JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Nissinen, Kari; Rautopuro, Juhani; Puhakka, Eija

Title: PISA-tutkimuksen metodologiasta

Year: 2018

Version: Published version

Copyright: © Kirjoittajat & Suomen kasvatustieteellinen seura, 2018.

Rights: CC BY 4.0

Rights url: https://creativecommons.org/licenses/by/4.0/

Please cite the original version:

Nissinen, K., Rautopuro, J., & Puhakka, E. (2018). PISA-tutkimuksen metodologiasta. In J. Rautopuro, & K. Juuti (Eds.), PISA pintaa syvemmältä : PISA 2015 Suomen pääraportti (pp. 345-378). Suomen kasvatustieteellinen seura. Kasvatusalan tutkimuksia, 77.
http://www.urn.fi/URN:ISBN:978-952-5401-82-0

Kari Nissinen,
Juhani Rautopuro &
Eija Puhakka

14. PISA-tutkimuksen metodologiasta

Johdanto

PISA on laaja kansainvälinen tutkimushanke, jonka menestyksellinen toteuttaminen edellyttää korkean tason menetelmällistä ja sisällöllistä asiantuntemusta sekä tutkimuksen suunnittelussa ja valmistelussa että kerätyn aineiston käsittelyssä. Hankkeen korkea laatu rakentuu osallistujamaiden yhdessä sopimille laatustandardeille, joiden toteutumista tutkimuksesta vastaava konsortio seuraa tarkoin prosessin eri vaiheissa. Tutkimuksen toteutuksen yhdenmukaisuutta ja luotettavuutta tuetaan yksityiskohtaisella ohjeistuksella, valvonnalla ja eri vaiheiden tarkalla dokumentoinnilla.

Tutkimusta varten laaditaan kansainvälisen asiantuntijaryhmän voimin arvioinnin viitekehys. Asiantuntijaryhmä koostuu johtavista tutkimuksen painopistealueiden oppimisen ja oppimistulosten arvioinnin asiantuntijoista, joiden vastuulla on tuottaa arvioinnissa käytettävät tehtävät ja kyselylomakkeet. Tehtäviä arvioidaan tarkastelemalla muun muassa niiden vaikeutta, vastaavuutta maan opetussuunnitelmaan ja sopivuutta 15-vuotiaiden maailmaan. Viitekehyksen laadinnan eri vaiheissa tutkimukseen osallis-

tuvat maat voivat kommentoida ja arvioida luonnoksia. Tällä pyritään parantamaan arvioinnin kulttuurista edustavuutta ja tasapuolisuutta.

Tutkimukseen esitettyjä tehtäviä esitestataan useita kertoja eri vaiheissa ja lopulliseen arviointiin valitaan noin 10–20 prosenttia esitetyistä tehtävistä. Tavoitteena on saada koe, joka kattaa mitattavan alueen mahdollisimman laaja-alaisesti ja on kulttuurisesti mahdollisimman tasapuolinen osallistujamaiden kesken. Lisäksi tehtävillä tulee olla kyllin hyvä erottelukyky ja niiden vaikeustason tulee vaihdella riittävästi.

PISA-tutkimuksen avoimuutta ja läpinäkyvyyttä lisää se, että tutkimusaineisto on vapaasti käytettävissä sen jälkeen, kun ensimmäiset tulokset on julkaistu. Aineistosta on luonnollisesti poistettu yksittäisiä oppilaita ja kouluja koskevat tunnistetiedot.

Vaikka PISA-tutkimuksilla on takanaan jo yli 15 vuoden historia aina vuodesta 2000 alkaen, niissä käytetyistä aineiston keruuseen, käsittelyyn ja analysointiin liittyvistä menetelmävalinnoista ei ole ollut saatavilla suomenkielistä yleisesitystä. Tämän artikkelin tavoite on paikata tätä aukkoa. Suomenkielisen yleisesityksen puute on saattanut olla osaltaan vaikuttamassa siihen, että PISA-tutkimusten mahdollisuuksista ja rajoista esiintyy vakaviakin väärinkäsityksiä. Myöskään PISA-aineistojen validin tilastollisen analyysin vaatimuksia ei ole kaikilta osin ymmärretty. Seurauksena on, että aineistoista on saatettu julkaista tilastollisia analyysejä, joissa ei ole osattu tai tiedetty käyttää aineiston ominaisuuksien edellyttämiä menetelmiä ja joiden tulokset ovat sen myötä virheellisiä.

Koska PISA-tutkimukset ovat saaneet julkisuudessa tavattoman paljon huomiota, on paikallaan selvittää vähintäänkin yleisellä tasolla, miten aineistot kerätään, keitä aineistojen keruu koskee, miten kokeet järjestetään, miten niistä määritetään tulokset ja mitä nämä tulokset kertovat. Mahdollisimman edustavan ja luotettavan aineiston hankinnassa ja edelleen analyysissä joudutaan huomioimaan ja kontrolloimaan monenlaisia seikkoja.

PISA-tutkimusten metodologiasta ei ole helppoa kirjoittaa yleistajuista esitystä. Erityisesti PISAn tilastotieteellinen metodo-

logia on monimutkaista ja sen aito ymmärtäminen vaatii syventävän tason tietoja tilastotieteestä. Tällaisessa artikkelissa ei ole mahdollisuuksia edes kohtuullisen tarkkaan tilastotieteellisen metodologian esittelyyn. Täsmällinen esitys vaatisi tilastomatemaattisen formalismin ts. laskukaavojen käyttöä, mikä ei ole perusteltua tällaisessa laajahkolle yleisölle suunnatussa kokoomateoksessa.

PISA kuten muutkin laajat kansainväliset arviointitutkimukset, esimerkiksi PIAAC, PIRLS ja TIMSS, soveltavat likimain samanlaista metodologiaa, joka koostuu hyvin monenlaisista osa-alueista aina otannan ja kokeen suunnittelusta ja tehtävien laadinnasta numeeristen tietokantojen rakentamiseen, osioanalyyseihin ja tulosjakaumien laskentaan. Aihepiiristä on olemassa useampiakin ulkomaisia menetelmäkäsikirjoja, ja alueen laajuutta kuvaa se, että näiden teosten tyypillinen laajuus on 500–600 sivua. Tämän artikkelin lähdeluettelossa on mainittu muutamia keskeisiä teoksia. Lisäksi arviointitutkimusten omista teknisistä raporteista ja muista vastaavista dokumenteista on saatavissa tietoa niiden menetelmällisistä ratkaisuista.

Tämä artikkeli tasapainoilee yleistajuisuuden ja täsmällisyyden välissä. Kaikki aiheet on pyritty käsittelemään niin tarkasti kuin se verraten tiiviissä ja pääosin yleiskielisessä esityksessä on mahdollista. Matemaattiset kaavat on haluttu välttää kokonaan. Sen sijaan on yritetty kuvailla verbaalisesti ne tavoitteet, joihin PISAssa omaksutuilla tilastomatemaattisilla menetelmillä pyritään, ja ne lähtökohdat ja perusideat, joita nämä menetelmät noudattavat. Tämän vuoksi tilastotieteellistä terminologiaa ei ole kaikilta osin pystytty välttämään – oikean kuvan saaminen eräistä käsitellyistä asioista edellyttää lukijalta jonkin verran pohjatietoja tilastotieteestä. Aiheista enemmän ja eksaktimpaa tietoa haluavia lukijoita varten tekstissä on annettu keskeisiä kirjallisuusviitteitä.

Artikkelissa esitellään aluksi PISA 2015 -tutkimuksen otantaasetelma ja otannan toteuttaminen Suomessa. Sen jälkeen kuvataan oppilaiden tekemän kokeen järjestelyä ja siihen olennaisesti kuuluvaa koetehtävien hajauttamista. Seuraavissa luvuissa tarkastellaan osioanalyysin teoriasta ja Bayes-tilastotieteestä nousevaa sangen monimutkaista menetelmäkokonaisuutta, jolla kerätystä

havaintoaineistosta jalostetaan kansalliset PISA-tulokset. Samalla pyritään valaisemaan sitä, mitä nämä tulokset itse asiassa ovat ja mitä päätelmiä niistä on mahdollista tehdä. Lopuksi tarkastellaan niitä tilastoanalyyseihin liittyviä erityiskysymyksiä, jotka PISA-aineistojen oikeaoppisissa analyyseissä tulee ottaa huomioon.

Otannan toteutus Suomen PISA 2015 -tutkimuksessa

PISA-tutkimusten kohdepopulaation muodostavat oppilaat, jotka ovat tutkimuksen ajankohtana noin 15-vuotiaita ja koulussa vähintään 7. luokalla (OECD 2017, luku 4). Vuoden 2015 PISA-tutkimuksen kohdejoukko määriteltiin Suomessa siten, että siihen kuuluivat kaikki 1.2.1999–31.1.2000 syntyneet nuoret, jotka olivat kevätlukukaudella 2015 perusopetuksessa luokilla 7–10 tai toisen asteen oppilaitoksessa. Suomessa tähän kohdejoukkoon kuului noin 59 000 nuorta. Kohdejoukosta tuli otantaa varten tavoittaa vähintään 95 prosenttia. Ennen otantaa Suomessa rajattiin kuitenkin kohdeperusjoukosta pois ne koulut, joiden virallinen opetuskieli oli muu kuin suomi tai ruotsi. Tämän rajauksen vaikutus kohdeperusjoukkoon oli vähäinen: tällaisia kouluja oli vuonna 2015 vain kahdeksan ja niissä opiskeli noin 300 oppilasta.

Kustakin osallistujamaasta tuli poimia otokseen vähintään 150 koulua, ja koulukohtainen tavoiteotoskoko oli 42 oppilasta. Siten PISA 2015 -tutkimuksen kansallinen tavoiteotoskoko oli vähintään 6 300 oppilasta. PISA-otanta suoritetaan kaksivaiheisesti: ensimmäisessä vaiheessa poimitaan satunnaisotannalla kohdejoukkoa vastaavien koulujen (ts. ylä- ja yhtenäiskoulujen sekä toisen asteen oppilaitosten) joukosta ne koulut, joiden oppilaat aiotaan testata. Toisessa vaiheessa suoritetaan oppilasotanta kunkin poimitun koulun 15-vuotiaiden oppilaiden luettelosta. Toisin kuin eräissä muissa kansainvälisissä arviointitutkimuksissa (esim. PIRLS, TIMSS) PISAssa ei poimita kokonaisia luokkia, vaan koulun oppilaat tulevat otokseen satunnaisesti eri rinnakkaisluokilta (mikäli näitä koulussa on).

Suomessa PISA-koulut valittiin peruskouluista sekä lukioista ja ammatillisista oppilaitoksista. Myös erityiskoulut sisältyivät otantaan. PISAssa käytetään koulujen poiminnassa ositettua otantaa, jolla pyritään varmistamaan kansallisten aineistojen edustavuus esimerkiksi maantieteellisesti ja kieliryhmien suhteen. Osittamiskriteerit vaihtelevat jonkin verran maittain. Suomessa PISA 2015 -tutkimuksen osittamisperusteina olivat EU:n ja Tilastokeskuksen yleisesti käyttämät suuralueet, koulun opetuskieli ja kuntaryhmä. EU:n aluejakoon perustuvat suuralueet ovat Suomessa Helsinki-Uusimaa, Etelä-Suomi, Länsi-Suomi sekä Pohjois- ja Itä-Suomi. Osituksessa käytetyt opetuskielet olivat suomi ja ruotsi. Suuralueista ja ruotsinkielisistä kouluista muodostetut osajoukot jaettiin osituksessa vielä kuntaryhmittelyn mukaan kaupunkimaisiin ja maaseutumaisiin kuntiin. Kaupunkimaiset kunnat käsittivät myös taajaan asutut kunnat.

