KANSALLINEN KOULUTUKSEN ARVIOINTIKESKUS

SUOMEN KIELEN JA KIRJALLISUUDEN PERUSOPETUKSEN OPPIMISTULOS-ARVIOINTI 2019

Menetelmäliite

Jukka Marjanen

JULKAISUT 15:2020

SUOMEN KIELEN JA KIRJALLISUUDEN PERUSOPETUKSEN OPPIMISTULOSARVIOINTI 2019

Menetelmäliite

Jukka Marjanen

Kansallinen koulutuksen arviointikeskus
Julkaisut 15:2020

JULKAISIJA Kansallinen koulutuksen arviointikeskus

KANSI JA ULKOASU Juha Juvonen (org.) & Sirpa Ropponen (edit) TAITTO PunaMusta

ISBN 978-952-206-609-1 pdf
ISSN 2342-4184 (verkkojulkaisu)

PAINATUS PunaMusta Oy, Tampere

Sisällys

1 Johdanto

Tässä liitteessä esitellään suomen kielen ja kirjallisuuden perusopetuksen oppimistulosarvioinnissa käytetyt otanta- ja analyysimenetelmät. Menetelmien teoriataustaa ei kuvata yksityiskohtaisesti, mutta viitteet kattavampiin lähteisiin annetaan tekstissä. Liitteessä kuvataan lyhyesti myös aineistojen käsittelyn vaiheet.

2 Otanta

Otanta toteutettiin kahdessa vaiheessa. Ensin suoritettiin koulutason otanta, joka tehtiin satunnaisesti koulun sijaintikunnan aluehallintoviraston (tästä eteenpäin AVI-alue) ja tilastollisen kuntatyypin (kaupunki, taajama, maaseutu) mukaisista ositteista. Otoksen pohjana toimi Tilastokeskuksen oppilaitosrekisteri vuodelta 2017. Perusjoukko sisälsi kaikki ne peruskoulut sekä perus- ja lukioasteen koulut, joiden opetuskieli oli suomi ja joissa opetettiin vuosiluokkaa 9. Lapista otokseen poimittiin 30 % kouluista, ja muiden AVI-alueiden kouluista mukaan otettiin hieman alle viidennes. Lapin koulujen yliotostuksella pyrittiin parantamaan Lapin oppilaita koskevien tulosten tilastollista luotettavuutta. Otoskoulujen jakautuminen AVI-alueittain ja kuntatyypeittäin on esitetty taulukossa 1.

Taulukko 1. Oppimistulosarvioinnin koulutason otanta

	Koulujen lukumäärä otosrekisterissä				**Koulujen lukumäärä otoksessa**				**Otoskoulujen osuus kaikista kouluista**			
AVI-alue	**K***	**T**	**M**	**Yh-teensä**	**K**	**T**	**M**	**Yh-teensä**	**K**	**T**	**M**	**Yhteensä**
Etelä-Suomi	170	17	24	**211**	31	3	4	**38**	18 %	18 %	17 %	**18 %**
Lounais-Suomi	36	21	22	**79**	7	4	4	**15**	19 %	19 %	18 %	**19 %**
Itä-Suomi	37	8	35	**80**	7	1	7	**15**	19 %	13 %	20 %	**19 %**
Länsi-ja Sisä-Suomi	64	34	44	**142**	12	6	8	**26**	19 %	18 %	18 %	**18 %**
Pohjois-Suomi	34	15	26	**75**	6	3	5	**14**	18 %	20 %	19 %	**19 %**
Lappi	13	4	16	**33**	4	1	5	**10**	31 %	25 %	31 %	**30 %**
Yhteensä	**354**	**99**	**167**	**620**	**67**	**18**	**33**	**118**	**19 %**	**18 %**	**20 %**	**19 %**

*K = kaupunkikoulut, T = taajamakoulut, M = maaseutukoulut

Koulut saivat itse päättää, osallistuivatko kaikki vai osa yhdeksäsluokkalaisista oppimistulosarviointiin. Jos vain osa koulun oppilaista teki arvioinnin, osallistujat valittiin otokseen sukunimen mukaan aakkostetusta oppilaslistasta tasavälisellä otannalla. Koulukohtaisen otoksen koko määräytyi koulun oppilasmäärän mukaan seuraavasti:

- Jos koulussa oli 1–50 yhdeksäsluokkalaista, kaikki oppilaat kuuluivat otokseen.
- Jos koulussa oli 51–100 yhdeksäsluokkalaista, **joka toinen oppilas poimittiin otokseen.**
- Jos koulussa oli 101 yhdeksäsluokkalaista tai enemmän**, joka kolmas oppilas poimittiin otokseen.**

Taulukossa 2 on esitetty otosoppilaiden lukumäärät sukupuolittain ja AVI-alueittain. AVI-alueittaiset oppilasmäärät eivät välttämättä täsmää tyttöjen ja poikien yhteismäärän kanssa, sillä joiltakin oppilailta puuttui sukupuolitieto.

Taulukossa 2 otoksen oppilasmääriä verrataan arvioon kaikkien yhdeksäsluokkalaisten määrästä. Oppilasmäärien arvioimiseksi etsittiin Tilastokeskuksen vuoden 2019 esi- ja peruskoulutustilastosta (SVT 2020) suomenkielisten yhdeksäsluokkalaisten määrät kunnittain ja sukupuolittain eriteltynä. Kuntakohtaisten lukujen avulla saatiin laskettua yhdeksäsluokkalaisten tyttöjen ja poikien määrä AVI-alueittain. Näistä luvuista vähennettiin S2-oppilaiden (suomi toisena kielenä) määrä, joka arvioitiin Vipusesta (Vipunen 2020) saatujen, vuotta 2018 koskevien maakuntakohtaisten tietojen avulla. Tässä arvioinnissa oletettiin, että S2-oppilaiden prosentuaalinen osuus oli säilynyt AVI-alueittain ennallaan vuonna 2019. Tämä osuus poistettiin Tilastokeskuksen tietoihin perustuvista kokonaisoppilasmääristä, jolloin päädyttiin taulukossa 2 esitettyihin lukuihin.

Taulukko 2. Otosoppilaiden lukumäärät sukupuolittain ja AVI-alueittain

	Arvio oppilasmäärästä AVI-alueittain			**Oppilasmäärät arviointiaineistossa**			**Otosoppilaiden osuus arvioidusta oppilasmäärästä**		
AVI-alue	**Tytöt**	**Pojat**	**Yhteensä**	**Tytöt**	**Pojat**	**Yhteensä**	**Tytöt**	**Pojat**	**Yhteensä**
Etelä-Suomi	10140	10506	**20646**	1126	1033	**2376**	11 %	10 %	**12 %**
Lounais-Suomi	3234	3487	**6721**	382	331	**765**	12 %	9 %	**11 %**
Itä-Suomi	2678	2785	**5464**	177	182	**396**	7 %	7 %	**7 %**
Länsi-ja Sisä-Suomi	5879	6223	**12102**	626	642	**1419**	11 %	10 %	**12 %**
Pohjois-Suomi	2913	3158	**6070**	290	309	**560**	10 %	10 %	**9 %**
Lappi	879	933	**1812**	197	218	**528**	22 %	23 %	**29 %**
Yhteensä	**25723**	**27092**	**52815**	**2798**	**2715**	**6044**	**11 %**	**10 %**	**11 %**

Taulukosta 1 nähdään, että arvioinnin otos oli koulutasolla alueellisesti kattava ja tasapainoinen. Oppilastasolla otanta kattoi noin 11 % kaikista yhdeksäsluokkalaisista (taulukko 2). Lapista mukana oli 29 % oppilaista, kun taas Itä- ja Pohjois-Suomen AVI-alueiden oppilaat olivat aineistossa jonkin verran aliedustettuina. Nämä seikat pyrittiin huomioimaan tuloksissa käyttämällä oppilasmäärien kokonaisarvioita painokertoimina analyyseissa (luku 4).

