuus ja epäidiomaattisuus.

Suoritukseen liittyvien tekijöiden lisäksi arvioijat nimesivät kohdassa "Muu syy" epävarmuuden aiheuttajiksi myös ulkoisia häiriötekijöitä. Arvioinnin aikana tehdyissä muistiinpanoissa kaikki näistä koskivat nauhoitteen kuunneltavuutta (yhteensä 23 mainintaa), kuten hyvin hiljaista puheääntä ("lopusta en kerta kaikkiaan saa selvää, vaikka äänet on täysillä"), nauhoitteen heikkoa äänenlaatua tai häiritsevää taustaääntä ("puhujan hengitystä suoraan mikrofoniin").

Arvioijaan itseensä liittyviä epävarmuuden syitä kuvattiin arvioinnin aikana harvoin (f = 12). Kuvatut syyt olivat hyvin moninaisia, ja kustakin löytyy koko aineistossa vain yksittäisiä mainintoja. Tällaisia olivat alkuun pääsemisen vaikeus (f = 3), epävarmuus oman arviointilinjan ankaruudesta/lempeydestä (f = 2) tai johdonmukaisuudesta, keskittymisvaikeudet sekä epävarmuus koulutusnäytteen arvioimisesta tai kaksoisarvioinnin tekemisestä (1 mainin kussakin). Lisäksi kerran mainittuja epävarmuuden syitä olivat palaaminen jo tehtyyn arvioon ("muutin arviota, kun kuuntelin uudestaan"), epäilys väärästä tulkinnasta sekä arviointitauko ("Minulla oli perhesyistä tauko arvioinnissa ja omaa linjaa täytyy vähän taas tarkistaa.").

Vaikka esimerkiksi tehtävänannon kannalta riittävän abstraktiotason ja riittävän ymmärrettävyyden pohdinta liittyvät arviointikriteerien soveltamiseen, arviointikriteerien piirteitä itsessään ei arvioinnin yhteydessä kirjoitetuissa muistiinpanoissa kertaakaan mainittu epävarmuuden syyksi. Tehtävänanto puolestaan nimettiin epävarmuuden syyksi kerran tilanteessa, jossa arvioija koki tehtävänannon liian helpoksi ja siksi keskitason tutkinnossa puhumisen taitotasoja huonosti erottelevaksi.

5.3.2 Arvioinnin jälkeen

Arvioijia pyydettiin erittelemään epävarmuuden syitä myös irrallaan arviointitilanteesta välittömästi jokaisen arviointirupeaman jälkeen[5]. Avokysymyksiin saadut vastaukset (f = 131) toistavat pääosin teemoja, jotka ilmenivät myös alkukyselyn avovastauksissa sekä arvioinnin aikana tehdyistä muistiinpanoista. Suoritukseen liittyvät syyt (f = 85, 64,9 %) olivat pääosassa, ja useimmin mainittiin rajatapaukset, epävarmuus tehtävänannon täyttymisestä sekä suorituksen epätasaisuus (ks. Liite 1). Arviointitilanteen jälkeen suorituksiin liittyviä syitä pohdittiin kuitenkin myös holistisemmin tuomalla esille epävarmojen arviointien kertyminen tietyntyyppisiin tehtäviin (f = 5) – usein avoimiin kertomistehtäviin –, sekä edeltävien suoritusten tason vaikutus arviointiin (f = 4). Useiden heikkojen suoritusten jälkeen vahvemman suorituksen tason arvioiminen esimerkiksi saattoi tuntua epävarmalta.

Selkeämpi ero arvioinnin aikana tehtyihin muistiinpanoihin verrattuna oli kuitenkin arvioijaan itseensä liittyvien syiden käsittelyssä. Arvioinnin jälkeen kokoavassa pohdinnassa käsiteltiin huomattavasti useammin sisältöjä, jotka liittyivät arvioijan omaan arviointilinjaan, arviointirytmiin pääsyyn tai arvioijan tunnistamiin piirteisiin itsestään arvioijana (f = 31, 23,7 % maininnoista). Useita kertoja mainittiin esimerkiksi epävarmuus ensimmäisistä arvioinneista, omasta ankaruudesta/lempeydestä sekä kaksoisarviointiin tulleista ja kaksoisarviointiin lähtevistä arvioinneista. Haastavaksi ei toisin sanoen koettu ainoastaan koko arviointikerran ensimmäisiä arviointeja, vaan myös tauon jälkeen uuden arviointirupeaman aloittaminen vaati käynnistelyä. Tämä voi osaltaan selittää suhteellisen pieniä muutoksia arviointivarmuudessa eri arviointirupeamien välillä (kuvio 4 edellä). Kaksoisarvioitavia suorituksia puolestaan pidetään usein vaikeina arvioida. Vaikka arvioijat kertoivat kaksoisarviointien yleisesti tuovan arviointiin turvaa, niiden kerrottiin juuri oletetun vaikeuden vuoksi myös lisänneen epävarmuutta. Lisäksi myös tietoisuus arviointien vertaamisesta jännitti ja herätti epävarmuutta ainakin yhdessä arvioijassa.