PISA-otannan ensimmäisessä vaiheessa poimittiin koulut ositteittain systemaattisella PPS-otannalla (probability proportional to size), jossa koulun poimintatodennäköisyys on sitä suurempi, mitä suurempi on koulun koko. Toisin sanoen mitä enemmän koulussa on kohdejoukkoon kuuluvia oppilaita, sitä todennäköisemmin koulu osuu otokseen. Kouluotanta suoritettiin muutamia kuukausia ennen varsinaisen aineistonkeruun aloittamista, ja sen pohjana käytettiin senhetkisiä Tilastokeskuksen oppilaitosrekistereitä. Otantaa ei suoritettu kotimaassa, vaan sen toteutti kaikkien osallistujamaiden osalta keskitetysti yhdysvaltalainen tutkimuslaitos Westat sille toimitetuista oppilaitosluetteloista.

PISAn kohdejoukkoon kuuluvien koulujen koon määrittämistä vaikeutti se, että Tilastokeskuksen oppilaitosrekistereissä ei ole käytettävissä peruskoulujen oppilaiden syntymäaikoja. Sen vuoksi peruskoulujen kokoa mitattiin otannassa niiden 9.-luokkalaisten määrällä. Toisen asteen oppilaitosrekistereistä oli sen sijaan saatavissa oppilaitoksen 15-vuotiaiden oppilaiden lukumäärä, joka on PISA-otannan kannalta tällaisen oppilaitoksen kokonaisoppilasmäärää mielekkäämpi koon mitta (ja yleensä hyvin pieni). Koska koulujen oppilaita koskevat rekisteritiedot eivät välttämättä olleet ajantasaisia, kouluotannan kehikkoon otettiin varmuuden

vuoksi mukaan myös sellaiset toisen asteen oppilaitokset, joissa Tilastokeskuksen rekisterien mukaan ei ollut 15-vuotiaita oppilaita. Näiden oppilaitoksen kooksi asetettiin otannassa yksi. Jos otannan jälkeen osoittautui, että poimitussa koulussa ei ollut kohdejoukkoon kuuluvia oppilaita, koulu jäi otoksesta pois.

Tiedot ositteeseen kuuluvista oppilaitoksista ja niiden oppilasmääristä saatiin siis Tilastokeskuksen rekistereistä. Koulujen kokoa mittaavilla lukuarvoilla arvioitiin myös kohdepopulaation koko ositteissa. Ositteesta poimittavien koulujen lukumäärä määrättiin oppilastasolla sovelletun suhteellisen kiintiöinnin (proportional allocation) periaatteella: ositteesta poimittiin niin monta koulua, että tavoiteotoskokojen täyttyessä ositteen osuus kansallisesta oppilasotoksesta vastasi sen osuutta kohdepopulaatiossa. Jos ositteeseen kuului esimerkiksi 10 prosenttia oppilaspopulaatiosta, siitä poimittiin niin monta koulua, että ositteesta saatu oppilasotos kattoi likimain 10 prosenttia koko maan oppilasotoksesta. Suomen PISA 2015 -otannassa ositteista valittiin otokseen 2–41 koulua ositteen koosta riippuen.

Suomessa on muutamilla PISA-kierroksilla poikettu suhteellisen kiintiöinnin periaatteesta. Perusteena tälle on ollut se, että kiinnostuksen kohteena on ollut sellaisia vähemmistöryhmiä, joista suhteellisen kiintiöinnin mukainen otos jäisi luotettavan tilastollisen analyysin kannalta liian pieneksi. Vuosina 2003, 2009 ja 2012 poimittiin ruotsinkielisistä kouluista yliotos siten, että kaikki ne, joissa oli 15-vuotiaita oppilaita, tulivat aineistoon mukaan. Osituksen näkökulmasta tämä tarkoitti sitä, että otoksessa oli merkittävästi enemmän oppilaita ruotsinkielisten koulujen ositteista kuin heidän suhteellisen osuutensa mukaan pitäisi olla. Esimerkiksi vuoden 2009 PISA-otoksessa 24 prosenttia oppilaista oli ruotsinkielisistä kouluista, kun heidän osuutensa oppilaspopulaatiossa oli noin 6 prosenttia. Vuonna 2012 poimittiin ruotsinkielisten lisäksi yliotos maahanmuuttajataustaisista oppilaista siten, että kaikki koulut, joissa oli vähintään viisi 15-vuotiasta maahanmuuttajataustaista oppilasta, otettiin mukaan, ja näistä kouluista testattiin kaikki maahanmuuttajaoppilaat. Yliotostuksen tuloksena 15 prosenttia Suomen PISA 2012 -aineiston oppilaista oli maa-

hanmuuttajataustaisia, kun heidän osuutensa oppilaspopulaatiossa oli noin 3,5 prosenttia. Vuoden 2015 PISA-otannassa ei ollut yliotostusta.

Yliotostuksesta aiheutuu aineistoon vääristymä, joka korjataan aineiston analyysivaiheessa otanta-asetelmasta johdetuilla oppilaskohtaisilla painokertoimilla. Painokerroin saadaan havaintoyksikölle (oppilaalle ja koululle) otanta-asetelmasta määritetyn poimintatodennäköisyyden käänteislukuna, ja se kuvaa, kuinka monta perusjoukon havaintoyksikköä yksi otokseen poimittu havaintoyksikkö edustaa. Jos esimerkiksi tietyltä alueelta poimitun oppilaan painokerroin on 8, hän edustaa otoksessa kahdeksaa alueensa oppilasta. Erityisesti yliotostettaessa osittamisella on keskeinen rooli, sillä se varmistaa, että eri ositteisiin kuuluvat oppilaat tulevat asianmukaisella tavalla painotetuiksi. Vuoden 2012 kouluotannassa käytettiin ylimääräisenä ositemuuttujana maahanmuuttajaoppilaiden määrää koulussa luokiteltuna kolmeen luokkaan (paljon / vähän / ei lainkaan maahanmuuttajataustaisia oppilaita). Vuoden 2012 kaltaisessa yliotostustilanteessa, jossa poimittiin kaikki maahanmuuttajaoppilaat, maahanmuuttajaoppilaan painokerroin on 1, eli hän edustaa vain itseään. Kun PISA 2012 -aineisto analysoidaan painokertoimia käyttäen, maahanmuuttajaoppilaiden osuus palautuu heidän todellista osuuttaan vastaavaan 3,5 prosenttiin. Yliotostuksen ohella painokertoimilla tasapainotetaan myös muunlaisia otanta-asetelmasta tai vastauskadosta johtuvia aineiston vinoumia. Siksi PISA-aineistojen analyyseissä on syytä käyttää painokertoimia rutiininomaisesti.

Otannan toisessa vaiheessa kunkin poimitun koulun kaikkien PISAn ikäkriteerin täyttävien oppilaiden luettelosta valittiin systemaattisella tasaväliotannalla 42 oppilasta. Mikäli oppilaita on tätä vähemmän, otettiin mukaan heidät kaikki. Luettelot hankittiin otannan ensimmäisessä vaiheessa poimituilta kouluilta. Oppilasotoksen edustavuuden parantamiseksi luettelot järjestettiin poimintaa varten luokka-asteen ja sukupuolen mukaan. Oppilasotanta suoritettiin Koulutuksen tutkimuslaitoksessa PISA-konsortion omalla KeyQuest-poimintaohjelmistolla (OECD 2014, luku 10).

Suomen PISA 2015 -otokseen poimittiin 168 koulua ja näistä alun perin kaikkiaan 6 431 oppilasta. Mahdolliseen vastauskatoon varauduttiin poimimalla vähimmäisvaatimusta suurempi määrä kouluja. Oppilaista 137 suljettiin kokeesta pois ennalta sovittujen kriteerien perusteella. Kriteerien mukaan oppilas suljettiin pois, jos hänen ikänsä osoittautui vääräksi, hänen kielitaitonsa oli riittämätön kokeen itsenäiseen suorittamiseen, hän ei olisi kyennyt selviytymään kokeesta ilman avustajaa tai hän oli vaihtanut koulua otannan suorittamisen jälkeen. Jäljelle jääneistä 6 294 oppilaasta PISA-kokeeseen osallistui lopulta 5 882 oppilasta, joten oppilaiden osallistumisaste vuoden 2015 tutkimukseen oli 93 prosenttia. Osallistuneista noin 87 prosenttia oli 9.-luokkalaisia, noin 13 prosenttia 8.-luokkalaisia ja vajaa prosentti 7.-luokkalaisia. Peruskoulun 10.-luokkalaisia, lukiolaisia ja ammattikoululaisia oli kokeessa vain muutama. Tavallisin syy kokeesta poisjäännille oli, että oppilas oli testipäivänä poissa koulusta. Muutamassa yksittäistapauksessa koe jäi toteuttamatta koulun tietokoneongelmien takia.

Kun otokseen poimittavien oppilaiden määrä on (likimain) sama koulusta riippumatta, PPS-kouluotannan ja systemaattisen oppilasotannan yhdistelmällä on hyviä otantateoreettisia ominaisuuksia verrattuna esimerkiksi siihen, että koulut poimittaisiin ositteista yksinkertaisella satunnaisotannalla. PPS-kouluotanta esimerkiksi skaalaa otoksen siten, että ositteen otantapainotettu havaintomäärä vastaa ositteen oppilaspopulaation kokoa. Käytännössä esimerkiksi vastauskato aiheuttaa tähän kuitenkin pientä eroa. Edelleen, saatu oppilasotos on periaatteessa ns. itsepainottuva (self-weighting), toisin sanoen jokaisen oppilaan todennäköisyys tulla kaksivaiheisessa otannassa poimituksi otokseen on ositteen sisällä sama. Muun muassa osituksen tai yliotostuksen seurauksena koulujen ja oppilaiden poimintatodennäköisyydet vaihtelevat kuitenkin jonkin verran, mikä voi aiheuttaa otoksen kokoonpanoon vinoumaa kohdeperusjoukkoon verrattuna. Tämä vinouma, samoin kuin mahdollisesta vastauskadosta tai yliotostuksesta johtuvat vääristymät, korjataan tilastollisissa analyyseissä käyttämällä otanta-asetelmasta kouluille ja oppilaille johdettuja paino-

kertoimia. Painokertoimien avulla otoksen kokoonpano saadaan laskennallisesti vastaamaan kohdeperusjoukossa vallitsevaa tilannetta. Samalla varmistetaan otantaan liittyvien seikkojen osalta otosaineistosta laskettujen tulosten vertailukelpoisuus sekä kansainvälisesti että aikaisempiin PISA-kierroksiin nähden. Vuoden 2015 tutkimuksen painokertoimien laskennan toteutti Westat-tutkimuslaitos.