3 Aineistojen esikäsittely, tarkistaminen ja yhdistäminen

Oppilaat käsittävä arviointiaineisto kerättiin Karvin sähköisessä arviointijärjestelmässä. Oppilaat suorittivat arvioinnin kahdessa osassa, ja aineistot yhdistettiin ennen datan jatkokäsittelyä. Yhdistämisen jälkeen oppilasaineistosta poistettiin ylimääräiset rivit, joita olivat esimerkiksi opettajien luomat testitunnukset ja aineistoon ennen arviointia luodut varatunnukset.

Seuraavaksi tarkistettiin koneellisesti, että sähköisen järjestelmän tuottama automaattinen pisteytys oli toteutunut oikein. Tämä tehtiin vertaamalla oppilaan raakavastauksia oikeaan vastausriviin, jolloin varmistettiin, että järjestelmä oli antanut pisteitä vain oikeista vastauksista. Samalla tarkastettiin, että tyhjäksi jätetyt vastaukset olivat jääneet tyhjiksi myös automaattisessa pisteytyksessä.

Oppilaiden äidinkielen ja kirjallisuuden päättöarvosanat saatiin Koski-tietovarannosta (https://www.oph.fi/fi/palvelut/koski-tietovaranto). Suurin osa Koski-aineistosta saatiin yhdistettyä KODA-aineistoon oppilaskohtaisten OID-tunnusten avulla. Osa KODA-aineiston OID-tunnuksista oli kuitenkin virheellisiä, ja näiden yhdistäminen tehtiin oppilaan etu- ja sukunimen sekä koulun nimen perusteella. Yhdistäminen onnistui lähes täydellisesti, sillä oppilasaineiston 6044 oppilaasta päättöarvosana jäi puuttumaan vain 34 oppilaalta. Myös nämä oppilaat olivat kuitenkin mukana kaikissa analyyseissa, jotka eivät koskeneet päättöarvosanoja.

Rehtori- ja opettajakyselyt toteutettiin Webropol -kyselytyökalulla (https://webropol.fi/). Oppilastunnusten luomisen yhteydessä opettajille luotiin yksilölliset opettajanumerot, joiden avulla opettajakyselyn aineisto yhdistettiin oppilasaineistoon. Rehtorikyselyn yhdistämisessä käytettiin koulun nimeä.

4 Sensorointi ja osioanalyysi

Opettajat vastasivat niiden arviointitehtävien pisteyttämisestä, joiden automaattinen tarkistaminen ei ollut mahdollista. Tällaisia olivat arvioinnin osan 1 pitkät kirjoitustehtävät ja osan 2 lyhyemmät avotehtävät. Pisteytyksen ja ennen kaikkea pisteytysohjeiden luotettavuuden arvioimiseksi oppilaiden vastauksista satunnaiset kymmenen prosenttia valittiin sensoroitaviksi. Vähintään yksi sensori pisteytti kaikki sensorointiin valikoituneet vastaukset, ja kolmanneksen vastauksista pisteytti myös toinen sensori. Sensoreina toimivat oppiaineen asiantuntijat. Sensoroitavia osioita oli yhteensä 24, sillä myös pitkien kirjoitustehtävien arviointikriteereitä nimitetään tässä osioiksi). Keskimäärin sensorointiin valikoitui 540 oppilaan vastaukset kustakin osiosta.

Sensoroinnin tavoitteena oli tarkastella, kuinka yhdenmukaisesti opettajat ja sensorit pisteyttivät oppilaiden antamat vastaukset. Arviointien yhdenmukaisuuden tarkastelemiseen on kehitetty useita tilastollisia tunnuslukuja, joista tunnetuin lienee Cohenin Kappa, jonka painotettua versiota käytettiin tässä arvioinnissa yhtenä pisteytysten yhdenmukaisuuden mittana (Cohen 1968). Koska Kappa voidaan laskea vain, jos vastausta kohden on tasan kaksi pistemäärää, laskettiin Kappa-arvot käyttäen opettajan pisteytystä ja yhtä sensorien antamista pisteytyksistä. Landisin ja Kochin (1977) antamien viitearvojen mukaisesti pisteytyksen johdonmukaisuutta pidettiin huomattavana (substantial), jos Kappa ylitti arvon 0,6, kohtuullisena (moderate), jos Kappa oli välillä 0,41–0,60, ja kelvollisena (fair), jos Kappa sai arvoja välillä 0,21–0,40. Cohenin Kappa-arvojen laskemiseen käytettiin R-ohjelmiston psych-laajennusosaa (Revelle 2019).

Yksi Kappa-luvun ongelmista on sen riippuvuus pisteytysten jakaumasta. Jos jokin pistekategoria on selvästi yleisempi kuin muut, Kappa aliarvioi pisteytysten todellista johdonmukaisuutta. Jos taas joidenkin vastausten pisteytykset poikkeavat toisistaan systemaattisesti arvioijien välillä, Kappa yliestimoi todellista yhdenmukaisuutta. Siksi sensorointiaineistosta laskettiin Kappojen lisäksi myös niin sanotut sisäkorrelaatiot (intraclass correlation), joiden saamiseksi osiokohtaiseen aineistoon sovitettiin ensin kaavan (1) mukainen yleistetty lineaarinen sekamalli (generalized linear mixed model) (Nelson & Edwards 2015). Pisteytyksen luotettavuuden mittana käytetty sisäkorrelaatio laskettiin tämän mallin varianssiestimaattien avulla kaavan (2) mukaisesti.

$$\Pr\left(Y_{is} \leq c \middle| u_i, v_j\right) = \Phi(\alpha_c - (u_i + v_s)) \qquad (1)$$

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + 1} \qquad (2)$$

Kaavassa (1) mallinnetaan todennäköisyyttä sille, että pisteyttäjä *s* (opettaja tai sensori) antaa oppilaan *i* vastaukselle korkeintaan pistemäärän *c*. Todennäköisyys riippuu oppilaan vastauksen tasosta u_i ja pisteyttäjän taipumuksesta antaa korkeita tai matalia pistemääriä v_j. Oppilaiden tason u_i oletetaan noudattavan normaalijakaumaa, jonka keskiarvo on 0 ja varianssi σ_u^2. Vastaavasti pisteyttäjien vaikutuksen oletetaan noudattavan normaalijakaumaa, jonka keskiarvo on nolla ja varianssi σ_v^2. Symboli Φ tarkoittaa standardoidun normaalijakauman kertymäfunktiota, ja α_c kontrolloi pistekategorian *c* yleisyyttä.

Ajatuksena on, että jos pisteyttäjät ovat omaksuneet yhdenmukaisen linjan, heidän välinen vaihtelunsa on pisteytyksessä pientä, ja pistemäärien vaihtelu on lähinnä oppilaiden osaamiseen liittyvää vaihtelua. Tällöin ICC lähestyy arvoa 1, joka on reliabiliteetin teoreettinen maksimi. Jos taas pisteytykset vaihtelevat huomattavasti opettajien ja sensorien välillä, pisteyttäjiin liittyvä varianssi kasvaa, ja ICC pienenee kohti arvoa 0. Portneyn (2020) antamien nyrkkisääntöjen mukaisesti, hyvänä (good) reliabiliteettina pidettiin tässä arvioinnissa ICC arvoa 0,75 ja kohtuullisena (moderate) arvoja välillä 0,5–0,74. GLMM-mallinnuksessa käytettiin Mplus 8 -ohjelmistoa (Muthén & Muthén 1998–2017).