5.4 Epävarmuuden vaikutus arvioinnin laatuun

Arvioinnin laadun mittareina käytettiin arvioijan ankaruutta/lempeyttä sekä johdonmukaisuutta.

5 Sessiokyselyn kysymys: *Jos koit tämän arviointisession aikana epävarmuutta, mistä se mielestäsi johtuu? Kerro vapaamuotoisesti.*

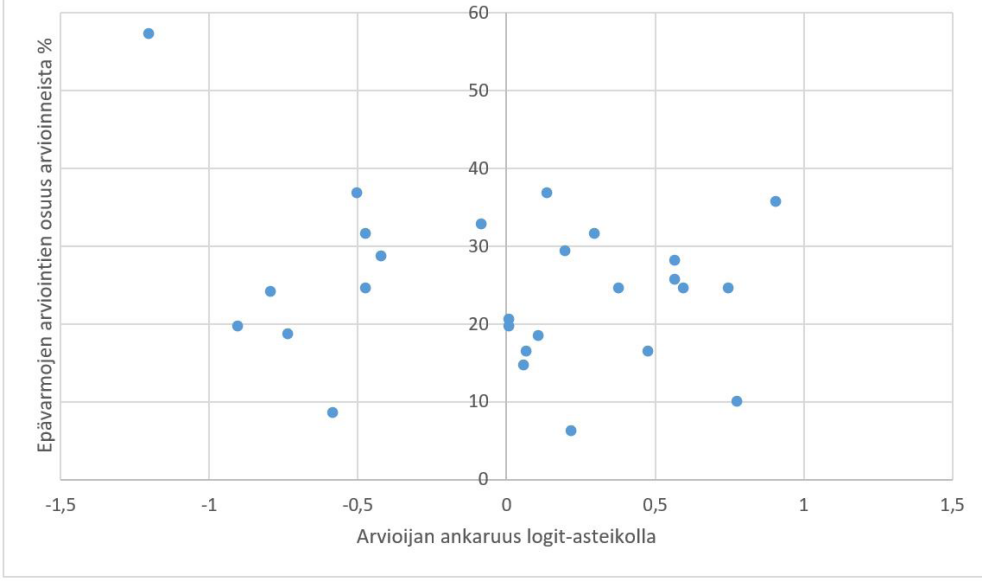
Näitä kuvaavat tunnusluvut saatiin analysoimalla arviointiaineisto (12059 datapistettä) Facets-ohjelmalla nk. kolmen facetsin (arvioija, suorittaja ja tehtävä) skaalamuuttujien (*rating scale*) mallilla. Tuloksia (ks. yhteenveto tuloksista, Liite 2) voidaan pitää tarpeeksi luotettavina, koska aineisto sopi malliin riittävän hyvin[6] ja arvioijat käyttivät arviointiskaalaa koko laajuudessaan ja johdonmukaisesti. Koska tutkimuksen keskiössä ovat arvioijat, tässä esitellään lyhyesti vain arvioijia koskevat tulokset.

Arvioijan logit-asteikollinen ankaruusparametri keskitettiin nollan ympärille ja sen arvot vaihtelivat välillä -1,20–0,91. Mitä pienempi ankaruusparametrin arvo on, sen ankarammasta arvioijasta on kyse. Mittaustarkkuus oli hyvä, koska ankaruusparametrien keskivirheet olivat varsin pieniä (0,10–0,12). Arvioija-facetsin reliabiliteetti (0–1) kuvaa, kuinka erilaisia ankaruusparametrit ovat. Arvo on lähellä nollaa, kun arvioijat ovat ankaruudeltaan samankaltaisia. Arvioijilla oli ankaruudessa eroja, koska ankaruusparametrien reliabiliteetti oli korkea 0,96.