Saatava aineisto on rakenteeltaan kaksitasoinen ja klusteroitunut: oppilastaso on hierarkkisesti koulutason alla. Samaan kouluun kuuluvat oppilaat eivät ole toisistaan riippumattomia, mikä on vastoin tavanomaisten tilastollisten analyysimenetelmien perusoletuksia. PISAn kaltaisten hierarkkisten aineistojen virheetön tilastollinen analyysi edellyttääkin aineiston monitasoisuuden huomioon ottavia menetelmiä. Tällaisia ovat esimerkiksi monitasomallit (multilevel models, mixed models, ks. esim. Goldstein 2010; Snijders & Bosker 2012) tai ryväsotannalla poimittujen aineistojen analyysiin soveltuvat survey-menetelmät (survey statistical methods, ks. esim. Lehtonen & Pahkinen 2004).

Oppilaita koskevan tiedonkeruun lisäksi koulujen rehtoreille suunnattiin erillinen verkkokysely, jossa kysyttiin muun muassa koulun resursseista ja käytänteistä. Joissakin osanottajamaissa on toteutettu PISAn yhteydessä myös opettaja- ja vanhempainkyselyjä, mutta näin ei ole tehty Suomessa.

PISA 2015 -tutkimuksen koejärjestelyt

Vuodesta 2015 alkaen PISA-koe ja sen taustakysely on toteutettu useimmissa osanottajamaissa (myös Suomessa) tietokonepohjaisesti. PISA-kokeen kesto on kaksi tuntia, ja siitä puolet käytetään pääarviointialueen tehtäviin. Toinen puoli käytetään sivuarviointialueiden tehtäviin. Vuonna 2015 pääarviointialueena olivat luonnontieteet ja sivuarviointialueina lukutaito, matematiikka ja yhteistoiminnallinen ongelmanratkaisu. Kokeen puolivälissä on lyhyt tauko. Kaikki oppilaat saavat tehdäkseen pääarviointialueen tehtäviä, mutta sivuarviointialueiden tehtävät hajautetaan oppilaille

etukäteen laaditun koejärjestelyn mukaan satunnaisesti niin, että oppilas tekee vain yhden tai kahden sivuarviointialueen tehtäviä. Siten vuonna 2015 lukutaidon, matematiikan ja yhteistoiminnallisen ongelmanratkaisun tehtäviä teki vain osa kokeeseen valituista oppilaista. Lisäksi tehtävien järjestys satunnaistettiin oppilaille siten, että puolet oppilaista tekee pääarviointialueen tehtävät ennen taukoa ja toinen puoli tauon jälkeen. Koejärjestelyä kuvataan tarkemmin jäljempänä.

Varsinaisen kokeen jälkeen oppilas vastasi noin 35 minuuttia kestävään taustakyselyyn, jossa kartoitetaan muun muassa oppilaan kotitaustaa, kiinnostuksen kohteita sekä opiskelua ja koulua koskevia kokemuksia ja asenteita erityisesti pääarviointialueeseen liittyen. Useimmissa osallistujamaissa, myös Suomessa, PISA 2015 -tutkimuksen oppilaskyselyyn liitettiin erillinen tietotekniikkakysely, jolla kerättiin tietoa oppilaan tietotekniikan käytöstä ja siihen liittyvistä asenteista.

Erityiskoulujen oppilaat saivat tehdäkseen lyhyemmän, tunnin mittaisen ns. Une Heure (UH) -kokeen (kaksi kertaa 30 minuuttia, välissä lyhyt tauko) ja sen jälkeen 15 minuutin mittaisen suppean taustakyselyn.

PISA-tutkimuksissa arvioidaan oppilaiden osaamista laajalla, kansainvälisen sisältöasiantuntijaryhmän valitsemalla tehtäväkokoelmalla. Laajalla tehtävävalikoimalla pyritään siihen, että arvioitavan aihealueen kansallinen osaamistaso saadaan kartoitettua mahdollisimman monipuolisesti. Tehtävien aihepiirit ja vaikeustasot vaihtelevat, ja erityisesti oppilaan saamien tehtävien vaikeusaste otetaan huomioon hänen lopullista koetulostaan laskettaessa. Tehtävien toimivuus testataan vuotta ennen varsinaista tutkimusta toteutettavalla esikokeella, ja kansainvälisten osioanalyysien avulla karsitaan huonosti toimivat tehtävät pois varsinaisesta pääkokeesta. Tehtävä hylätään esimerkiksi, jos sen kysymyksenasettelu osoittautuu moniselitteiseksi tai jos se ymmärretään eri maissa tai eri kielillä merkittävässä määrin eri tavalla. Aihealueesta riippumatta PISA-tehtäviin liittyy johdantoteksti, jonka pituus vaihtelee tehtävästä toiseen ja joka voi sisältää taustamateriaalia, kuten kuvioita tai numeerisia taulukoita. Tehtävät

pohjautuvat tekstiin ja annettuun oheismateriaaliin. Ne ovat pääasiassa soveltavia, eikä vastaus niihin löydy suoraan tekstistä, vaan edellyttää omaa päättelyä. Tehtävät jakautuvat usein alakohtiin tai osioihin, joista osa on monivalintoja ja osa edellyttää avovastausta. Monivalintakysymykset pisteytetään sähköisesti, kun taas avovastaukset pisteyttää koulutettu arvioijaryhmä kansainvälisen ohjeistuksen mukaisesti.

Vuoden 2015 PISA-tutkimuksessa oli luonnontieteistä 83 eri tehtävää, matematiikasta 45 tehtävää, lukutaidosta 29 tehtävää ja yhteistoiminnallisesta ongelmanratkaisusta 7 tehtävää. Koska oppilaan ei ole mahdollista tehdä kahden tunnin koeajassa kaikkia mahdollisia tehtäviä, hän sai vastatakseen neljän ns. tehtäväklusterin kokonaisuuden. Yhteensä tämä merkitsi noin 25 tehtävän tekemistä; tehtävien määrä vaihtelee hieman klusterista toiseen. Yhteen klusteriin vastaamisen on laskettu vievän noin puoli tuntia. Se, mitkä tehtäväklusterit kukin oppilas sai, määräytyi satunnaisesti, kuitenkin niin, että kaikki oppilaat tekivät kaksi klusteria pääarviointialueen tehtäviä. Sivuarviointialueiden tehtäväklusterit satunnaistettiin oppilaille etukäteen määrätyn asetelman mukaisesti.

Vuonna 2015 luonnontieteiden tehtävät ryhmittyivät 12 erilaiseen klusteriin. Matematiikan ja lukutaidon tehtävät ryhmittyivät kumpikin kuuteen klusteriin ja yhteistoiminnallisen ongelmanratkaisun tehtävät kolmeen klusteriin. Koska ajan mittaan oppilaiden elin- ja opiskeluympäristöt muuttuvat ja arvioitavat aihepiirit kehittyvät, osa aiemmilla PISA-kierroksilla käytetyistä koetehtävistä vanhentuu, ja ne korvataan uusilla ajanmukaisilla tehtävillä. Valittu joukko vanhoja tehtäviä säilytetään ns. trendi- tai ankkuritehtävinä, joiden avulla uusimmat koetulokset linkitetään vertailukelpoisiksi aiempien PISA-kierrosten tulosten kanssa. Luonnontieteiden 12 klusterista puolet koostui aiemmissa PISA-tutkimuksissa käytetyistä trenditehtävistä. Loput luonnontieteiden klusterit sisälsivät uusia tehtäviä. PISA 2015:ssä kaikki matematiikan ja lukutaidon tehtävät olivat trenditehtäviä. Lukutaidon tehtävät päivitetään seuraavan kerran vuoden 2018 tutkimukseen, jolloin lukutaito on pääarviointialue. Yhteistoiminnallinen ongelmanratkaisu oli uusi arviointialue, joten kaikki sen tehtävät olivat uusia.

Neljän klusterin tehtäväkokonaisuuksia oli kaikkiaan 66 erilaista. On siis mahdollista, että PISA 2015 -tutkimuksessa ei kahdelle saman koulun oppilaalle osunut täsmälleen samanlaista tehtäväkokonaisuutta. Tehtäväkokonaisuudet ryhmittyivät kuuteen erilaiseen kategoriaan:

- 33 prosenttia kokonaisuuksista sisälsi 2 klusteria luonnontieteitä ja 2 lukutaitoa
- 33 prosenttia kokonaisuuksista sisälsi 2 klusteria luonnontieteitä ja 2 matematiikkaa
- 22 prosenttia kokonaisuuksista sisälsi 2 klusteria luonnontieteitä ja 2 ongelmanratkaisua
- 4 prosenttia kokonaisuuksista sisälsi 2 klusteria luonnontieteitä, 1 lukutaitoa ja 1 matematiikkaa
- 4 prosenttia kokonaisuuksista sisälsi 2 klusteria luonnontieteitä, 1 lukutaitoa ja 1 ongelmanratkaisua
- 4 prosenttia kokonaisuuksista sisälsi 2 klusteria luonnontieteitä, 1 matematiikkaa ja 1 ongelmanratkaisua.

Kaikki oppilaat tekivät siis kaksi klusteria luonnontieteiden tehtäviä. Luonnontieteiden tehtäväklusterit jakautuivat oppilaiden kesken edelleen siten, että puolet teki sekä trenditehtäviä että uusia tehtäviä, 42 prosenttia teki pelkästään uusia tehtäviä ja loput 8 prosenttia pelkästään trenditehtäviä. Oppilaista 41 prosenttia teki luonnontieteiden lisäksi lukutaidon tehtäviä, samoin 41 prosenttia teki luonnontieteiden lisäksi matematiikkaa. Yhteistoiminnallista ongelmanratkaisua teki 30 prosenttia oppilaista. Oppilaista valtaosa, 88 prosenttia, teki vain yhden sivuarviointialueen tehtäviä, loput tekivät kahden sivuarviointialueen tehtäviä.

Klusterit järjestettiin tehtäväkokonaisuuksiksi tasapainotetusti siten, että kukin klusteri, olipa se mitä arviointialuetta tahansa, esiintyi kaiken kaikkiaan yhtä usein kokeen ensimmäisenä, toisena, kolmantena tai neljäntenä klusterina. Näin pyrittiin eliminoimaan mahdollinen harha, joka aiheutuisi siitä, että tietyt tehtävät ja aihealueet esiintyisivät tyypillisesti kokeen alkuosassa ja toiset kokeen loppuosassa, jossa oppilaan suoriutuminen on usein kokeen alkuosaa heikompaa ja jossa aikapula voi vaikuttaa onnistumiseen.

PISA-tutkimusten koejärjestelyt tähtäävät ennen kaikkea siihen, että oppilaiden kansallinen osaamistaso saadaan selvitettyä kattavasti. Tulosten luotettavuuden kannalta on olennaista, että kunkin maan oppilasjoukko tekee kokonaisuutena laajan valikoiman erilaisia koetehtäviä. Tämä tulee mahdolliseksi hajauttamalla tehtävät eri oppilaille edellä kuvatulla tavalla. Asetelman kääntöpuolena on se, että erilaisten tehtäväkokonaisuuksien vuoksi yksittäisten oppilaiden koetulokset eivät sellaisenaan ole välttämättä vertailukelpoisia. Oppilaiden kesken vertailukelpoisemmat tulokset saataisiin toki antamalla kaikille oppilaille täsmälleen samat koetehtävät, mutta kokeeseen käytettävissä olevan ajan rajallisuuden vuoksi tehtäväkokoelma jäisi tällöin kansallisen tason luotettavan arvioinnin kannalta liian suppeaksi. Koetehtävien määrän lisääminen ja sen myötä kokeen pitkittäminen ei välttämättä parantaisi kansallisten tulosten tarkkuutta: pitempi ja oppilaalle rasittavampi koe heikentää tyypillisesti hänen testimotivaatiotaan ja lisää alisuorittamisen todennäköisyyttä, varsinkin kun oppilas ei saa PISA-kokeessa menestymisestä mitään henkilökohtaista hyötyä.