Sensoroiduista osioista 9 ylsi hyvään tasoon ICC-luvun perusteella ja huomattavaan tasoon Kappa-arvon perusteella. Yhteensä 21 osiota ylsi molempien tunnuslukujen valossa vähintään kohtuulliseen reliabiliteettiin. Vain kaksi osiota jäi tämän tason alle Kappa-luvun mukaan (alimmillaan 0,34) ja yksi osio ICC-luvun perusteella (alimmillaan 0,495). Heikoimman Kappa-arvon saanut osio poistettiin varsinaisissa analyyseissa käytettyjen osioiden joukosta. Muilta osin avotehtävien pisteytyksen luotettavuutta voidaan pitää vähintään kohtuullisena.

Sensoroinnin lisäksi, arviointitehtävien toimivuutta tarkasteltiin klassisen osioanalyysin ja niin sanotun IRT-analyysin (Item Response Theory) avulla (esim. deAyala 2009). Varsinaisissa analyyseissa käytettyjen osioiden joukosta poistettiin tässä vaiheessa ne osiot, joiden erottelukyky oli heikko. Erottelukyvyn mittana käytettiin IRT-analyysin erottelukykyparametria sekä osion korrelaatiota muista osioista laskettuun yhteispistemäärään (ns. item-rest korrelaatio) (esim. Nunnally 1978).

Koska varsinaisissa analyyseissa hyödynnettiin IRT-analyysia laajasti, poistettiin aineistosta myös ne osiot, jotka eivät sopineet yhteen IRT-mallin kanssa. Mallin ja aineiston yhteensopivuutta arvioitiin PV-Q1-tunnusluvun (Chalmers & Ng 2017) ja graafisten tarkastelujen avulla. Koska arviointiaineisto on suuri, heikosti malliin sopivia osioita ei poistettu suoraan PV-Q1 lukuun liittyvän p-arvon perusteella. Sen sijaan osioiden joukosta etsittiin ne, joiden PV-Q1 luku oli huomattavasti suurempi kuin muiden osioiden. Graafisissa analyyseissa käytettiin yleistettyjä additiivisia malleja (generalized additive model). Osioanalyysit toteutettiin R-ohjelmistolla, ja IRT-mallinnuksessa hyödynnettiin sen mirt-laajennusosaa (Chalmers 2012).

Osioanalyysien perusteella aineistosta poistettiin kaikkiaan 7 osiota.

Samaan tekstiin, kuvaan tai äänikatkelmaan liittyvien osioiden vastaukset korreloivat usein keskenään, vaikka niiden taustalla oleva osaaminen olisi huomioitu (ns. residuaalikorrelaatiot). Tämä voi aiheuttaa vääristymää IRT-analyyseissa etenkin, jos mallinnuksessa halutaan käyttää yksiulotteista faktorimallia. Residuaalikorrelaatioiden minimoimiseksi samaan tehtävään liittyviä osioita yhdistettiin ennen varsinaisia analyyseja summamuuttujiksi, joita käsiteltiin jatkoanalyyseissa osioina. Yhdistelmäosioiden sopivuus IRT-malliin tarkistettiin vielä ennen varsinaisia analyyseja edellä kuvatuilla menetelmillä.

5 Analyysimenetelmät

Tässä arvioinnissa käytetyt menetelmät vaihtelivat jonkin verran sen mukaan, olivatko kyseessä alueellisiin eroihin ja oppilaan sukupuoleen vai esimerkiksi oppilaiden asenteisiin, harrastuneisuuteen tai muihin oppilaita ryhmitteleviin taustatekijöihin liittyvät analyysit. Ensin mainituissa analyyseissa käytetyt menetelmät kuvataan luvussa 4.1 ja jälkimmäiset luvussa 4.2.

5.1 Tyttöjen ja poikien väliset osaamiserot ja erot AVI-alueiden välillä

Tulkintatehtäviin liittyvät analyysit esitellään luvussa 3.1.1, tuottamistehtäviä koskevat analyysit luvussa 3.1.2 ja arvioinnin kokonaistuloksen analysoimisessa käytetyt menetelmät luvussa 3.1.3. Ellei toisin mainita, kaikki näissä luvuissa kuvatut analyysit toteutettiin Mplus-ohjelmistolla.

5.1.1 Tulkintatehtävät

Analyysien ensimmäisessä vaiheessa yksittäisistä osioista muodostettujen yhdistelmäosioiden faktorilataukset ja osioiden vaikeutta kuvaavat tunnusluvut määritettiin IRT-analyysin avulla. Analyysissa käytetty malli on kuvattu kaavassa (3):

$$\Pr\left(Y_{ij} \leq c \middle| \theta_{1i}, \theta_{2i}\right) = \Phi(\tau_{jc} - \left(\lambda_{1j}\theta_{1i} + \lambda_{2j}\theta_{2i}\right)) \quad (3)$$

Kaavan (3) mukaan todennäköisyys sille, että oppilas *i* saa osiosta *j* korkeintaan pistemäärän c, riippuu kategorian *c* vaikeustasosta, osion erottelukyvystä (eli faktorilatauksista) ulottuvuuksilla 1 ja 2 (λ_{1j} ja λ_{2j}) sekä oppilaan osaamisesta ulottuvuuksilla 1 ja 2 (θ_{1i} ja θ_{2i}). Symboli Φ tarkoittaa standardoidun normaalijakauman kertymäfunktiota.

Mallinnuksessa käytettiin niin sanottua bi-faktorimallia, jossa kaikkien osioiden määritellään olevan yhteydessä yleiseen osaamisulottuvuuteen 1 ($\lambda_{1j} > 0$ kaikilla osioilla) ja joidenkin osioiden ajatellaan mittaavan myös muuta osaamista ($\lambda_{2j} > 0$ joillain osioilla). Esimerkiksi monimediaisten tekstien tulkitsemistehtävissä malliin muodostettiin yleisfaktorin ohelle myös avotehtävistä koostuva faktori, mutta tarvittaessa lisäfaktoreita voi olla useampia kuin yksi.

Tässä arvioinnissa lisäfaktorien tarkoituksena oli kontrolloida muiden osaamisulottuvuuksien kuin pääfaktorin vaikutusta oppilaiden vastauksiin, eikä niihin liittyviä tuloksia raportoida tekstissä tilan säästämiseksi. Osioparametrien estimoinnissa käytettiin suurimman uskottavuuden (maximum likelihood) menetelmää, ja mallin identifioimiseksi osaamisulottuvuuksien keskiarvoksi määritettiin 0 ja varianssiksi 1. Yleisfaktorin ja lisäfaktori(e)n välisiksi korrelaatioiksi määritettiin 0.

Mallinnuksen seuraavassa vaiheessa kullekin oppilaalle simuloitiin useita arvioita heidän osaamisestaan niin sanottua plausible values (PV-arvo) -menetelmää käyttäen (von Davier, Gonzalez & Mislevy 2009). Tässä arvioinnissa oppilaille tuotettiin sata PV-arvoa kultakin osa-alueelta. PV-arvot poimittiin satunnaisesti todennäköisyysjakaumasta (ns. posteriorijakauma), jonka keskiarvo ja hajonta perustuivat yhtäältä oppilaan suorituksiin osaamistehtävissä, toisaalta rakenneyhtälömallinnuksen (esim. Muthén & Asparouhov 2018, Bollen 1989) perusteella muodostettuun ennusteeseen oppilaan keskimääräisestä osaamisesta (ns. priorijakauma). Koska mallinnuksen seuraavissa vaiheissa

käytettiin monitasomallinnusta, myös rakenneyhtälömallinnus tehtiin monitasoisena. Siten myös koulun vaikutukselle poimittiin jokaiselta osa-alueelta omat PV-arvonsa. Käytetty malli on kuvattu kaavoissa (4–8):