Arvioijan johdonmukaisuutta kuvaavaksi arvoksi valittiin lähipainotettu keskineliöpoikkeama (*Infit mean square*), jonka hyväksymisrajoiksi asetettiin 0,70–1,30 (odotusarvo 1) (Eckes, 2011). Sen keskiarvo aineistossa oli 1,00 (KH = 0,10). Arvioijakohtaiset arvot vaihtelivat välillä 0,82–1,28, eli kaikki arvioijat olivat arvioinneissaan johdonmukaisia. Tämän perusteella voidaan todeta, että epävarmuudella ei ole tämän aineiston perusteella yhteyttä arvioinnin johdonmukaisuuteen.

Arvioinnissa ilmenevän epävarmuuden yhteyttä arvioijan ankaruuteen selvitettiin Spearmanin järjestyskorrelaatiokertoimen ja Pearsonin korrelaatiokertoimen avulla. Arvioijan ankaruudella ei ollut tilastollisesti merkitsevää yhteyttä eri tavoin mitattuun epävarmuuteen: epävarmaksi merkittyjen arviointien suhteellisella määrällä, arvioijien yleisellä arviointivarmuudella, arviointivarmuudella arviointien alkaessa tai arviointisessiokohtaisten arviointivarmuuksien keskiarvolla. Spearmanin sekä Pearsonin korrelaatiokertoimet olivat alhaisia (r_s = -0,040–0,128; r_p = -0,237–0,096) ja ne eivät eronneet tilastollisesti merkitsevästi nollasta ($p > 0,05$). Kuvio 7 havainnollistaa, kuinka epävarmojen arviointien määrän ja arvioijan ankaruuden välillä ei ole riippuvuutta.

KUVIO 7.
Ankaruuden yhteys epävarmojen arviointien osuuteen



6 4,8 % standardoitujen jäännöksien itseisarvoista oli ≥ 2 (raja-arvo 5 %) ja 1,0 % ≥ 3 (raja-arvo 1 %).

Epävarmuuden vaikutusta arvioinnin ankaruuteen tutkittiin lisäksi Facets-analyysin yhteydessä tehtyjen Bias-analyysien avulla. Bias-analyysien avulla tutkittiin, vaihteleeko arvioijien ankaruus sen suhteen, arvioivatko he varmoiksi vai epävarmoiksi merkitsemiään suorituksia. Havaittiin, että arvioijilla ei ollut tässä tilastollisesti merkitseviä eroja (Bias-termien t-arvot välillä -2–2).

6 PÄÄTÄNTÖ

Tutkimuksemme tavoite oli selvittää arvioijien arviointivarmuutta puhumisen arvioinnissa. Käytimme aineistona Yleisten kielitutkintojen arvioijilta suomen kielen keskitason tutkinnossa kerättyä kyselyaineistoa sekä arvioinnin aikana tehtyjä muistiinpanoja ja arviointikirjauksia.

Tulokset osoittavat, että yksikään arvioija ei kokenut epävarmuutta hallitsevana tunteena yleisesti eikä myöskään sillä arviointikerralla, jonka aikana tutkimusaineisto kerättiin. Koulutetut arvioijat eivät myöskään pitäneet arviointia Yleisissä kielitutkinnoissa erityisen vaikeana tehtävänä, ja lisäksi heillä näyttää olevan keinoja selvittää vaikeitakin arviointeja niin, että lopullinen arviointitulos tuntuu varmalta.

Arvioinnin aikana suurin osa suorituksista koettiin kokonaisuudessaan varmoiksi arvioida ja silloin, kun epävarmuutta esiintyi, se kohdistui yleensä vain yhteen tehtävään. Kaikki tutkimukseen osallistuneet arvioijat kuitenkin jäivät ainakin satunnaisesti epävarmaksi antamistaan puhumisen taitotasoarvioista, mikä vastaa oletuksia epävarmuuden ilmenemisestä subjektiiviseen päätöksentekoon perustuvassa arvioinnissa (esim. Anderson ym., 2019). Tästä huolimatta epävarmuus oli tutkitussa aineistossa yksilöllistä: epävarmuuden määrässä oli suuria eroja eri arvioijien välillä ja epävarmuus kohdistui pääosin eri suorituksiin. Arvioijat myös aloittivat työskentelyn eri aikaan ja rytmittivät työtään eri tavoin.