PISA-tulosten laskennan perusajatus

PISA-tutkimusten päämääränä on määrittää 15-vuotiaiden oppilaiden lukutaidon, matematiikan, luonnontieteiden ja vaihtuvien aihealueiden osaamisen kansallinen *jakauma* eri maissa. Erityisesti kiinnostavat osaamisen keskimääräinen taso ja vaihtelu sekä eritasoisten osaajien määrät paitsi koko oppilaspopulaatiossa myös sen tärkeimmissä (oppilasmäärältään riittävän suurissa) osajoukoissa. Tällaisia osajoukkoja ovat esimerkiksi eri sukupuolet, kieliryhmät tai maan eri alueet.

Oppilaan osaaminen tai kyky (proficiency, ability) on latentti ominaisuus, jota ei voida havaita, mutta jota voidaan mitata oppilaan tekemästä kokeesta saadun informaation avulla. Maan, alueen tai osajoukon PISA-tulos on siihen kuuluville oppilaille estimoidun kansallisen osaamisjakauman keskiarvo, ja osaamisen vaihtelua mitataan tämän jakauman varianssilla tai keskihajonnalla.

Osaamisen kansallinen jakauma muodostetaan PISA-kokeen tehneille oppilaille lasketuista heidän henkilökohtaisen osaamisensa tasoa arvioivista todennäköisyysjakaumista. Nämä jakaumat johdetaan oppilaiden pisteytetyistä koevastauksista osioanalyysin ja tilastollisen mallintamisen keinoin. Käytetty metodologia seuraa olennaisesti ns. *Bayes-tilastotieteen* periaatteita. Ennen PISA-kokeen suorittamista tehdään ennakko-oletus siitä, mikä on oppilaan luultava osaamisen taso, kun hänen koetuloksensa ei ole vielä tiedossa. Tämä oletus puetaan todennäköisyysjakauman muotoon – Bayes-tilastotieteessä puhutaan *priorijakaumasta*. Koska oppilaan osaamistaso on tuntematon ja siihen liittyy paljon epävarmuutta (uncertainty), ennakkoarvio on mielekkäämpää pukea jakauman kuin yksittäisen lukuarvon muotoon. Priorijakauman muoto kertoo, millä välillä oppilaan taso etukäteen ajateltuna todennäköisimmin on. Jakauman varianssi mittaa sitä epävarmuutta, joka oppilaan osaamisen tasosta vallitsee ennen kokeen suorittamista – suuri varianssi vastaa suurta epävarmuutta, toisin sanoen oppilaan mahdollisten osaamistasojen kirjo on laaja. Jakauman odotusarvo (eli keskiarvo) kertoo sen, mikä on oppilaan todennäköisin osaamisen taso. PISAssa osaamisen priorijakaumaksi on valittu normaalijakauma. Toisin sanoen oletetaan, että oppilaiden osaaminen noudattaa kansallisen populaation tasolla Gaussin käyrää. Etukäteen ajatellen, kun oppilaan koemenestystä ei vielä tiedetä, on siis todennäköisintä, että oppilaan osaaminen on lähellä kansallista keskitasoa. Etukäteistodennäköisyys sille, että oppilaan osaamistaso on erityisen korkea tai erityisen matala, on puolestaan pieni. Lisäksi oletetaan, että oppilaan priorijakauma riippuu lineaarisen regressiomallin tapaan suuresta joukosta taustamuuttujia. Toisin sanoen oppilaan, jolla on tietyt taustamuuttujien arvot, odotettavissa oleva osaamisen taso voi olla eri kuin oppilaan, jolla on jotkin toiset taustamuuttujien arvot. Taustamuuttujien käytön tarkoituksena on parantaa laskennan lopputuloksena saatavan kansallisen jakauman tarkkuutta: hyödyntämällä tietoa siitä, miten taustamuuttujien arvot jakautuvat oppilaspopulaatiossa, parannetaan tulosten kansallista edustavuutta. Taustamuuttujien vaikutuksia osaamiseen ei tässä vaiheessa spesifioida täsmällisemmin, vaan

ne estimoidaan kerätystä datasta varsinaisen tulosten laskennan yhteydessä.

Priorijakauman asettamisen jälkeen kerätään havaintoaineisto: otokseen valitut oppilaat tekevät PISA-kokeen, ja heidän vastauksensa pisteytetään. Bayes-tilastotieteen perusajatuksen mukaisesti priorijakaumaa päivitetään havaintoaineistosta saadun informaation (koetulokset sekä tiedot taustamuuttujien arvoista) nojalla. Tuloksena saadaan niin sanottu *posteriorijakauma* oppilaan osaamiselle. Muistettakoon, että oppilaan osaaminen on latentti suure, jota ei voi sellaisenaan havaita. Posteriorijakauma kertoo jakauman muodossa, mikä on oppilaan todennäköinen osaamisen taso, kun tiedetään hänen taustamuuttujiensa arvot ja se, miten hän onnistui PISA-kokeessa tekemissään, vaikeustasoiltaan tunnetuissa tehtävissä. PISA-tutkimuksessa myös posteriorijakauma on normaalijakauma. Oppilaskohtaisena posteriorina normaalijakauma kuvaa sitä, että todennäköisimmin oppilas on suorittanut omalla tasollaan, kun taas äärimmäinen yli- tai alisuorittaminen taas on hyvin epätodennäköistä. Jakauman varianssi mittaa oppilaan osaamista koskevan arvioinnin epävarmuutta, toisin sanoen sitä, kuinka leveä on se virhemarginaali, joka oppilaan osaamisarvioon liittyy. Jakauman keskiarvoa voidaan käyttää oppilaan PISA-pistemääränä, joskin se on oppilaan osaamisen mittarina epätäsmällinen. PISAn käyttämän metodologian tuottamat yksittäisten oppilaiden pistemäärät eivät ole osaamisarvioina harhattomia, vaan ne saattavat yliarvioida tai aliarvioida oppilaan todellisen osaamisen, oppilaan taustamuuttujien arvoista riippuen (OECD 2017, luku 9). Tämä johtuu siitä, että posteriorin laskennassa käytettävät taustamuuttujat korjaavat oppilaan pistemääräarviota kohti niiden oppilaiden keskiarvoa, joilla on samat taustamuuttujien arvot (Wu 2005, 125). Hieman paradoksaalisesti oppilaskohtaisten osaamisarvioiden tarkkuudesta tinkiminen johtaa kansallisen jakauman, erityisesti sen varianssin, luotettavampaan arvioon. Kansainvälisissä arviointitutkimuksissa halutaan erityisesti välttää tulosta, jossa osaamisen kansallinen vaihtelu aliarvioidaan. Simulointikokeilla on voitu osoittaa, että menetelmät, joilla PISA-asetelmassa voidaan tuottaa suhteellisen tarkat oppilaskoh-

taisen osaamisen estimaatit, johtavat kansallisen vaihtelun aliarviointiin (Wu 2005, 118–121).

On olennaista huomata, että PISA-koeaineistosta laskettavat tulokset ovat nimenomaan estimoituja todennäköisyysjakaumia, eivät yksittäisiä lukuarvoja sinänsä. Saaduista todennäköisyysjakaumista voidaan kuitenkin tuottaa yksittäisiä PISA-tulosta kuvaavia tunnuslukuja, tärkeimpänä kansallisen osaamisjakauman keskiarvo, jota usein kutsutaan maan PISA-pistemääräksi.

Oppilaille laskettujen posteriorijakaumien avulla voidaan siis estimoida osaamisen kansallinen jakauma ja vaihtelu varsin luotettavasti etenkin tutkimuksen pääarviointialueella (PISA 2015:ssa luonnontieteet). Useilla PISA-kierroksilla käytettyjen trenditehtävien ansiosta osaamisjakaumat pystytään skaalaamaan vertailukelpoisiksi myös eri vuosien PISA-tutkimusten välillä. Edellä mainittujen seikkojen ohella arvio yksittäisen oppilaan kokonaisosaamisesta jää kuitenkin verraten epätäsmälliseksi myös siksi, että oppilas tekee kokeessa tyypillisesti vain alle viidesosan kaikista mahdollisista tehtävistä. Latentin ilmiön posteriorijakaumaan jää tällöin väkisinkin paljon epävarmuutta, joskin tämä on olennaisesti pienempää kuin priorijakaumassa, koska oppilaan koetulokset tarkentavat osaamisarviota. Myös koulukohtaiset tulokset jäävät melko epätarkoiksi koulujen pienen otoskoon (42 oppilasta/koulu) takia: tehtävien hajauttamisen seurauksena voi käydä niin, että tietyn PISA-tehtävän tekee kokeessa vain pari koulun oppilasta.

Vuoden 2015 PISA-tutkimuksen kaikki osioanalyysit, posteriorijakaumien ja niistä johdettujen pistemäärien laskennat sekä niiden pohjana olleen kansainvälisen tietokannan rakentamisen suoritti yhdysvaltalainen tutkimuslaitos Educational Testing Service (ETS).

PISA-laskentametodologian teoreettinen tausta

Yksittäisten oppilaiden osaamista arvioivien jakaumien laskennassa käytetään psykometrian alaan kuuluvaa ns. IRT-metodologiaa (item-response theory; esim. van der Linden 2016a, 2016b; de Ayala 2009; Rutkowski, von Davier & Rutkowski 2014). Sen avulla

oppilaiden tulokset voidaan skaalata yhteismitalliselle asteikolle, sekä kansallisesti että kansainvälisesti, vaikka oppilaat eivät olisi tehneet kokeessa täsmälleen samanlaista tehtäväkokonaisuutta. Menetelmä perustuu olennaiselta osin informaatioon niiden tehtävien vaikeustasosta, jotka oppilas oli saanut tehdäkseen. Vaikeista tehtävistä suoriutuminen johtaa korkeampaan osaamisarvioon kuin helpoista tehtävistä suoriutuminen.

Menetelmän lähtökohtana on perinteisesti ollut Raschin osioanalyysimalli, jossa oppilaan todennäköisyys ratkaista tehtävä oikein riippuu logistisen funktion mukaisesti hänen (latentista) kyvystään ja tehtävän vaikeustasosta. Mitä korkeammalla tasolla oppilaan kyky on verrattuna tehtävän vaikeustasoon, sitä todennäköisemmin hän osaa ratkaista tehtävän oikein. PISA-tutkimuksessa tehtävien vaikeustasot määritetään kaikkien osallistujamaiden yhteisestä aineistosta, ja trenditehtävissä käytetään hyväksi myös tietoja niille aikaisemmilla PISA-kierroksilla saaduista vaikeustasoista. Tässä otetaan kuitenkin huomioon, että tehtävien vaikeustasot eivät välttämättä pysy vuodesta toiseen muuttumattomina. Tehtävien vaikeustasoissa voi olla myös kansallisia ja kulttuurisidonnaisia poikkeavuuksia: tehtävä voi olla jonkin maan oppilaille vaikeampi kuin jonkin toisen maan oppilaille. Jos tällainen kansallinen poikkeama on jossakin tehtävässä merkittävän suuri, se otetaan huomioon käsittelemällä kyseinen tehtävä kansallisena erikoistapauksena siten, että sille määritetään kyseisessä maassa oma, muista maista poikkeava vaikeustaso, joka huomioidaan lopullisessa pistemäärälaskennassa[1]. Ongelmalliset tehtävät pyritään kuitenkin karsimaan varsinaisesta tutkimuksesta pois jo esikoevaiheessa, ja kokonaisuutena tällaisia hyvin poikkeavasti käyttäytyviä tehtäviä onkin esiintynyt PISA-pääkokeissa vain vähän.