$$\Pr\left(Y_{ijl} \leq c \middle| \theta_{1i}, \theta_{2i}\right) = \Phi(\hat{\tau}_{jc} - \left(\hat{\lambda}_{1j}\theta_{1il} + \hat{\lambda}_{2j}\theta_{2il}\right)) \quad (4)$$

$$\theta_{1il} = \beta_{10l} + \sum \beta_{11r} d_{1il} + \sum \beta_{11a} x_{1il} + \varsigma_{1il} \quad (5)$$

$$\theta_{2il} = \beta_{20l} + \sum \beta_{22r} d_{2il} + \sum \beta_{22a} x_{2il} + \varsigma_{2il} \quad (6)$$

$$\beta_{10l} = \gamma_{100} + \eta_{1l} \quad (7)$$

$$\beta_{20l} = \gamma_{200} + \eta_{2l} \quad (8)$$

Malli muistuttaa kaavan (3) mallia, mutta nyt koulun *l* oppilaan *i* osaamista selitetään koulun sijaintikunnan AVI-alueen ja oppilaan sukupuolen yhdistelmää indikoivilla dummy-muuttujilla (d-muuttujat) sekä oppilaan äidinkielen ja kirjallisuuden päättöarvosanaa indikoivilla dummy-muuttujilla (x-muuttujat). Oppilaan sukupuoli, koulun sijaintikunnan AVI-alue ja suomen kielen ja kirjallisuuden päättöarvosana sisällytettiin malliin, koska ne ovat keskeisiä seuraavien vaiheiden analyyseissa, joissa käytetään mallin pohjalta muodostettuja PV-arvoja. Arvioinnissa oli mukana kouluja kuudelta AVI-alueelta, ja sukupuoli-muuttuja oli kolmiluokkainen (tyttö, poika, puuttuva tieto). Siten sukupuolen ja AVI-alueen yhdistelmää kuvaavia dummy-muuttujia oli mallissa 17 (koska ensimmäistä ryhmää käytettiin referenssiryhmänä). Päättöarvosanoissa referenssiryhmänä käytettiin arvosanaa kahdeksan, ja dummy-muuttujia oli yhteensä 6.

Dummy-muuttujien lisäksi malli sisälsi koulun *l* vaikutusta kuvaavat termit η_{1l} ja η_{2l}. Niiden oletettiin noudattavan normaalijakaumia, joiden keskiarvot ovat 0 ja varianssit $\sigma^2_{\eta 1}$ ja $\sigma^2_{\eta 2}$. Oppilasvarianssia kuvaavien jäännöstermien ς_{1il} ja ς_{2il} oletettiin vastaavasti noudattavan normaalijakaumia, joiden keskiarvot ovat 0 ja varianssit $\sigma^2_{\varsigma 1}$ ja $\sigma^2_{\varsigma 2}$.

Kaavojen (7) ja (8) termit γ_{100} ja γ_{200} kuvaavat referenssiryhmän keskiarvoja (tässä tapauksessa ne Etelä-Suomen tytöt, joiden äidinkielen ja kirjallisuuden päättöarvosana oli 8) osaamisulottuvuuksilla 1 ja 2. β-termit puolestaan kertovat oppilasryhmän tai saadun päättöarvosanan keskimääräisen eron referenssiryhmän keskimääräiseen tulokseen.

Osioparametrit on kaavassa (4) merkitty symboleilla $\hat{\lambda}$ ja $\hat{\tau}$ eikä λ ja τ. Tällä kuvataan sitä, että mallinnuksessa käytettiin edellisestä analyysista saatuja osioparametrien estimaatteja sen sijaan, että osioparametrit olisi estimoitu vapaasti tämän vaiheen analyysimallissa. Mallinnuksessa ja PV-arvojen muodostamisessa käytettiin Bayes-estimointia (esim. Palomo, Dunson & Bollen 2007).

Vaikka PV-arvot eivät ole parhaita arvioita yksittäisten oppilaiden tai koulujen tasosta, niiden avulla voidaan tuottaa parempi kuva koko oppilasjoukon tai koulujen osaamisen jakautumisesta kuin esimerkiksi summapistemääriä tai ratkaisuprosentteja käyttämällä. (von Davier, Gonzalez & Mislevy 2009.) Toisin kuin suoraan osiopistemääristä lasketut summat tai ratkaisuprosentit, PV-arvot eivät myöskään sisällä mittavirhettä, jolloin muun muassa koulujen osuus tulosten kokonaisvaihtelusta

tulee arvioiduksi tarkemmin. Tätä niin sanottua sisäkorrelaatiota käytetään useissa oppimistuloksia koskevissa arvioinneissa keskeisenä koulutuksellisen tasa-arvon mittarina.

Analyysien seuraavassa vaiheessa PV-arvoja käytettiin ryhmäkohtaisten osaamiskeskiarvojen ja varianssien estimoimiseen. Näiden laskemiseen käytettiin moniryhmäistä (multiplegroup) monitasomallia (Asparouhov & Muthén 2012) (Kaavat 9 ja 10). Koska koulujen määrä oli pieni eri ryhmissä (esim. AVI-alueet), erillisten kouluvarianssien estimointi olisi ollut epätarkkaa. Siksi koulutason varianssi määritettiin samaksi kaikissa oppilasryhmissä.

$$y_{ilr} = PV_l + PV_{ilr} \qquad (9)$$

$$y_{ilr} = \bar{y}_r + u_l + \varepsilon_{ilr} \qquad (10)$$

Kaavassa (9) PV_l tarkoittaa koulun *l* plausible value-arvoa ja PV_{ilr} koulun *l* ryhmään *r* kuuluvan oppilaan i plausible value -arvoa. Kaavassa (10) $\bar{y}_r$ on ryhmän *r* keskiarvo, u_l koulun *l* efekti ja ε_{ilr} tarkoittaa koulun *l* ryhmään *r* kuuluvan oppilaan *i* eroa koulun keskiarvosta. Termien u_l ja ε_{ilr} oletetaan olevan normaalijakautuneita keskiarvolla 0 ja variansseilla σ_u^2 ja $\sigma_{\varepsilon r}^2$. Koska koulutason vaihtelu määritettiin samaksi kaikissa ryhmissä, kouluefektin vaihtelua kuvaavassa termissä ei ole ryhmään liittyvää alaindeksiä *r*.
Koulut saivat itse päättää, osallistuivatko kaikki vai vain osa niiden oppilaista arviointiin. Tästä syystä ryhmäkohtaisten keskiarvojen estimoinnissa käytettiin painokertoimia, jotta kunkin koulun oppilaat tulisivat edustetuksi analyyseissa oikeassa suhteessa. Painokertoimet määritettiin jakamalla koulun todellinen oppilasmäärä arviointiin osallistuneiden oppilaiden määrällä (kaava 11). Koulun todellinen oppilasmäärä saatiin Koski-tietokannasta. Ennen analyyseja painokertoimet normalisoitiin siten, että niiden summa vastasi arviointiin osallistuneiden oppilaiden kokonaismäärää (kaava 12). Koska mallinnuksessa käytettiin painokertoimia, analyysissa käytettiin MLR-estimaattoria (Mplus-ohjelman oletusarvo painotetuille analyyseille).

$$w_{il} = \frac{N_l}{n_l} \qquad (11)$$

$$w_{il}^* = w_{il} * \frac{\sum n_l}{\sum w_{il}} \qquad (12)$$

Kaavoissa (11) ja (12) w_{il} tarkoittaa koulun *l* oppilaan *i* normalisoimatonta painokerrointa, ja w_{il}^* on saman oppilaan painokerroin normalisoituna. N_l tarkoittaa koulun *l* oppilasmäärää Koski-tietovarannossa ja n_l arviointiin osallistuneiden oppilaiden lukumäärää.
Raportin tulosluvussa 2 ryhmäkohtaisia keskiarvoja verrataan kansalliseen keskiarvoon ja ajoittain myös toisiinsa. Kansallinen keskiarvo laskettiin ryhmäkohtaisten keskiarvojen painotettuna keskiarvona kaavan (13) mukaisesti:

$$\hat{\bar{y}} = \frac{\sum w_r \hat{\bar{y}}_r}{\sum w_r} \qquad (13)$$

Kaavassa (13) w_r tarkoittaa ryhmän *r* (esim. Lapin tytöt) saamaa painokerrointa. Painokertoimia käytettiin, jotta esimerkiksi Lapin oppilaiden yliedustus ja Itä-Suomen oppilaiden hienoinen aliedustus tulisivat huomioiduksi tuloksissa. Painokertoimina käytettiin luvun 1 taulukossa 2 esitettyjä arvioita oppilaiden kokonaismääristä.