Epävarmuuden määrässä ilmeni vaihtelua arviointijakson eri vaiheissa. Kyselyaineisto osoitti, että arviointivarmuus oli suurimmillaan yhteisen koulutuksen jälkeen ennen arvioinnin aloittamista ja pienimmillään ensimmäisen arviointirupeaman aikana. Arviointisessioiden välillä oli hienoisia eroja siten, että arviointivarmuus hieman lisääntyi arviointien edetessä, vaikkei aivan saavuttanut lähtötilannetta. Arvioinnin reaaliaikaisen aikaleima-aineiston perusteella arviointivarmuudessa tapahtuva muutos ei kuitenkaan ollut lineaarinen vaan aaltoileva. Epävarmojen suoritusten merkintä esimerkiksi selvästi väheni päivinä, jolloin arvioija kohdisti huomiotaan kaksoisarviointiin lähetettävien suoritusten valintaan ja toisaalta lisääntyi voimakkaasti päivinä, jolloin arvioitiin kaksoisarviointiin tulleita suorituksia. Kaksoisarviointeihin kasautuva epävarmuus on odotuksenmukaista, sillä arvioijat pitävät kaksoisarvioitavia suorituksia usein haasteellisina arvioida, ja lisäksi tietoisuus arviointien vertaamisesta jo itsessään huolettaa ainakin yksittäisiä arvioijia. Arviointivarmuus oli kaikissa tarkastelupisteissä ja eri metodein tarkasteltuna kuitenkin yleisesti hyvä ja erot tarkastelupisteiden välillä pieniä.

Arvioinnin aikana epävarmuuden syiksi nimettiin, osin keruumetodin vuoksi, erityisesti suoritukseen liittyviä tekijöitä, kuten kahden taitotason rajalla olevat tai epätasaiset suoritukset, epävarmuus tehtävänannon täyttymisestä, arvioitavan suorituksen niukkuus. Myös aiemman tutkimuksen mukaan tällaiset suoritukset koetaan yleisesti vaikeiksi arvioida (Ahola, 2016). Lisäksi epävarmuutta herättivät jotkin suorituksesta riippumattomat syyt, joista useimmin mainittiin heikosti kuuluva puheääni ja äänitteen heikkolaatuisuus. Tutkimuksen perusteella epävarmuutta koskeva reflektio ennen arviointia ja sen jälkeen kuitenkin antaa osin erilaisen kuvan epävarmuuden määrästä ja syistä kuin arvioinnin aikana kerätty aineisto. Esimerkiksi arvioijaan itseensä ja arviointitehtävän vastuullisuuteen liittyviä syitä sekä puhujaan liittyviä erityispiirteitä joko reflektoidaan tai raportoidaan herkemmin arviointitilanteen ulkopuolella kuin arvioinnin aikana: Monet arvioijat kertoivat kokeneensa epävarmuutta esimerkiksi omasta arviointilinjastaan sekä arvioitavaan suoritukseen ja puhujaan liittyvistä erityispiirteistä, kuten voimakkaasta aksentista tai vironkielisyydestä, tai omasta taustastaan. Tähän liittyi muun muassa pelko siitä, että oma pitkä kokemus kielenoppijoiden parissa saa ymmärtämään suorituksia liiankin hyvin. (Samankaltaisia havaintoja myös Ahola, 2016; Tarnanen, 2002.) Arvioinnin aikana tämänkaltaisia syitä ei – osin mahdollisesti erilaisesta kysymyksenasettelusta johtuen – yhtä selvästi tuotu

esille, vaan arvioijat rajasivat huomionsa selvemmin päätöksentekotehtävän ytimeen eli puhesuoritukseen sekä arviointiohjeisiin ja -kriteereihin.