Dikotomisille (oikein / väärin) muuttujille tarkoitettua Raschin mallia ja siitä useampiluokkaisille (täysin oikein / osittain oikein / väärin) muuttujille laajennettua partial credit -mallia (PCM) käy-

1. Tämä menettelytapa otettiin käyttöön vuoden 2015 tutkimuksessa. Vuosina 2000–2012 ongelmalliseksi osoittautunut tehtävä eliminoitiin sen maan tulosten laskennasta (OECD 2016, Annex A5).

tettiin PISA-pistemäärien laskennassa aina vuoden 2012 tutkimukseen asti. Menetelmäkehityksen edetessä ja Raschin malliin kohdistetun kritiikin (mm. Kreiner 2011) myötä PISAssa siirryttiin vuodesta 2015 alkaen käyttämään aineistoihin paremmin sopivaa 2-parametrista logistista (2PL) osioanalyysimallia ja yleistettyä PCM-mallia (GPCM), joissa tehtävien vaikeustason ohella mallinnetaan niiden diskriminaatiokyky eli kyky erotella toisistaan eritasoisia oppilaita. Näissä malleissa otetaan lisäparametrina huomioon, että eri tehtävät erottelevat oppilaita eri tavoin.

Aiemmilla PISA-kierroksilla käytetyn laskentamallin ja menettelytapojen korvaaminen uusilla menetelmillä herätti kysymyksiä vuoden 2015 tulosten vertailukelpoisuudesta aiempien kierrosten tuloksiin nähden. Tämän vuoksi ETS-tutkimuslaitos laski uudelleen vuosien 2006–2012 PISA-tulokset vuonna 2015 käyttöön otetuilla menettelytavoilla. Osoittautui, että muutamassa maassa vanhan laskentamallin puutteilla on saattanut olla vaikutusta tuloksiin (OECD 2016, Annex A5). Tällaisista tapauksista mainittakoon Korean, Tanskan ja Viron PISA 2009 -lukutaitotulokset. Korean ja Tanskan kohdalla vanha laskentamalli antoi tuloksen, joka todennäköisesti yliarvioi maan todellisen lukutaitotason. Viron julkaistu tulos oli taas todennäköisesti aliarvio maan tasosta. Suomessa ja valtaosassa muita maita eri laskentamallien antamat tulokset olivat sen sijaan hyvin yhdenmukaiset: erot olivat korkeintaan muutamia pisteitä, ja ne mahtuivat tilastolliseen virhemarginaaliin.

Oppilaskohtaisten, oppilaan latenttia osaamista kuvaavien posteriorijakaumien määrittäminen tapahtuu yhdistämällä kaksi tilastollista mallia samaan analyysiin. 2PL- tai GPCM-osioanalyysimalli kuvaa oppilaan todennäköisyyden onnistua tehtävässä, kun oppilaalla on tietty latentti kyky tai osaamistaso ja tehtävällä on tietyt ominaisuudet (vaikeustaso ja erottelukyky). Toinen PISA-laskennoissa käytettävä malli on niin sanottu latentti regressiomalli, joka kuvaa latentin kyvyn jakauman oppilaspopulaatiossa ja joka toimii Bayes-analyysissä priorijakaumana.

Latentissa regressiomallissa oletetaan, että oppilaiden kyky on populaatiossa normaalijakautunut, mutta samalla sen odotusarvo riippuu suuresta joukosta potentiaalisia taustamuuttujia (conditio-

ning variables). Pistemääriä laskettaessa käytännöllisesti katsoen kaikki PISA-taustakyselystä sekä koe- ja otanta-asetelmasta saatavat muuttujat (esimerkiksi tehtäväklusterit ja otannan ositteet) asetetaan mallin selittäviksi muuttujiksi. Tällä on kaksi tarkoitusta. Ensiksi, kuten jo todettiin, taustamuuttujien avulla parannetaan aineistosta lasketun kansallisen osaamisjakauman tarkkuutta. Toiseksi, taustamuuttujien hyödyntäminen jo PISA-pistemääriä konstruoitaessa vähentää PISA-aineistosta myöhemmin tehtävien, samoja taustamuuttujia käyttävien data-analyysien, erityisesti regressioanalyysien mahdollista harhaa takaamalla analyyseissä käytetyn mallin tilastollisen oikeellisuuden (Wu 2005, 125). Jos osoittautuu, että jokin käytetty taustamuuttuja on yhteydessä oppilaan osaamiseen, käytetty laskentatapa varmistaa, että se näkyy myös oppilaiden PISA-pistemäärissä. Jos taas näin ei ole, taustamuuttujan kertoimeksi latentissa regressiossa tulee (likimain) nolla eikä se vaikuta tuloksiin. PISA-posteriorijakaumia määrättäessä latentti regressiomalli on vain laskentaa parantava tekninen apuväline: regressiomallin kertoimista ei olla sinänsä kiinnostuneita eikä niitä pyritä tulkitsemaan. Olennaista on saavuttaa posteriorijakauman kannalta numeerisesti mahdollisimman vakaa laskentatulos. Kun mallissa on hyvin suuri määrä taustamuuttujia, tämä ei ole itsestään selvää. Epävakausongelmat vältetään käyttämällä latentissa regressioanalyysissä selittäjinä PISA-taustakyselyn muuttujista johdettuja keskenään korreloimattomia pääkomponentteja. Kunkin maan kohdalla käytetään niin monta pääkomponenttia, että ne kattavat yhdessä 80 prosenttia alkuperäisten taustamuuttujien kokonaisvarianssista, toisin sanoen valtaosan taustamuuttujien sisältämästä informaatiosta (OECD 2017, luku 9). Tämä vähentää regressioanalyysissä tarvittavien muuttujien määrää merkittävästi ja lisää sen myötä laskennan vakautta. Koska analyysin tarkoituksena ei ole muodostaa tulkinnallisia selitysmalleja, pääkomponenttien korreloimattomuus ja mahdollinen heikko tulkittavuus eivät muodosta tässä ongelmaa.

Ennen varsinaisten posteriorijakaumien laskentaa määritetään 2PL- tai GPCM-osioanalyysimallien parametrien (ts. tehtävien vaikeustasojen ja erottelukykyjen) estimaatit kaikkien PISA-

osallistujamaiden yhteisestä datasta. Soveltuvin osin hyödynnetään myös aikaisempien PISA-kierrosten kansainvälistä dataa. Saatuja estimaatteja käytetään posteriorijakaumien laskennassa kiinteinä vakioina (vrt. ns. empiirinen Bayes -menetelmä), jolloin estimoitavaksi jäävät osioanalyysimallien osalta oppilaan osaaminen (kykyparametri) ja latentin regression osalta pääkomponenttien regressiokertoimet ja virhevarianssi. Nämä parametrit estimoidaan datasta samanaikaisesti, ja estimoinnin tuloksena saadaan lopulta jokaiselle oppilaalle henkilökohtainen posteriorijakauma. Estimointimenetelmänä käytetään epätäydellisten, tyypillisesti puuttuvaa tietoa sisältävien aineistojen käsittelyyn kehitettyä EM-algoritmia (Expectation-Maximization Algorithm; sen soveltamisesta arviointitutkimukseen ks. Mislevy 1985).

Posteriorijakaumat eivät ole ilmaistavissa suljetussa kaavamuodossa, mutta niistä voidaan tuottaa numeerisia realisaatioita eli otoksia simuloimalla. Näitä realisaatioita kutsutaan nimellä *plausible values* (lyhennettynä PV) – kelvollista tai vakiintunutta suomennosta ei termille ole. Kaikki PISA-aineistoista laskettavat kansalliset ja kansainväliset tulokset perustuvat näihin PV-lukuarvoihin. Vuodesta 2015 alkaen posteriorijakaumista on generoitu data-analyysejä varten 10 PV-arvoa, joiden avulla haarukoidaan oppilaiden todennäköinen osaamistaso ja sen arviointiin liittyvää epävarmuutta kuvaava satunnaisvaihtelu. Aiemmilla kierroksilla käytettiin viittä PV-arvoa. Ensimmäisessä, vuoden 2000 PISA-tutkimuksessa PV-arvoille määriteltiin kansainvälinen skaalaus siten, että osallistuneiden OECD-maiden keskiarvoksi asetettiin 500 ja keskihajonnaksi 100. Näin skaalattujen PV-arvojen mittayksikköä kutsutaan yleisesti PISA-pistemääräksi. Myöhempien PISA-kierrosten PV-arvot on skaalattu tälle samalle pistemääräasteikolle trenditehtävien avulla.

Plausible value -arvot edustavat siis jakaumaa, joka kuvaa PISA-pistemääräasteikolla, mikä on oppilaan todennäköinen osaamisen taso, kun tiedetään, miten hän onnistui tekemissään vaikeustasoltaan ja erottelukyvyltään tunnetuissa tehtävissä, ja lisäksi otetaan huomioon se, millainen on taustamuuttujien suhteen hänen kaltaistensa oppilaiden tyypillinen osaamistaso. Kansalliset tulok-

set saadaan yhdistämällä näistä oppilaiden posteriorijakaumista saadut henkilökohtaiset tulokset.

Jos oppilas ei ole tehnyt lainkaan jonkin arviointialueen, esimerkiksi matematiikan, tehtäviä, hänelle tuotetaan laskennallinen posterioriarvio mallintamalla hänen osaamisensa muiden arviointialueiden tehtävistä saatujen tulosten avulla. Mallintamisen mahdollistaa se, että luonnontieteiden, lukutaidon ja matematiikan osaamisen väliset korrelaatiot ovat PISA-tutkimuksissa erittäin korkeita. Suomen PISA 2015 -aineistossa ne olivat sekä luonnontieteiden ja lukutaidon että luonnontieteiden ja matematiikan välillä 0,87. Lukutaidon ja matematiikan välinen korrelaatio oli 0,79.

ETS-tutkimuslaitos käyttää posteriorijakaumien (PV-arvojen) ja latentin regressioanalyysin vaatimiin laskentoihin DGROUP-tietokoneohjelmistoa (Rogers, Tang, Lin & Kandathil 2006).

PISA-tutkimuksissa oppilaat ryhmitellään arvioidun osaamisensa perusteella 6–7 suoritustasolle, joista korkein edustaa huippuosaamista ja matalin erittäin heikkoa osaamista. Luonnontieteissä ja lukutaidossa suoritustasoja on seitsemän, matematiikassa kuusi. Oppilaan suoritustaso määräytyy PV-arvoille asetettujen pisterajojen mukaan: mitä korkeamman pisterajan oppilaan PV-arvo ylittää, sitä korkeampi on hänen suoritustasonsa.