Yksittäisen ryhmän ero kokonaiskeskiarvosta saatiin vähentämällä ryhmäkohtaisesta keskiarvosta $\bar{y}_r$ kokonaiskeskiarvon estimaatti (kaava 14).

$$\hat{d}_r = \hat{\bar{y}}_r - \hat{\bar{y}} \quad (14)$$

Tyttöjen ja poikien välinen ero puolestaan laskettiin tyttöjen ja poikien painotettujen keskiarvojen erotuksena kaava (15):

$$\hat{d}_{sukup} = \frac{\sum w_{rt}\hat{\bar{y}}_{rt}}{\sum w_{rt}} - \frac{\sum w_{rp}\hat{\bar{y}}_{rp}}{\sum w_{rp}} \quad (15)$$

Kaavassa (15) w_{rt} tarkoittaa tyttöjen ryhmän *r* (esim. Lapin tytöt) painokerrointa ja $\hat{\bar{y}}_{rt}$ kyseisen ryhmän osaamiskeskiarvoa. Poikien vastaavat luvut on merkitty kaavaan alaindekseillä *rp*.

Koska analyyseissa käytettiin PV-arvoja, mallinnukset tehtiin 100 kertaa eli jokaiselle PV-arvolle erikseen. Parametrien lopulliset estimaatit $\hat{p}$ ovat yksittäisistä PV-arvoista $\hat{p}_i$ laskettujen estimaattien keskiarvoja (kaava 16). (OECD 2009, 118–119; Nissinen, Rautopuro & Puhakka 2018.) Parametrilla tarkoitetaan tässä esimerkiksi yksittäisen ryhmän eroa kokonaiskeskiarvosta (kaava 14) tai tyttöjen ja poikien välistä osaamiseroa (kaava 15).

$$\hat{p} = \frac{1}{100}\sum \hat{p}_i \quad (16)$$

Yksittäisiä PV-arvoja koskevat mallinnukset toteutettiin Mplus 8 -ohjelmistolla, joka myös tuotti hajontaestimaatit (eli keskivirheet) parametrien arvoille ($\sigma_{\hat{p}_i}$). Arvio varsinaisten parametriestimaattien keskivirheestä saatiin lisäämällä yksittäisistä PV-arvoista laskettujen hajontaestimaattien keskiarvoon lisätermi, joka muodostuu yksittäisten parametriestimaattien välisestä hajonnasta (kaava 17): (OECD 2009, 118–119; Nissinen, Rautopuro & Puhakka 2018.)

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{1}{100}\sum \sigma^2_{\hat{p}_i} + \left(\left(1 + \frac{1}{100}\right)\frac{1}{99}(\hat{p}_i - \hat{p})^2\right)} \quad (17)$$

Parametrin tilastollista merkitsevyyttä kuvaava p-arvo määritettiin parametrin ja sen keskivirheen avulla kaavan (18) mukaisesti. Kaavassa pystyviivat || tarkoittavat itseisarvoa.

$$p - arvo = \left(1 - \Phi\left(\left|\frac{\hat{p}}{\sigma_{\hat{p}}}\right|\right)\right) * 2 \quad (18)$$

Varsinaisessa raportissa parametrien tilastollinen merkitsevyys on kielennetty seuraavasti:
p < 0,001 tilastollisesti erittäin merkitsevä
p < 0,01 tilastollisesti merkitsevä
p < 0,05 tilastollisesti melkein merkitsevä

Tilastollinen merkitsevyys ei vielä kerro erojen käytännön merkittävyydestä, sillä suurilla aineistoilla pienetkin erot ovat lähes väistämättä tilastollisesti merkitseviä. Siksi tämän raportin tulosluvuissa ilmoitetaan myös erojen efektikoot. Efektikoko saatiin jakamalla tarkasteltavana olevan eron estimaatti ryhmäkohtaisista kokonaisvariansseista lasketulla jäännöshajonnalla kaavat (19) – (20). Näin laskettu efektikoko muistuttaa Hedgesin (2009) monitasomalleille johtamaa efektikoon mittaa δ_T. Kaavassa (19) ero $\hat{d}$ on joko yksittäisen ryhmän ero kansallisesta keskiarvosta, kahden AVI-alueen keskiarvojen erotus tai tyttöjen ja poikien keskiarvojen välinen ero. Vaikka ero koskisi vain yhtä tai kahta ryhmää, kaavan (19) jakajassa käytettiin aina kaikkien ryhmien tuloksiin perustuvaa hajontaestimaattia (kaava 20).

$$\hat{E} = \frac{\hat{d}}{\hat{\sigma}_R} \qquad (19)$$

$$\hat{\sigma}_{Res} = \sqrt{\frac{\sum(w_r - 1)\hat{\sigma}_r^2}{(\sum w_r) - 1}} \qquad (20)$$

$$\hat{\sigma}_r^2 = \hat{\sigma}_u^2 + \hat{\sigma}_{\varepsilon r}^2 \qquad (21)$$

Kaavan (20) ryhmäkohtaiset varianssit $\hat{\sigma}_r^2$ saatiin lisäämällä kunkin ryhmän oppilasvarianssiin koulutason varianssiestimaatti $\hat{\sigma}_u^2$ (kaava 21). Koulutason varianssi oli kaikille ryhmille sama, kuten edellä on kuvattu.

Arviointiraportissa erojen efektikoot on kielennetty seuraavasti:
<0,20 pieni ero
0,50 keskisuuri ero
> 0,80 suuri ero

Efektikokojen lisäksi ryhmien välisten erojen merkittävyyttä tarkastellaan selitysosuuksien kautta. Selitysosuudet laskettiin kaavan mukaisesti (22):

$$\hat{R}^2 = \frac{\hat{\sigma}_{Tot}^2 - \hat{\sigma}_{Res}^2}{\hat{\sigma}_{Tot}^2} \qquad (22)$$

$$\hat{\sigma}_{Tot} = \sqrt{\frac{\sum(w_r - 1)\hat{\sigma}_r^2 + \sum w_r\left(\hat{\bar{y}}_r - \hat{\bar{y}}\right)^2}{(\sum w_r) - 1}} \qquad (23)$$

Kaavassa (22) esiintyvä kokonaisvarianssi $\hat{\sigma}_{Tot}^2$ laskettiin ryhmäkohtaisista variansseista ja keskiarvoista kaavalla (23).