Arviointikokemuksen pituudella ei tutkitussa arvioijaryhmässä ollut yhteyttä arviointivarmuuteen, eli epävarmuus ei näytä poistuvan tai systemaattisesti vähentyvän arviointivuosien karttuessa. Tämä koski sekä kyselyillä selvitettyä yleistä arviointivarmuutta että epävarmaksi jääneiden arviointien määrää arvioinnin aikana. Tulos on sikäli yllättävä, että arvioijista yli puolet (n = 15) mainitsi alkukyselyssä kokemuksen arviointivarmuutta lisääväksi tekijäksi. Myös Aholan (2016, 2022) sekä Tarnasen (2002) aiemman haastattelututkimuksen aineistossa esiintyi kuvauksia siitä, kuinka kokemuksen myötä arvioinnista tulee varmempaa, kun arviointikriteerit tulevat tutuimmaksi, arviointi muuttuu alkuaikoja holistisemmaksi ja eri taitotasojen väliset eroavaisuudet selkeytyvät.

Tulosten eroja voi selittää esimerkiksi se, että arviointitilanteen ulkopuolella arvioijat reflektoivat arviointitoimintaansa kokonaisuutena, eivät yksittäisten suoritusarviointien kautta. Kokeneilla arvioijilla voi myös olla enemmän keinoja käsitellä arviointiin liittyviä epävarmuuden tunteita, vaikka epävarmuus itsessään ei kokemuksen myötä poistukaan (myös Ahola, 2016; Tarnanen, 2002). On myös huomioitava, että vaikka arvioijien välillä oli juuri Yleisiin kielitutkintoihin liittyvässä arviointikokemuksessa suuria eroja, useimmat olivat hankkineet arviointikokemusta muualtakin esimerkiksi opettajan työssä, mitä tutkimusasetelmamme ei huomioi. Sekin, että tutkimuksen osallistujat ovat valikoitunut joukko ja toimivat arvioijina aktiivisesti, voi vaikuttaa tuloksiin: vaikka puhumisen arvioijat kyselyaineistossamme yleisesti kokivat kokemuksen lisäävän arviointivarmuutta, he eivät samalla tavalla tuoneet esille kokemuksen puutetta epävarmuutta aiheuttavana tekijänä. Lisäksi tulee huomioida sekin mahdollisuus, että arviointitoimintaan on syntynyt erityisesti kokeneen arvioijan varmuutta painottava asiantuntijadiskurssi, jota osin ehkä perusteettomastikin kierrätetään tai ei haluta julkisesti kyseenalaistaa.

Tutkitussa aineistossa epävarmuus ei vaikuttanut arvioinnin laatuun eli sen johdonmukaisuuteen tai ankaruuteen/lempeyteen, mikä tukee aiempaa tutkimusta (Bosshardt ym., 2016). Epävarmojen arvioijien arvioimat henkilöt saivat siis yhtä oikeudenmukaisia arvioita kuin varmojen arvioijien arvioimat henkilöt. Vaikka suorituksen arviointi toisinaan vei paljon aikaa, epävarmat arvioinnit eivät voimakkaasti myöskään vaikuttaneet arviointirytmiin.

Arvioinnin laatutekijöistä ja riittävästä arviointivarmuudesta huolehtiminen on kuitenkin tärkeää, jotta epävarmuus ei kuormita arvioijia liikaa ja jotta arviointitiedon luotettavuus säilyy. Arvioijat tulee valikoida ja kouluttaa, ja lisäksi tehtävien toimivuus sekä arviointiohjeiden ja kriteeristön oikea käyttö tulee varmistaa. (Myös esim. McNamara ym., 2019; Fan & Yan, 2020.) Tämä tarkoittaa muun muassa arvioijatoiminnan seuraamista ja palautteen antamista. Myös teknisiin ongelmiin tulee puuttua. Jos puhumista arvioidaan äänitallenteelta, ääneen laatu on varmistettava. Arvioijan puolestaan tulee huolehtia riittävästä virkeydestä, arviointiolosuhteista sekä sopivasta arviointirytmistä.

Arvioijat itse pitivät tärkeinä arviointiharjoituksia ja keskustelua sisältäviä arviointikoulutuksia jokaisen arviointikerran yhteydessä. Tämä vastaa aiempaan kansainväliseen tutkimukseen pohjautuvaa näkemystä siitä, että hyvän arviointilinjan muodostumisessa koulutuksilla ja konkreettisten näytteiden arviointiharjoituksilla on tärkeä rooli (Fan & Yan, 2020; Suomessa mm. Ahola, 2022; Tarnanen, 2002). Esimerkiksi tutkintoon osallistujan kannalta usein merkityksellinen ”kolmosen raja” on arviointivarmuuden varmistamisen kannalta erityisen tärkeä, sillä tämän tutkimuksen tilastollisten analyysien perusteella epävarmuus yhdistyy useammin taitotasoarvioon alle 3 ja varmuus vastaavasti taitotasoarvioon 4. Myös arviointiin liittyvien tunteiden jakaminen on aineistomme perusteella arvioijille merkityksellistä ja auttaa niiden tunnistamisessa ja käsittelyssä (myös Ahola, 2016).