Pisterajat määritellään ns. standard setting -menetelmän avulla. Menetelmä on yleisessä käytössä varsinkin vieraiden kielten arvioinneissa, joissa sovelletaan standard settingiin perustuvia eurooppalaisia taitotasoasteikkoja (Ahola, Neittaanmäki & Hirvelä 2016; Kaftandjieva 2010). PISAssa sisältöalueen asiantuntijaryhmä pohtii etukäteen, mitkä seikat ovat olennaisia eri suoritustasoja vastaavassa osaamisessa ja määrittelee sen perusteella koetehtävät vastaamaan kukin tiettyä suoritustasoa. Sen jälkeen kaikille suoritustasoille laaditaan sanallisesti kuvatut kriteerit sen pohjalta, millaisia kullekin suoritustasolle määritellyt tehtävät ovat. Pisterajojen laatiminen on aina jonkin verran subjektiivista, mutta virheluokituksen riski on mahdollista estimoida. Pisterajojen määrittäminen vaatii riittävän määrän tehtäviä (Olsen & Nilsen 2017).

PISAssa pisterajat on siis johdettu tehtäville etukäteen kiinnitettyjen suoritustasojen perusteella, ja kunkin suoritustason

määrittelyyn vaaditaan keskimäärin 18 tehtävää. Eri tehtävät vaativat oppilaalta eritasoisia valmiuksia, ja oppilaat saavat tehdäkseen kaikentasoisia tehtäviä. Huippuosaamista vastaavan pisterajan ylittäminen edellyttää, että oppilas on suoriutunut onnistuneesti riittävän monesta korkeimman suoritustason tehtävästä. Alempien suoritustason rajat määritellään vastaavasti.

Standard setting -menetelmän avulla voidaan suoritustasojen lisäksi määrittää taso, joka on edellytys hyväksytylle suoritukselle, jos sellaista jossain oppiaineessa tai testissä tarvitaan. PISAn standard setting edesauttaa tulosten ja niiden kehityksen tarkastelua eri mittauskertojen välillä. Lisäksi pisterajojen avulla voidaan määrittää kynnystaso, joka oppilaan tulisi saavuttaa, jotta hänellä voidaan arvioida olevan riittävät tiedot ja taidot selvitä arkielämän tilanteissa ja jatko-opinnoissa (Olsen & Nilsen 2017). PISAssa tällaiseksi kynnystasoksi on määritetty suoritustaso 2.

PISA-tulosten data-analyysi

PISAn laskentametodologia tuottaa jokaiselle tutkimukseen osallistuneelle oppilaalle jokaisesta arviointialueesta – luonnontieteet, matematiikka, lukutaito sekä vaihtuvat alueet – 10 PV-arvoa, jotka mittaavat oppilaan osaamisen tasoa kyseisessä arviointialueessa. ETS on laskenut valmiiksi kunkin arviointialueen PV-arvot julkaistuihin PISA-datoihin. Jokainen kymmenestä PV-arvosta on yhtä hyvä arvio oppilaan osaamisesta, ja niiden varianssi kuvaa sitä epävarmuutta, joka liittyy oppilaan latentin kyvyn arviointiin rajallisella tehtävävalikoimalla. Tosin saman oppilaan PV-arvot ovat usein varsin lähellä toisiaan, varsinkin jos ne eivät ole osaamisjakauman ääripäissä. PV:t korreloivat keskenään hyvin voimakkaasti: vuoden 2015 pääarviointialueen luonnontieteiden kymmenen PV:n korrelaatiot olivat välillä 0,91–0,92. Periaatteessa PV-arvojen keskiarvo on paras aineistosta saatavissa oleva estimaatti oppilaan osaamiselle, sillä se estimoi oppilaan posteriorijakauman odotusarvoa tarkemmin kuin yksittäiset PV-arvot. Keskiarvon käyttöä PISA-aineiston jatkoanalyyseissä ei kuitenkaan suositella, sillä se joh-

taa osaamisen kansallisen vaihtelun aliarviointiin ja edelleen liian merkitseviin analyysituloksiin. Sen sijaan PISA-konsortio suosittelee data-analyyseissä käytettäväksi lähestymistapaa, joka soveltaa alun perin puuttuvien havaintojen paikkaamiseen kehitettyä ns. moni-imputointia (OECD 2009a, luku 8).

Moni-imputointi (multiple imputation, Rubin 1987) on puuttuvien havaintojen käsittelyyn liittyvä lähestymistapa, jossa pyritään aidosti ottamaan huomioon aineistossa olevien puuttuvien tietojen paikkaamiseen (eli imputointiin) liittyvä epävarmuus. Jos tuntematon havaintoarvo korvataan yhdellä lukuarvolla, tämä on tavallisen tilastollisen analyysin näkökulmasta yhtä luotettava muuttujan arvo kuin oikea havaittu arvo olisi. Periaatteessa sama pätee, jos puuttuva havainto korvattaisiin jonkin kelvollisen paikkausmenetelmän sijasta millä tahansa lukuarvolla. Moni-imputoinnissa puuttuva havainto paikataan yhden arvon sijasta useilla vaihtoehtoisilla lukuarvoilla, ja tutkitaan, kuinka paljon analyysitulokset vaihtelevat sen mukaan, mitä lukuarvoa paikkaamisessa on käytetty. Tällä vaihtelulla mitataan puuttuvien tietojen paikkaamiseen liittyvää epävarmuutta. Rubin on osoittanut, että yhdistämällä vaihtoehtoisilla lukuarvoilla suoritettujen analyysien tulokset keskiarvotuloksiksi ja lisäämällä analyysitulosten välinen vaihtelu, ns. imputointivarianssi, tilastollisiin data-analyyseihin ylimääräiseksi vaihtelulähteeksi saadaan puuttuvia tietoja sisältävien aineistojen analyyseistä luotettavia tuloksia. Arviointitutkimuksissa laskettavat PV-arvot voidaan ymmärtää moni-imputoinnin hengessä oppilaan osaamista koskevaa puuttuvaa tietoa paikkaaviksi vaihtoehtoisiksi arvoiksi.

PISA-aineistojen analyysiin sovellettuna moni-imputointiin perustuva data-analyysi on pelkistetysti seuraava. Haluttu tilastollinen analyysi lasketaan jokaiselle kymmenelle PV:lle erikseen. Saadaan kymmenen analyysitulosta, joista lasketaan keskiarvo, ja tämä keskiarvo on analyysin varsinainen tulos. Esimerkiksi Suomen virallinen kansallinen keskiarvo on kymmenen eri PV-keskiarvon keskiarvo. Se on kansallisen osaamisjakauman odotusarvon (eli todennäköisyysjakauman keskiarvon) estimaatti. Vastaavasti Suomen virallinen kansallinen varianssi saadaan laskemalla ensin kunkin

PV-arvon varianssi Suomen oppilasaineistossa ja sen jälkeen näiden kymmenen varianssin keskiarvo.

Estimoitujen tunnuslukujen, joista edellä esimerkkeinä annettiin kansallinen keskiarvo ja varianssi, estimointitarkkuutta mittaavat otantavarianssit lasketaan siten, että kymmenelle eri PV-muuttujalle laskettujen tunnuslukujen varianssiestimaatit (ts. 10 varianssia) yhdistetään keskiarvoksi, ja tähän keskimääräiseen varianssiin lisätään imputointivarianssi, joka on laskettujen kymmenen varianssiestimaatin välinen varianssi. Tunnusluvun keskivirhe on näin lasketun otantavarianssin, eli kymmenen varianssiestimaatin keskiarvon ja imputointivarianssin summan, neliöjuuri. Imputointivarianssilla otetaan huomioon se lisäepävarmuus, joka johtuu siitä, että oppilaan tarkkaa osaamistasoa ei tiedetä, vaan sitä joudutaan kartoittamaan kymmenen vaihtoehtoisen arvion välityksellä (se, mitä oppilaan osaamisesta voidaan aineiston perusteella sanoa, on todennäköisyysjakauma, ei yksittäinen pistemäärä, ja jokainen PV-arvo on siis yhtä hyvä arvio oppilaan todellisesta osaamisesta). Imputointivarianssista aiheutuvan komponentin lisääminen tunnuslukujen keskivirheisiin ja sitä myötä edelleen merkitsevyystesteihin ja virhemarginaaleihin on olennaista PISA-tuloksia raportoitaessa. Imputointivarianssi on korjaustermi, jota ilman PISA-tulosten tarkkuus tulisi tilastollisissa analyyseissä yliarvioiduksi. Tällöin esimerkiksi merkitsevyystestit antaisivat virheellisen merkitseviä tuloksia, toisin sanoen p-arvot olisivat virheellisen pieniä.

Samaa menettelyä sovelletaan, kun lasketaan tuloksia koko oppilaspopulaation sijasta sen osajoukoille, esimerkiksi tytöille ja pojille. Yksinkertaisten perustunnuslukujen ohella menetelmää voidaan soveltaa periaatteessa mihin tahansa tilastolliseen analyysiin, esimerkiksi regressioanalyysiin.

Monissa mutkikkaammissa tilastollisissa analyyseissä, esimerkiksi rakenneyhtälömallien rakentamisessa, moni-imputoinnin periaatteen mukainen laskenta saattaa osoittautua varsin työlääksi, varsinkin jos mallien rakentaminen on monivaiheinen prosessi ja edellyttää monia erilaisia kokeiluja. Käyttämällä kymmenen PV:n sijaan vain yhtä PV-arvoa analyysin riippuvana muuttujana

voidaan suorittaa harhaton analyysi siinä mielessä, että tunnuslukujen varianssit ja keskivirheet eivät aliarvioidu. Toki analyysitulos on tällöin vain approksimaatio sille tulokselle, joka saataisiin kymmenen PV-arvon oikeaoppisella moni-imputaatioanalyysillä. Saatujen kokemusten mukaan yhdellä PV:llä saatu likimääräinen tulos on sisällöllisten päätelmien tekemiseen kuitenkin yleensä varsin hyvä. Sen sijaan PV-arvojen keskiarvon käyttäminen tilastollisissa analyyseissä selitettävänä muuttujana johtaisi varianssien ja keskivirheiden aliarviointiin ja tulosten näennäiseen tarkkuuteen.