Edellä kuvatuissa analyyseissa oppilaiden osaamista kuvaavat PV-arvot olivat asteikolla, jonka keskiarvo on (likimain) 0 ja keskihajonta 1. Tulkittavuuden helpottamiseksi tulokset esitetään raportissa kuitenkin PISA-arvioinneista tutulla asteikolla sekä kouluarvosanoiksi muunnettuina. PISA-asteikolla osaamispisteiden kansallinen keskiarvo on aina 500 ja keskihajonta 100. Arvosana-asteikolla koko aineiston keskiarvo on 7,9 ja keskihajonta 1,2. Nämä arvot laskettiin äidinkielen ja kirjallisuuden

päättöarvosanoista, jotka saatiin Koski-tietovarannosta. Alkuperäisellä asteikolla ilmaistut ryhmäkohtaiset keskiarvot muunnettiin uusille asteikoille kaavan (24) mukaisesti:

$$\hat{\bar{y}}_r^* = \frac{\sigma_{PISA/arv}}{\hat{\sigma}_{Tot}} (\hat{\bar{y}}_r - \hat{\bar{y}}) + \bar{y}_{PISA/arv} \qquad (24)$$

Kaavassa (24) $\sigma_{PISA/arv}$ tarkoittaa PISA-asteikon tai kouluarvosana-asteikon keskihajontaa ja $\bar{y}_{PISA/arv}$ niiden keskiarvoa. $\hat{\sigma}_{Tot}$ tarkoittaa kaavan (23) mukaista oppilaiden osaamispistemäärien kokonaishajontaa.

Myös ryhmäkohtaiset kokonaishajonnat muunnettiin PISA-asteikolle. Tähän käytettiin kaavaa (25):

$$\hat{\sigma}_r^{2*} = \frac{\hat{\sigma}_r^2 - \hat{\sigma}_{Tot}^2}{\hat{\sigma}_{Tot}^2} * 100+100 \quad (25)$$

Koulujen välisiä eroja tarkasteltiin sisäkorrelaatioiden avulla (kaava 26). Sisäkorrelaatioiden laskemiseksi yksittäisiin PV-arvoihin sovitettiin malli, jossa estimoitiin ainoastaan koulutason ja oppilastason varianssit ($\hat{\sigma}_u^2$ ja $\hat{\sigma}_\varepsilon^2$). Aineistoa käsiteltiin tässä analyysissa yhtenä kokonaisuutena, eikä oppilaita ryhmitelty AVI-alueen tai sukupuolen mukaan. Raportissa esitetyt sisäkorrelaatiot ovat yksittäisistä PV-arvoista laskettujen sisäkorrelaatioiden keskiarvoja.

$$\widehat{ICC}_\iota = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2} \quad (26)$$

Ryhmäkohtaisten keskiarvojen välisiin eroihin keskittyvien analyysien lisäksi raportissa tarkastellaan oppilaiden osaamisen jakaumista myös tarkemmin. Näissä analyyseissa oppilaat jaettiin osaamisluokkiin PV-arvojen perusteella. Osaamisluokkien rajat perustuivat kaavojen (27) ja (28) mukaiseen monitasoregressioanalyysiin, jossa oppilaiden arvioinnissa osoittamaa osaamista (PV-arvot) selitettiin äidinkielen ja kirjallisuuden päättöarvosanoja kuvaavilla dummy-muuttujilla (x-muuttujat). Koska arvosanoja 4 oli koko aineistossa vain muutama, ne yhdistettiin arvosanakategoriaan 5.

$$y_{il} = \beta_{0l} + \sum \beta_{1a} x_{il} + \varepsilon_{il} \qquad (27)$$

$$\beta_{0l} = \gamma_{00} + u_l \qquad (28)$$

Koska analyysin referenssikategoriana käytettiin arvosanaa kahdeksan, mallin vakiotermi γ_{00} kuvaa arvosanaa kahdeksan vastaavaa keskimääräistä osaamista. Muita arvosanoja vastaavat osaamiskeskiarvot saatiin lisäämällä vakiotermiin kuhunkin arvosanaan liittyvä β-termi. Lopullisina luokkarajoina käytettiin yksittäisistä PV-arvoista laskettujen raja-arvojen keskiarvoja. Raja-arvojen avulla oppilaiden osaamispisteet (eli PV-arvot) voitiin jakaa seitsemään ryhmään seuraavasti:

1. Arvosanaa 5–tai vastaava sitä heikompi osaamistaso
2. Arvosanoja 5–6–vastaava osaamistaso
3. … 6–7–…
4. … 7–8–…
5. … 8–9–…

6. … 9–10- …
7. Arvosanaa 10 vastaava osaamistaso

Raportissa esitetään, kuinka suuri osuus oppilaista ylsi kuhunkin näistä ryhmistä paitsi kansallisella tasolla myös eri oppilasryhmissä. Näiden tarkastelujen avulla voidaan tuottaa tarkempaa tietoa osaamisen jakautumisesta eri oppilasryhmissä kuin pelkkiä keskiarvoja tarkastelemalla. Koska tulokset perustuvat PV-arvoihin, raportissa esitetyt prosenttiosuudet ovat yksittäisiin PV-arvoihin perustuvien prosenttiosuuksien keskiarvoja.

5.1.2 Tuottamistehtävät

Pitkien kirjoitustehtävien (arvioinnin 1. osa) analysoimisessa ei käytetty IRT-analyyseja eikä niihin perustuvia PV-arvoja. Sen sijaan oppilaiden osaamisen mittana käytettiin suoraan opettajien antamia pistemääriä. Arviointi sisälsi kaksi pitkää kirjoitustehtävää, joissa molemmissa opettajat pisteyttivät oppilaiden tekstit kriteeriperusteisesti. Ensimmäisessä kirjoitustehtävässä kriteereitä oli 5 ja toisessa 4. Kriteerikohtaiset maksimipistemäärät vaihtelivat välillä 2–4.

Analyyseissa kirjoitustehtäviä käsiteltiin sekä erikseen että yhtenä kokonaisuutena. Tehtäväkohtaisissa analyyseissa osaamispistemääränä käytettiin yksittäisten kriteerien pistemääristä laskettua summaa. Kun kirjoitustehtäviä käsiteltiin yhdessä, molempien tehtävien pistemäärät standardoitiin ensin vähentämällä oppilaskohtaisista pistemääristä koko oppilasjoukon pistekeskiarvo ja jakamalla nämä luvut koko oppilasjoukon pistemäärien keskihajonnalla. Tämän jälkeen oppilaille laskettiin kahden kirjoitustehtävän standardoitujen pistemäärien keskiarvo, jota käytettiin analyyseissa. Keskiarvon laskemisessa käytettiin standardoituja pistemääriä, jotta molemmat tehtävät saisivat yhteistuloksessa saman painoarvon.

Seuraavaksi aineistoon sovitettiin kaavan (10) mukainen moniryhmäinen (multiplegroup) monitasomalli, jossa selitettävänä muuttujana oli joko yksittäisen kirjoitustehtävän summapistemäärä tai kahden tehtävän standardoiduista pistemääristä laskettu keskiarvo. Ryhmäkohtaisten keskiarvojen, kansallisen kokonaiskeskiarvon sekä ryhmäkeskiarvojen välisten erojen laskemiseen käytettiin kaavoja (11) (15). Koska analyysissa ei käytetty PV-arvoja, parametrien varsinaiset estimaatit saatiin suoraan yhdestä analyysista eikä useiden yksittäisiin PV-arvoihin perustuvien estimaattien keskiarvoista. Myös parametrien keskivirheet saatiin suoraan ohjelmistosta, eikä niiden laskemiseen tarvittu kaavoja (16) ja (17). Muut kirjoitustehtäviä koskevat tunnusluvut laskettiin edellisestä luvusta tutuilla kaavoilla (18) (26).