Epävarmuuden tunteita tuskin voi kielitaidon arvioinnista poistaa kokonaan (myös Tarnanen, 2014; päätöksenteosta yleisesti Anderson ym., 2019). Epävarmuus ei myöskään ole pelkästään kielteinen ilmiö, vaan kohtuullisena pysyessään sillä voi aineistomme perusteella olla arviointiin myös monenlaisia myönteisiä vaikutuksia, kuten huomion kohdentaminen, merkityksellisyyden kokemukset, mahdollisuus kehittyä arvioijana jne. Arviointiin liittyvästä päätöksenteosta tarvi-

taan kuitenkin lisätutkimusta. Tutkimusta olisi hyvä laajentaa arvioijien pitkittäisseurantaan ja myös muiden kuin suomen kielen arviointiin sekä testiarvioinnista myös muihin puhumisen arvioinnin konteksteihin. Kielikoulutukseen liittyvää puhumisen arviointia koskeva kohdennettu tutkimus esimerkiksi tuottaisi tärkeää tietoa eri oppilaiden ja ryhmien välisen arvioinnin yhdenmukaisuudesta sekä arvioinnin kuormittavuudesta. Tämän tutkimuksen perusteella satunnaista tai harvakseltaan esiintyvää epävarmuuden tunnetta ei tarvitse pelätä tai peitellä, vaan epävarmuus on inhimillistä ja kuuluu myös kielitaidon testiarviointiin. Epävarmuus muutti muotoaan arvioinnin edetessä ja kohdistui eri tilanteissa sekä eri arvioijilla osin erilaisiin asioihin mutta ei heikentänyt arvioinnin laatua.

LÄHTEET

Ahola, S. (2016). Puhetta arvioinnista: Yleisten kielitutkintojen arvioijien käsityksiä arvioinnista. Teoksessa A. Huhta & R. Hildén (toim.), *Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e,* 9, (s. 89–109). AFinLA. http://journal.fi/afinla/article/view/60848

Ahola, S. (2022). *Rimaa hipoen selviää tilanteesta: yleisten kielitutkintojen suomen kielen arvioijien käsityksiä kielitaidon arvioinnista ja suullisesta kielitaidosta.* Jyväskylän yliopisto. http://urn.fi/URN:ISBN:978-951-39-9005-3

Ahola, S. & Halonen, M. (2022). Ensivaikutelmat kielitaidon arvioinnissa. *Kieli, koulutus ja yhteiskunta, 13*(2). https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-maaliskuu-2022/ensivaikutelmat-kielitaidon-arvioinnissa

Anderson, E.C., Carleton, R.N., Diefenbach, M. & Han, P.K.J. (2019). The relationship between uncertainty and affect. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.02504

Aryadoust, V., Ng, L.Y. & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Bhise, V., Rajan, S.S., Sittig, D.F., Morgan, R.O., Chaudhary, P. & Singh, H. (2018). Defining and measuring diagnostic uncertainty in medicine: A systematic review. *Journal of General Internal Medicine, 33*(1), 103–115. https://doi.org/10.1007/s11606-017-4164-1

Bosshardt H., Packman, A., Blomgren, M. & Kretschmann, J. (2016). Measuring stuttering in preschool-aged children across different languages: An international study. *Folia Phoniatrica et Logopaedica, 67*(5), 221–230. https://doi.org/10.1159/000440720

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. http://dx.doi.org/10.1191/1478088706qp063oa

Carter, R. & McCarthy, M. (2017). Spoken grammar: where are we and where are we going? *Applied Linguistics, 38*, 1–20. https://doi.org/10.1093/applin/amu080

Cumming, A., Kantor, R. & Powers, D.E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*, 67–96. https://doi.org/10.1111/1540-4781.00137

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135.