Koska PISA-aineisto kerätään verraten mutkikkaalla kaksivaiheisella otanta-asetelmalla, sen asianmukainen analysointi edellyttää tämän asetelman huomioon ottavia menetelmiä. Otanta-asetelmasta johtuen eri koulut ja oppilaat niiden sisältä poimitaan vaihtelevilla todennäköisyyksillä, mikä saattaa aiheuttaa vääristymää lopullisen oppilasotoksen kokoonpanoon. Tämän korjaamiseksi on PISAn data-analyyseissä syytä käyttää otantapainoja rutiininomaisesti. Tarvittavat painomuuttujat on laskenut ETS, ja ne ovat valmiiksi saatavilla PISA-datoissa. Erityisen oleellista on ottaa huomioon se, että oppilasaineisto on klusteroitunut, toisin sanoen saman koulun oppilaat muistuttavat toisiaan jossain määrin enemmän kuin eri kouluista satunnaisesti poimitut oppilaat. Tähän liittyen heidän tuloksensa eivät välttämättä ole toisistaan riippumattomia: heillä on sama kouluympäristö, samat opettajat, usein sama asuinalue jne. Oppilaiden klusteroitumisen astetta mitataan ns. sisäkorrelaatiolla (intra-cluster correlation), joka saa arvoja väliltä 0–1. Positiivinen sisäkorrelaatio tarkoittaa, että koulun oppilaat ovat homogeenisempia kuin täysin satunnaiset oppilaat – sitä enemmän, mitä suurempi korrelaation arvo on. Ääritapauksessa oppilaiden tulokset ovat täysin identtiset. Tällöin homogeenisuus on täydellinen ja sisäkorrelaation arvo on 1. Jos sisäkorrelaatio on nolla, koulun oppilaat ovat toisistaan riippumattomat eikä homogeenisuutta havaita. Suomessa, jossa koulujen väliset erot ovat kansainvälisesti ottaen vähäisiä, myös saman koulun oppilaiden homogeenisuus on verraten vähäistä. Suomessa sisäkorrelaatio on koulututkimuksissa, myös PISAssa, tyypillisesti suuruusluokkaa

0,05–0,10. Kouluilla on Suomessakin oppilaita homogenisoivaa vaikutusta, mutta tämä on varsin lievää.

Vaikka oppilaiden klusteroituminen on Suomen kouluaineistoissa hyvin vähäistä, sen huomiotta jättäminen vääristää tilastollisten analyysien tuloksia, koska saman koulun oppilaat eivät ole toisistaan riippumattomia. Yleiset tilastolliset perusmenetelmät olettavat, että havaintoyksiköt ovat täysin riippumattomia, ja jos tämä ehto ei ole voimassa, saatavat tulokset ovat näennäisen tarkkoja ja analyysit näennäisen voimakkaita (tunnuslukujen keskivirheet aliestimoituvat – sitä enemmän, mitä voimakkaampi on sisäkorrelaatio). PISAn kaltaista kaksitasoisella otanta-asetelmalla kerättyä klusteroitunutta aineistoa voidaan analysoida luotettavasti kahdella vaihtoehtoisella lähestymistavalla. Ensimmäinen on ns. malliperusteinen (model-based) lähestymistapa, jonka tyypillisin esimerkki ovat monitasomallit (multilevel models). Niissä aineiston hierarkkisuus ja klusteroituneisuus hoidetaan mallintamalla aineisto sopivasti. Toinen, laajojen kansainvälisten arviointitutkimusten yhteydessä suosittu tapa on käyttää tilastollisissa analyyseissä erityisiä survey-menetelmiä. Näiden alle luetaan ns. asetelmaperusteiset (design-based) varianssin estimointimenetelmät, joihin muun muassa kuuluu tilastotieteen piirissä tunnettu jackknife-lähestymistapa. Sen eräs variaatio on ns. Fay-modifioitu Balanced Repeated Replication (BRR)-menetelmä, joka on PISA-konsortion suosittelema ja tukema varianssin estimointimenetelmä (OECD 2009a, luku 4).

Jackknife on idealtaan yksinkertainen lähestymistapa varianssin estimointiin laskennallisesti muutoin hankalissa tilanteissa. Se on sovellettavissa monenlaisiin tilanteisiin, eikä se edellytä aineistolta minkäänlaista jakaumaoletusta (kuten normaalijakaumaa) toisin kuin malliperusteiset menetelmät. Jackknifen perusideana on uusiokäyttää jo poimittua otosta poistamalla siitä yksi havainto ja laskemalla sen jälkeen haluttu tunnusluku. Tätä toistetaan käymällä koko aineisto läpi siten, että siitä poistetaan vuoron perään eri havainto. Näin saadaan joukko ns. replikaattidatoja, jotka eroavat toisistaan vain yhden havainnon osalta. Jokaisesta replikaattidatasta lasketaan samalla tavalla sama tunnusluku, jonka arvo

vaihtelee hieman datasta toiseen. Näiden arvojen varianssilla estimoidaan alkuperäisestä kokonaisesta aineistosta lasketun tunnusluvun (otanta)varianssia. PISAssa käytetty Fay-modifioitu BRR-menetelmä on jackknife-idean muunnelma, jossa sen sijaan, että aineistosta poistettaisiin vuoron perään yksi havainto, havaintoyksiköiden otantapainoja varioidaan vuoron perään tietyn systeemin mukaan siten, että alkuperäisestä datasta tuotetut replikaattidatat eroavat toisistaan vain havaintojen otantapainojen osalta. Nämä replikaattidatat sisältävät siis täsmälleen samat alkuperäiset havainnot, ainoastaan näiden otantapainoissa eli ns. replikaattipainoissa on eroja. Jokaisesta replikaattidatasta lasketaan sen replikaattipainoja käyttäen sama tunnusluku, ja saatujen tunnuslukujen varianssilla estimoidaan halutun alkuperäisen tunnusluvun varianssia jackknife-periaatteen mukaisesti.

BRR-menetelmä varioi replikaattipainoja PISAn kaksitasoista, koulu- ja oppilasotannoista koostuvaa otanta-asetelmaa noudattaen. Siten asetelman ominaisuudet, mukaan lukien oppilaiden klusteroituminen, tulevat varianssien ja keskivirheiden laskennassa täysin huomioiduksi. PISA-konsortio on laskenut julkaistuihin PISA-datoihin BRR-estimoinnissa tarvittavat replikaattipainot valmiiksi. Näitä on aineistossa 80 kpl, toisin sanoen PISAn käyttämä BRR-estimointi perustuu 80:een valmiiksi muodostettuun replikaattidataan.

Koska PISA-tutkimuksia toistetaan kolmen vuoden välein, eri vuosien väliset muutostrendit ovat kiinnostuksen kohteena. Kuten jo aiemmin todettiin, eri vuosien PISA-tulokset pystytään skaalaamaan yhdenmukaiselle asteikolle hyödyntämällä ns. linkki- tai trenditehtäviä, joita käytetään useilla PISA-kierroksilla peräkkäin. Kaikki tehtävät eivät ole kuitenkaan samoja peräkkäisilläkään PISA-kierroksilla. Samoin tehtävien vaikeustasoissa voi tapahtua muutoksia vuosien välillä. Siten lukutaidon, matematiikan tai luonnontieteiden osaamisen mittareita ei voida pitää täysin identtisinä PISA-kierroksesta toiseen, mikä aiheuttaa muutostrendeihin ylimääräistä vaihtelua ja päätelmiin lisäepävarmuutta. Tämän vuoksi eri PISA-vuosien välisiä muutoksia mittaavien tunnuslukujen (tavallisesti keskiarvoerojen) keskivirheisiin tulee

lisätä vielä ylimääräinen komponentti, niin sanottu linkitysvirhe (linking error). PISA-konsortio on laskenut keskiarvovertailuihin liittyvien linkitysvirheiden numeeriset arvot valmiiksi kaikille tärkeimmille arviointialueille (lukutaito, matematiikka, luonnontieteet) eri PISA-kierrosten välisissä trendianalyyseissä käytettäväksi. Vuoden 2015 vertailuissa tarvittavat linkitysvirheiden arvot on julkaistu OECD:n PISA 2015 -raportin liitteessä (OECD 2016, Annex A5, taulukko A5.2).

Australialainen ACER-tutkimuslaitos on tuottanut PISA-aineistojen analysointia varten SAS- ja SPSS-tietokoneohjelmistojen yhteydessä käytettävissä olevia makroja. Makrot suorittavat sekä PV-arvojen moni-imputointianalyysin että BRR-varianssiestimoinnin edellä kuvatulla tavalla. Niillä voidaan laskea perustunnuslukuja sekä suorittaa regressio-, varianssikomponentti- ja sekamallianalyysejä. Tietoa näistä makroista saa OECD:n julkaisemista data-analyysioppaista (OECD 2009a; OECD 2009b). Uudempi vaihtoehto on käyttää muun muassa PIRLS- ja TIMSS-tutkimuksia toteuttavan IEA-järjestön Data Processing Centerin rakentamaa IDB Analyzer ohjelmistoa[2], joka toimii SPSS- ja SAS-ympäristöissä ja jolla voidaan toteuttaa valikkopohjaisesti edellä kuvatun kaltaisia data-analyysejä paitsi PISA-aineistoille myös monien muiden kansainvälisten arviointitutkimusten (PIRLS, TIMSS, PIAAC, TALIS, ICCS) aineistoille.

Lopuksi

Kansainvälisten arviointitutkimusten kuten PISAn metodologia kattaa hyvin monenlaisia aihealueita. Koululaisia tai opiskelijoita koskevissa arvioinneissa sovelletaan poikkeuksetta hierarkkista ryväsotanta-asetelmaa, jossa ensin poimitaan kouluja ja sen jälkeen kokonaisia luokkia tai yksittäisiä koulun oppilaita. Aineiston kansallisen edustavuuden varmistamiseksi kouluotanta toteutetaan ositettuna otantana. Ositetulla ryväsotannalla kerätyn aineis-

2. Ohjelmisto on ladattavissa osoitteesta http://idbanalyzerv3.software.informer.com/3.1/

ton pätevä tilastollinen analyysi edellyttää otanta-asetelman mukaisia menetelmiä; tavallisimmat perusmenetelmät (esimerkkeinä vaikkapa keskiarvojen t-testi tai yksisuuntainen varianssianalyysi) eivät sellaisenaan sovellu. Analyysimenetelmien tulee pystyä ottamaan huomioon sekä asetelmasta johtuvat havaintoyksiköiden vaihtelevat poimintatodennäköisyydet että havaintoaineiston hierarkkisuus ja samaan kouluun kuuluvien oppilaiden korreloituneisuus – kaikki nämä seikat ovat tilastollisten perusmenetelmien oletusten vastaisia. Arviointiaineistojen kerääminen ja menestyksellinen analysointi edellyttävätkin otantamenetelmien, survey-tilastotieteen ja siihen kuuluvien asetelmaperusteisten analyysimenetelmien hallintaa.

Arviointitutkimusten koejärjestelyt ovat monimutkaisia. Tutkimusten päätavoite on tuottaa mahdollisimman kattavaa ja luotettavaa tietoa koulutusjärjestelmän alaisuuteen kuuluvien oppilaiden osaamisen jakautumisesta. Tästä seuraa, että yksittäisten oppilaiden osaamisen arviointiin ei kiinnitetä suurta huomiota, vaan metodiset valinnat tähtäävät kansallisen osaamistason ja vaihtelun kokonaisvaltaiseen määrittämiseen. Kokonaisvaltainen arviointi edellyttää esimerkiksi sitä, että kansalliseen otokseen valittu oppilasjoukko saa tehdäkseen laajan valikoiman koetehtäviä, jotka vaihtelevat sekä aihepiireiltään että vaikeustasoiltaan. Koska kokeeseen käytettävä aika on syytä pitää kohtuullisena, yksi oppilas voi tehdä vain suppean osan kaikista mahdollisista tehtävistä. Hajauttamalla tehtävät siten, että eri oppilaat tekevät eri tehtäviä, ja kokoamalla näin saadut koetulokset yhteen saadaan monipuoliseen tehtäväkokoelmaan perustuva kansallisen tason osaamisarvio kohtuullisessa koeajassa. Kokeen luotettavuus ja kansallisen tason harhattomuus edellyttävät monien seikkojen suhteen tasapainoista koejärjestelyä. Erityyppisiä tehtäviä tulee esiintyä kokeessa riittävän paljon ja tasaisesti: mikään aihepiiri ja vaikeustaso ei saisi olla aineistossa yli- tai aliedustettu. Myös yhden oppilaan tulee saada tehdäkseen (kohtuuden rajoissa) mahdollisimman monentyyppisiä tehtäviä. Tehtäväkombinaatioiden tulee olla tasapainoisia siten, että mitkään tehtävätyypit eivät esiinny liian usein tai liian harvoin yhdessä. Myös tehtävien järjestys tulee tasapainottaa.