Suurin ero edellisessä luvussa kuvattuihin analyyseihin oli osaamisjakaumia koskevissa tarkasteluissa. Osaamisluokkien rajat muodostettiin kaavoissa (27) ja (28) kuvatun mallin perusteella, mutta PV-arvoja ei ollut käytettävissä oppilaiden luokittelemiseen. Oppilaita ei myöskään luokiteltu suoraan opettajien antamien pistemäärien perusteella, vaan oppilaiden osaamisjakaumia simuloitiin ryhmäkeskiarvojen ja keskihajontojen avulla. Simuloinnissa oletettiin, että kunkin oppilasryhmän osaamisjakauma noudattaa normaalijakaumaa, jonka keskiarvo on $\hat{\bar{y}}_r$ ja keskihajonta $\hat{\sigma}_r^2$. Kutakin ryhmää esittävästä jakaumasta poimittiin satunnaisesti sama määrä arvoja kuin kyseisen ryhmän oppilaita oli arviointiaineistossa. Nämä arvot toimivat sitten osaamisen mittana, ja ne luokiteltiin arvosanarajojen mukaisesti. Simulaatio toistettiin 100 kertaa, ja raportissa esitetään, kuinka suuri osuus eri ryhmien oppilaista kuului keskimäärin kuhunkin osaamisluokkaan näissä simulaatioissa.

5.1.3 Arvioinnin kokonaistulos

Oppilaiden kokonaistulosta kuvaava pistemäärä muodostettiin tekstien tulkitsemisentehtävien kahden osa-alueen tehtävistä (mediatekstit ja kirjallisuus), kielitiedon tehtävistä sekä kahden tuottamistehtävän yhteistulosta kuvaavasta pistemäärästä (ks. edellinen luku). Tulkintatehtävistä ja kielitiedon tehtävistä laskettiin ensin osa-alueittain 100 PV-arvon keskiarvot jokaiselle oppilaalle, minkä jälkeen nämä luvut standardoitiin, jotta eri osa-alueiden tulokset olisivat samalla mitta-asteikolla. Varsinaisissa analyyseissa oppilaiden kokonaisosaamisen mittana käytettiin keskiarvoa, joka laskettiin tulkintatehtävien kahden osa-alueen ja kielitiedon tehtävien standardoiduista pisteistä sekä kahden tuottamistehtävän standardoidusta yhteispistemäärästä. Siten tulkintatehtävien eri osa-alueet, kielitiedon tehtävät sekä tuottamistehtävät saivat kukin saman painoarvon (¼) oppilaiden kokonaistuloksessa. Kokonaistuloksen analysoimisessa käytettiin samoja menetelmiä kuin tuottamistehtävien analysoimisessa (ks. edellinen luku).

5.2 Osaaminen suhteessa taustamuuttujiin

Oppilaiden osaamista suhteessa joihinkin oppilastason taustamuuttujiin tarkasteltiin perinteisten yksitasoregressiomallien avulla. Kun kyseessä ovat oppilastason selittävät muuttujat, niiden tilastollinen merkitsevyys tulee estimoiduksi oikein, vaikka monitasomallinnusta ei käytettäisi. Selitettävänä muuttujana oli useimmissa tapauksissa arvioinnin kokonaistulos, mutta joissain analyyseissa selitettiin myös arvioinnin eri sisältöalueiden tuloksia. Tällöin käytettiin PV-arvoja, ja lopulliset tulokset ovat keskiarvo 100 yksittäisen analyysin tuloksista.

Osa taustamuuttujista oli ryhmitteleviä, kuten vanhempien koulutustausta, oppilaan jatko-opintosuunnitelmat tai lukitausta. Tällöin aineistoon luotiin oppilasryhmiä kuvaavat dummy-muuttujat, joita käytettiin analyyseissa selittävinä muuttujina. Muita taustamuuttujia, kuten oppilaiden asenteet ja harrastuneisuus, käsiteltiin analyyseissa yksinkertaisuuden vuoksi jatkuvina muuttujina, vaikka kaikki niistä eivät jatkuvia tarkkaan ottaen olisi. Jos taustamuuttujan yhteys oppilaiden osaamiseen ei ollut suoraviivainen, malliin lisättiin myös kvadraattinen termi (muuttujan arvot korotettuna toiseen potenssiin).

Analyysien ensimmäisessä vaiheessa malleihin sisällytettiin vain yksi selittävä tekijä. Taustamuuttujien ja oppilaiden osaamisen välisen yhteyden vahvuutta tarkastellaan raportissa selitysosuutta kuvaavan R^2-tunnusluvun avulla. Taustamuuttujien kohdalla käytettiin perinteisistä yksitasoregressioanalyyseista tuttua tunnuslukua eikä kaavassa (22) esitettyä monitasomallinnukseen soveltuvaa lukua. Jos selittävä muuttuja oli ryhmittelevä, ryhmien välisten erojen merkittävyyttä tarkasteltiin myös efektikokojen avulla. Efektikoot laskettiin muuten kaavan (19) mukaisesti, mutta jakajana käytettiin yksitasoregression tuottamaa jäännöshajontaa. Asenteiden, harrastuneisuuden ym. kohdalla tehtiin myös hierarkkista analyysia, jossa regressiomalliin lisättiin selittäviä taustamuuttujia ryppäittäin ja katsottiin, kuinka paljon muuttujien lisääminen malliin kasvatti selitysastetta.

Oppilaiden osaamisen erojen lisäksi raportissa tarkastellaan tyttöjen ja poikien välisiä eroja esimerkiksi oppiainetta koskevissa asenteissa ja harrastuneisuudessa. Tällöin käytettiin regressiomallia, joissa eri taustamuuttujia selitettiin oppilaan sukupuolen avulla. Myös asenteiden muutoksia eri

arviointivuosina (2005, 2010, 2014 ja 2019) tarkasteltiin regressiomalleilla. Tällöin selittävinä muuttujina käytettiin arviointivuosia kuvaavia dummy-muuttujia ja vertailukohtana vuotta 2019.

Kun oppilaiden päättöarvosanoja selitettiin arvioinnin kokonaistuloksella, analyysissa käytettiin monitasoregressiomallinnusta (esim. Hox 2010). Sen avulla saatiin tarkasteltua koulujen välisiä eroja arvosanojen antamisessa. Käytetty monitasomalli on kuvattu kaavoissa (28), (29) ja (30).

$$y_{il} = \beta_{0l} + \beta_{1l} x_{il} + \varepsilon_{il} \quad (28)$$

$$\beta_{0l} = \gamma_{00} + u_{0l} \quad (29)$$

$$\beta_{1l} = \gamma_{11} + u_{1l} \quad (30)$$

Kaavassa (28) oppilaan päättöarvosanaa y_{il} selitetään arvioinnin kokonaistuloksella x_{il}. Koska kokonaisosaamisen keskiarvo oli 0, mallin vakiotermi β_{0l} kuvaa keskimääräisen oppilaan keskimääräistä osaamista. Vakiotermin kuitenkin annettiin vaihdella kouluittain (engl. random intercepts, kaava 29), mikä tarkoittaa, että arvioinnissa keskitasoisesti suoriutunut oppilas sai erilaisia arvosanoja riippuen siitä, mitä koulua hän kävi. Vakiotermin vaihtelua kuvaavan termin u_{0l} oletetaan noudattavan normaalijakaumaa, jonka keskiarvo on 0 ja varianssi σ^2_{u0}. Vakiotermin lisäksi myös arvosanan ja arvioinnin kokonaistuloksen välisen yhteyden annettiin vaihdella kouluittain (engl. random slopes, kaava 30). Tällöin ankarimmin ja löyhimmin arvosanoja antavien koulujen välinen ero suhteessa arvioinnin kokonaistulokseen ei ole yhtä suuri esimerkiksi päättöarvosanan 5 ja päättöarvosanan 10 saaneilla oppilailla. Yhteyden voimakkuutta kuvaavan termin u_{1l} oletetaan mallissa noudattavan normaalijakaumaa, jonka keskiarvo on 0 ja varianssi σ^2_{u1}.