Duijm, K., Schoonen, R. & Hulstijn, J. H. (2017). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing, 35*(4), 501–527. https://doi.org/10.1177/02655322177125

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch Analysis. *Language Assessment Quarterly, 2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Euroopan neuvosto (2011). *Manual for language test development and examining.* Strasbourg: Council of Europe. https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b

Fan, J. & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. Systematic review article. *Frontiers in Psychology, 27*. https://doi.org/10.3389/fpsyg.2020.00330

Halonen, M., Huhta, A., Ahola, S., Hirvelä, T., Neittaanmäki, R., Ohranen, S. & Ullakonoja, R. 2020. Ensikielen tunnistamisen merkityksestä suullisen kielitaidon arvioinnissa Yleisissä kielitutkinnoissa. Teoksessa S. Grasz, T. Keisanen, F. Oloff, M. Rauniomaa, I. Rautiainen & M. Siromaa (toim.), *Menetelmällisiä käänteitä soveltavassa kielentutkimuksessa*. AFinLAn vuosikirja 78, (s. 56–70). AFinLA. https://doi.org/10.30661/afinlavk.89453

Honko, M., Neittaanmäki, R., Jarvis, S. & Huhta, A. (2023). Beyond literacy and competency – The effects of raters' perceived uncertainty on assessment of writing. *Assessing Writing, 57*. https://doi.org/10.1016/j.asw.2023.100768.

Honko, M., Neittaanmäki, R., Huhta, A. & Jarvis, S. (2023). Arvioijakäyttäytymisen mittari: tutkimuskyselyt sekä muistiinpanotaulukko Arvioinnin epävarmuus ja epävarmuuden vaikutukset arvioinnin laatuun Yleisissä kielitutkinnoissa (YKI) -tutkimuksesta. Jyväskylän yliopisto. https://doi.org/10.17011/jyx/dataset/88216

Huhta, A. & Hilden, R. (2013). Kielitaidon arvioinnin metodologisia vaihtoehtoja. Teoksessa A. Räisänen (toim.), *Oppimisen arvioinnin kontekstit ja käytännöt*, (s. 159–186). Opetushallitus. https://karvi.fi/publication/oppimisen-arvioinnin-kontekstit-ja-kaytannot/

Huhta, A. & Ahola, S. (2019). Arviointi kielipoliittisena ja kielikoulutuspoliittisena toimintana. Teoksessa T. Saarinen, P. Nuolijärvi, S. Pöyhönen & T. Kangasvieri (toim.), *Kieli, koulutus, politiikka*, (s. 287–313). Vastapaino.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2023). Facets computer program for many-facet Rasch measurement, version 3.85.1. Winsteps.com

Luoma, S. (2004). *Assessing Speaking*. Cambridge Language Assessment Series. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Lynch, B. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351–372. DOI: 10.1177/026553220101800403

McNamara, T., Knoch, U. & Fan, J. (2019). *Fairness, justice and language assessment.* Oxford University Press.

OPH 2003. Yleiset kielitutkinnot. Puhumisen arviointikriteerit. Opetushallitus. https://www.jyu.fi/hytk/fi/laitokset/solki/yki/yleista/tietoakielitutkinnoista/puhumisen_arviointikriteerit.pdf

Tarnanen, M. (2002). *Arvioija valokeilassa – Suomi toisena kielenä -kirjoittamisen arviointia*. Jyväskylän yliopisto.

Tarnanen, M. (2014). Arvioija taidon arvottajana. Teoksessa T. Leblay, T. Lammervo & M. Tarnanen (toim.), *Yleiset Kielitutkinnot 20 vuotta*, (s. 115–124). Opetushallitus.

Tossavainen, H. (2016). Kielitestien eettisyydestä ja oikeudenmukaisuudesta. Teoksessa A. Huhta & R. Hildén (toim.), *Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e,* 9, (s. 27–43). AFinLA

Wind, S. A., Jones, E. & Bergin, C. (2021). Principals' severity affects teacher evaluation: statistical adjustments mitigate effects. *School Effectiveness and School Improvement, 32*(3), 413–-429. https://doi.org/10.1080/09243453.2021.1892773

Winke, P. & Gass, S. (2011). *The Relationship Between Raters' Prior Language Study and the Evaluation of Foreign Language Speech Samples*. TOEFL iBT® Research Report TOEFL iBT–16. https://files.eric.ed.gov/fulltext/EJ1110379.pdf

Youn, S.-J. (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment, 7*(1), 32–60.

RATER UNCERTAINTY AND ITS EFFECTS ON THE QUALITY OF THE PROFICIENCY RATING IN THE FINNISH SPEAKING TEST ASSESSMENT

- Mari Honko, Centre for Applied Language Studies, University of Jyväskylä
- Reeta Neittaanmäki, Centre for Applied Language Studies, University of Jyväskylä

The study examines the uncertainty of trained raters in speaking test assessment and its effect on the quality of the proficiency ratings of adult learners of Finnish. The data consists of 12,059 task-specific speaking proficiency ratings and notes taken by the raters during the assessment, as well as questionnaire data. Qualitative content analysis as well as statistical methods were used to answer the research questions. In the studied group, certainty dominated the assessment. All the speaking raters remained, however, at least occasionally uncertain of the proficiency ratings they gave. When uncertainty did occur, it typically covered only one task-specific rating in the set of tasks of a given speaker. The raters named several reasons for their uncertainty. The level of uncertainty varied at different stages of the assessment, with uncertainty being highest at the beginning of the assessment and increasing also as the second rater's previously assessed performance entered the double-assessment. Uncertainty turned out to be largely individual, as in double-assessed performances, the raters' uncertainty was mainly focused on the performances of different speakers. The length of the rating experience did not explain the variation in rater (un)certainty. Perceived uncertainty did not affect the quality of the assessment, i.e. its consistency or severity/leniency. Yet, it is important to take care of raters and assessment quality factors and so that uncertainty does not burden raters and so that the validity and reliability of assessment is maintained.

Keywords: language assessment, language test, speaking, uncertainty

LIITE 1.
Epävarmuuden syyt välittömästi arvioinnin päätyttyä, sessiokyselyistä tehty sisällönerittely

Suoritukseen liittyvät	**85**
rajatapaus	22
tehtävänannon täyttyminen	15
suorituksen epätasaisuus	13
puutteeet rakenteissa / rakenteet ja sanasto	8
niukkuus	6
puhetapa	5
äidinkielen vaikutus	4
ääntäminen	3
abstraktiotaso	2
sisällöllinen epäselvyys tai monitulkintaisuus	2
ymmärrettävyys	2
kieli vs. tilanteesta selviäminen	1
ohjeiden seuraaminen	1
vilppi	1
Arviointipotti - esim. heikon yleislinjan jälkeen vähän parempi suoritus	**4**
Tiettyyn tehtävään liittyvät suoritukset	**5**
Arvioijaan liittyvät tekijät	**31**
alku vaikea, myös uuden session alku	10
lempeys/ankaruus	8
kaksoisarviointi	7
lopussa epävarmuus tehtyjen linjausten pohtimisesta	2
oma arviointilinja yleensä	1
ulkopuolelta tullut palaute omasta arviointilinjasta	1
koulutusnäytteen arviointi	1
vireys	1
Olosuhteet / ulkoiset tekijät	**3**
aikapaine (ympäröivä arki)	1
koulutus	1
tutkimus	1
Kohdentamaton	**3**
yht.	**131**

LIITE 2.

Facets-mallin yhteenvetotulokset suorittajille, arvioijille ja tehtäville

Statistics	*Suorittajat*	*Arvioijat*	*Tehtävät*
Mean measure	-1.40 (0.82)	0.00 (0.11)	0.00 (0.07)
SD measure	2.91 (0.37)	0.55 (0.00)	0.54 (0.01)
Adj. (True) SD	2.77	0.54	0.54
Min	-7.46	-1.20	-0.73
Max	7.68	0.91	1.01
Homogeneity index	11206.9 (df 970)***	640.9 (df 26)***	568.5 (df 9)***
Separation	3.09	4.71	7.39
Strata	4.45	6.61	10.19
Reliability	0.91	0.96	0.98
Mean Infit mean-square	0.97	1.00	0.98
SD Infit mean-square	0.45	0.10	0.11
Mean Outfit mean-square	0.97	0.99	0.97
SD Outfit mean-square	0.51	0.14	0.14
N	971	27	10

*Huomio: Arvioija ja tehtävä facets on keskitetty nollan ympärille. Tuloksissa mukana kaikki suorittajat. Suluissa keskivirheet. *** p=0,001.*