Monimutkaisimmillaan taustamuuttujia koskevat tarkastelut olivat niissä mallinnuksissa, joissa tarkasteltiin tuotetun tekstin sanamäärän (arvioinnin osa 1) ja tehtäviin käytetyn ajan (arvioinnin osa 2) yhteyttä tuloksiin. Näissä analyyseissa käytettiin polkumallia (esim. Bollen 1989), joka kuvataan kaavoissa (31) (33).

$$y_i = \alpha_0 + \alpha_t t_i + \alpha_s s_i + \sum \alpha_j x_{ij} + \varepsilon_i \quad (31)$$

$$t_i = \beta_0 + \beta_s s_i + \sum \beta_j x_{ij} + \epsilon_i \quad (32)$$

$$x_{ij} = \gamma_{0j} + \gamma_{1j} s_i + \varsigma_{ij} \quad (33)$$

Kaavassa (31) oppilaiden tulosta y_i selitetään oppilaan tuottamaa sanamäärää tai tehtäviin käyttämää aikaa (t_i), oppilaan sukupuolella s_i sekä asenteita, harrastuneisuutta ja mediankäyttöä kuvaavilla muuttujilla x_{ij}. Termi α_0 on mallin vakiotermi ja muut α:t selittäviin muuttujiin liittyviä regressiokertoimia.

Oppilaan tuottama sanamäärä tai tehtäviin käytetty aika t_i on sekin yhteydessä oppilaan sukupuoleen, asenteisiin ja harrastuneisuuteen (kaava 32). Oppilaiden asenteissa ja harrastuneisuudessa x_{ij} puolestaan on eroja tyttöjen ja poikien välillä (kaava 33).

Rakenneyhtälömallin avulla saatiin tietoa siitä, miten arviointiin panostaminen oli yhteydessä tuloksiin. Sen lisäksi mallin avulla oli mahdollista tarkastella, kuinka suuri osuus tyttöjen ja poikien arvioinnissa osoittaman osaamisen eroista oli selitettävissä arvioinnissa yrittämisellä sekä oppilaiden asenteilla ja harrastuneisuudella. Tämän selvittämiseksi, laskettiin ensin yhteen kaikki sukupuoleen liittyvät suorat ja epäsuorat vaikutukset kaavan (34) mukaisesti. Tämän jälkeen arviointiin panostamisen, ajankäytön ja harrastuneisuuden osuus tyttöjen ja poikien välisestä osaamisen erosta voitiin laskea kaavalla (35).

$$\hat{s}_{tot} = \hat{\alpha}_s + \hat{\alpha}_t\hat{\beta}_s + \sum \hat{\gamma}_{1j}\hat{\alpha}_j + \sum \hat{\alpha}_t\hat{\beta}_j\hat{\gamma}_{1j} \qquad (34)$$

$$\hat{s}_{ind\%} = 1 - \frac{\hat{\alpha}_s}{\hat{s}_{tot}} \qquad (35)$$

Kaavassa (34) $\hat{\alpha}_s$ on sukupuolen suora yhteys arvioinnin tulokseen, ja $\hat{\alpha}_t\hat{\beta}_s$ tarkoittaa sukupuolen epäsuoraa yhteyttä oppilaan tuottaman sanamäärän tai tehtäviin käyttämän ajan kautta. Termit $\hat{\gamma}_{1j}\hat{\alpha}_j$ ovat sukupuolen epäsuoria yhteyksiä asenteiden ja harrastuneisuuden kautta, ja termit $\hat{\alpha}_t\hat{\beta}_j\hat{\gamma}_{1j}$kuvaavat epäsuoria yhteyksiä tulokseen asenteiden ja harrastuneisuuden sekä tehtäviin käytetyn ajan tai tuotetun sanamäärän kautta.

Lähteet

Asparouhov, T. & Muthén, B. 2012. Multiple group multilevel analysis. Mplus Web Notes: No. 16. https://www.statmodel.com/examples/webnotes/webnote16.pdf

Bollen, K. 1989. Structural equations with latent variables. New York: John Wiley.

Chalmers, R. P. & Ng, V. 2017. Plausible-Value Imputation Statistics for Detecting Item Misfit. Applied Psychological Measurement, 41, 372-387.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial
credit. Psychological Bulletin 70 (4), 213–220.

deAyala, R. 2009. The theory and practice of item response theory. New York: The Guildford Press.

Hedges, L. 2009. Effect sizes in nested designs. Teoksessa H., Cooper, L., Hedges & J., Valentine (toim.) The handbook of research synthesis and meta-analysis. New York: Russel Sage Foundation.

Hox, J. 2010. Multilevel analysis. Techniques and applications. Second edition. New Yorl: Routledge.

Landis, J. & Koch, G. 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

R. Philip Chalmers. 2012. mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. doi:10.18637/jss.v048.i06

Muthén, B. & Asprouhov, T. 2018. Multidimensional, multilevel, and multi-timepoint item response modeling. Teoksessa W. van der Linden (toim.) Handbook of item response theory. Volume three. Applications. Boca Raton, FL: CRC Press.

Muthén, L.K. and Muthén, B.O. 1998-2017. Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Nelson, K. & Edward, D. 2015. Measures of agreement between many raters for ordinal classifications. Stat. Med. 34 (23), 3116–3132.

Nissinen, K., Rautopuro, J., & Puhakka, E. 2018. PISA-tutkimuksen metodologiasta. Teoksessa J. Rautopuro, & K. Juuti (toim.), PISA pintaa syvemmältä: PISA 2015 Suomen pääraportti (pp. 345–378). Kasvatusalan tutkimuksia, 77. Jyväskylä: Suomen kasvatustieteellinen seura.

Nunnally J. 1978. Psychometric theory (2nd ed.). New York, NY: McGraw-Hill.
OECD. 2009. PISA data analysis manual. SPSS Second edition. Paris: OECD Publishing.

Palomo, J., Dunson, D. & Bollen, K. 2007. Bayesian structural equation modeling. Teoksessa S.-Y. Lee (toim.) Hanbook of latent variable and related models. Amsterdam : Elsevier.

Portney, L. 2020. Foundations of clinical research. Fourth edition. Philadelphia, F. A. Davis Company.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revelle, W. 2019. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.9.12.

Suomen virallinen tilasto (SVT): Esi- ja peruskouluopetus [verkkojulkaisu]. ISSN=1799-3709. Helsinki: Tilastokeskus [viitattu: 18.6.2020]. Saantitapa: http://www.stat.fi/til/pop/index.html

Vipunen - opetushallinnon tilastopalvelu: 7–9 luokilla suomea tai ruotsia toisena kielenä opiskelleet. Saantitapa: https://vipunen.fi/fi-fi/_layouts/15/xlviewer.aspx?id=/fi-fi/Raportit/Perusopetus%20-%20ainevalinnat%20-%20suomi%20tai%20ruotsi%20toisena%20kielen%C3%A4%20-%207-9%20-%20maakunta.xlsb [Viitattu 18.6.2020]

von Davier, M., Gonzalez, E. & Mislevy R.J. 2009. What are plausible values and why are they useful? I verket IERI. 2017. Issues and methodologies in large-scale assessments. Hamburg, Germany: IER-Institute.

KARVI

Tässä liitteessä esitellään suomen kielen ja kirjallisuuden perusopetuksen oppimistulosarvioinnissa käytetyt otanta- ja analyysimenetelmät. Menetelmien teoriataustaa ei kuvata yksityiskohtaisesti, mutta viitteet kattavampiin lähteisiin annetaan tekstissä. Liitteessä kuvataan lyhyesti myös aineistojen käsittelyn vaiheet.

ISBN 978-952-206-609-1 pdf
ISSN 2342-4184 (verkkojulkaisu)

Kansallinen
koulutuksen arviointikeskus
PL 28 (Mannerheiminaukio 1 A)
00101 HELSINKI
Puhelinvaihde: 029 533 5500
Faksi: 029 533 550

karvi.fi