On tavallista, että oppilaan suoriutuminen heikkenee koeajan loppua kohti. Siksi koejärjestelyllä tulee huolehtia, että kaikkia tehtävätyyppejä ja aihealueita esiintyy tasaisesti kokeen alku- ja loppupuolella.

PISA-tulosten laskenta perustuu keskeisesti Bayes-tilastotieteeseen. Siinä oppilaiden osaamiselle asetetaan priorijakauma, joka on etukäteisoletus osaamisjakauman muodosta. PISAssa ja muissa arviointitutkimuksissa tämä jakauma on normaalijakauma: on realistista olettaa, että oppilaiden latentti osaaminen jakautuu kansallisessa kohdepopulaatiossa Gaussin käyrän tavoin. Kun oppilaan kokeessa osoittama osaaminen on arvioitu (ts. koetehtävät on pisteytetty ja tehtävien vaikeustasot on huomioitu), priorijakauma päivittyy posteriorijakaumaksi, josta kaikki kansalliset (ja erilaisia osajoukkoja koskevat) PISA-tulokset voidaan johtaa. Posteriorijakauman määrittämisessä hyödynnetään koetulosten ohella suurta joukkoa taustamuuttujia. Näin posteriorijakauma saadaan myös taustamuuttujien suhteen niin edustavaksi kuin sen kerätyn aineiston perusteella on mahdollista. Posteriorijakauma antaa arvion oppilaiden todennäköisestä osaamistasosta, kun tiedetään, miten he onnistuivat tekemissään vaikeustasoltaan ja erottelukyvyltään tunnetuissa tehtävissä.

Kansallisten tulosten määrittämisessä käytetään edellä mainittuja asetelmaperusteisia menetelmiä sekä ns. moni-imputointiin perustuvaa laskentatapaa. Menetelmävalintojen johtavana ajatuksena on ottaa aidosti huomioon kaikki se epävarmuus, joka liittyy latentin osaamisen arviointiin rajallisella kokeella ja otoksella, ja estää näin kansallisen osaamisen vaihtelun aliarviointi (ja samalla tulosten tarkkuuden yliarviointi).

PISA-tutkimukset ovat saaneet osakseen kritiikkiä niin ulkomailla kuin Suomessakin. Suuri osa kritiikistä kohdistuu tulosten tulkintaan ja tehtyihin johtopäätöksiin, mutta myös laskentametodeja on kyseenalaistettu. Kansallisten posteriorijakaumien keskiarvojen perusteella laaditut maiden rankinglistat ovat saaneet paljon huomiota. Vaikka suurin osa PISA-tulosten saamasta julkisuudesta lieneekin kohdistunut näihin listoihin, ne eivät kuitenkaan ole se päämäärä, johon PISA-tutkimusten tekemisellä pyritään.

Mikään PISAn metodinen valinta ei tähtää maiden paremmuusjärjestyksen määrittämiseen. Rankinglistat ovat kansainvälisten tulosten yksi sivutuote, joka ei tutkimuksellisesta näkökulmasta lainkaan kuulu tärkeimpien joukkoon. Rankinglistoihin vaikuttaa sitä paitsi oleellisesti se, mitkä maat ovat kulloisellakin kierroksella osallistuneet tutkimukseen.

PISAn oppilaskyselyjen luotettavuutta on myös epäilty, ja tälle kritiikille onkin jossain määrin aihetta. On huomattu, että etenkin heikoimmin kokeessa menestyneillä oppilailla on taipumusta vastata kysymyksiin ristiriitaisesti ja huolimattomasti tai jättää kokonaan vastaamatta. Näyttää myös siltä, että yllättävän moni nuori ei tiedä esimerkiksi vanhempiensa koulutustasoa.

PISA-tulosten perusteella on usein haluttu tehdä päätelmiä maiden koulutusjärjestelmistä ja erityisesti niissä tehtyjen uudistusten vaikuttavuudesta. Tällaiset päätelmät ovat kuitenkin helposti hätiköityjä. Etenkin kahden peräkkäisen PISA-kierroksen välinen kolmen vuoden jakso on liian lyhyt, jotta koulutuspolitiikassa tehtyjen muutosten mahdollinen vaikutus näkyisi tuloksissa. (Borman kumppaneineen (2003) on arvioinut, että uudistusten näkyminen oppimistuloksissa vie tyypillisesti 8 vuotta). Lisäksi koulutusjärjestelmien ohella muussa yhteiskunnassa tapahtuu paljon muutoksia, joilla kaikilla voi olla oma yhteytensä heikkeneviin tai paraneviin oppimistuloksiin. PISA-tutkimukset ovat sarja poikkileikkaustutkimuksia, eikä niiden avulla voida tehdä kausaalipäätelmiä. Tehtyjen uudistusten vaikuttavuutta on vaikea arvioida, koska samalta ajanjaksolta ei samasta maasta ole käytettävissä kontrolliaineistoa: millaista tulosten kehitys olisi ollut, jos uudistuksia ei olisi tehty. PISA-tulokset kertovat sangen luotettavasti, mikä on maan tai koulutusjärjestelmän senhetkinen tila PISA-kokeen testaamilla oppimistuloksilla mitattuna.

PISAssa käytettyjä tilastollisia menetelmiä, erityisesti osioanalyysimalleja, ovat kritisoineet muun muassa tanskalaiset Kreiner ja Christensen (2014). Tosin he keskittyvät tarpeettoman paljon tulosten käyttöön maiden rankinglistojen laadinnassa. PISAssa alun perin käytetyn Raschin osioanalyysimallin rajoitukset ovat kiistattomat, ja PISA-konsortio siirtyikin sittemmin käyttämään Raschin

mallia joustavampaa 2PL-osioanalyysimallia. PISAn keräämän aineiston valtava koko aiheuttaa kuitenkin ongelmia minkä tahansa tilastollisen mallin sovittamiseen: on tunnettua, että riittävän suuressa aineistossa sisällölliseltä kannalta merkityksetönkin poikkeama saadaan tilastollisesti merkitseväksi. Siten hyvin suureen aineistoon on käytännössä mahdoton löytää tilastollista mallia, joka sopisi siihen täydellisesti. Aineiston suuruus vaikuttaa myös siihen, että edistyneempien, entistä moniparametrisempien ja sen myötä aineistoon paremmin sopivien mallien vaatima laskentakuorma tulee käytännössä kohtuuttoman raskaaksi (vrt. Kreiner 2011). Toisaalta aineiston suuri koko ja varsinkin tehtävien suuri määrä auttavat kuitenkin lieventämään osioanalyysimallin mahdollisista puutteista johtuvia ongelmia – aineiston eri osissa esiintyvät virheet keskimääräistyvät pois. Tätä havainnollistaa se, että laskettaessa aiempien PISA-kierrosten tulokset sekä vanhalla että uudella mallilla tulosten väliset erot mahtuivat tilastolliseen virhemarginaaliin (OECD 2016, Annex A5). PISA-konsortion metodologisesta kehittämistyöstä vastaava ETS-tutkimuslaitos on suhtautunut vakavasti PISAn metodeja koskevaan kritiikkiin ja tehnyt laajoja kokeita erilaisilla malleilla ja laskentatavoilla. Todella vakavia, virheellisiin päätelmiin johtaneita puutteita ei PISAssa käytetyistä menetelmistä ole toistaiseksi löydetty.

Vaikka PISA-tutkimuksia hallinnoiva OECD ei ole tiedemaailmassa toimiva taho, se hyödyntää kaikessa PISA-tutkimuksiin liittyvässä toiminnassa aina tehtävien laadinnasta lopullisen datan analysointiin korkean tason akateemista asiantuntemusta. Käytettyä metodologiaa – sekä laskentamalleja että koejärjestelyjä – kehitetään PISA-konsortiossa jatkuvasti. Esimerkkinä tästä mainittakoon, että PISA 2018 -tutkimuksessa otetaan ensimmäisiä askelia adaptiivisen koejärjestelyn suuntaan. Siinä oppilaan saamat tehtävät määräytyvät kokeen kuluessa sen mukaan, miten hän onnistuu kokeen alussa tekemissään tehtävissä. Ajatuksena on määrittää yksittäisen oppilaan osaamisen taso entistä tarkemmin esimerkiksi siten, että hyvää osaamista osoittaneen oppilaan ei tarvitse käyttää koeaikaa hänelle selvästikin liian helppojen tehtävien tekemiseen. Toinen esimerkki on vuoden 2015 PISAssa käyttöön

otettu uudistus, jossa muutostrendien määrittämisessä tarvittavien linkkitehtävien vaikeustasojen skaalaamisessa käytetään hyväksi edellisen PISA-kierroksen datan lisäksi myös aikaisempien kierrosten dataa. Vaikka PISA-tutkimukset eivät ole täydellisiä ja niistä löytyy – kaikkien muiden empiiristen tutkimusten tavoin – kritisoitavaa, niiden aineistot kuuluvat varauksetta luotettavimpien ja korkeatasoisimpien koulutustutkimuksen piirissä saatavilla olevien aineistojen joukkoon.

Lähteet

Ahola, S., Neittaanmäki, R. & Hirvelä, T. 2016. Puheen ymmärtämisen tehtävien taitotasoille asettaminen yleisissä kielitutkinnoissa. Teoksessa A. Huhta & R. Hildén (toim.). *Kielitaidon arviointitutkimus 2000-luvun Suomessa*. AFinLA-e: Soveltavan kielitieteen tutkimuksia 9. Suomen soveltavan kielitieteen yhdistys.

de Ayala, R. J. 2009. The theory and practice of item response theory. New York: Guilford Press.

Borman, G. D., Hewes, G. M., Overman, L. T. & Brown, S. 2003. Comprehensive school reform and achievement: a meta-analysis. Review of Educational Research, 73, 125–230.

Goldstein, H. 2010. Multilevel statistical models. 4th edition. Chichester: Wiley.

Kaftandjieva, F. 2010. Methods for Setting Cut Scores in Criterion-referenced Achievement Tests. A comparative analysis of six recent methods with an application to tests of reading in EFL. European Association for Language Testing and Assessment. Arnhem: Cito.

Kreiner, S. 2011. Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment. Research report 11/1, Department of Biostatistics, University of Copenhagen.

Kreiner, S. & Christensen, K. B. 2014. Analyses of model fit and robustness: a new look at the PISA scaling model underlying rankings of countries according to reading literacy. Psychometrika, 79, 210–231.

Lehtonen, R. & Pahkinen, E. 2004. Practical methods for design and analysis of complex surveys. Chichester: Wiley.

van der Linden, W. J. (toim.) 2016a. Handbook of item response theory. Volume one: models. Boca Raton, FL: CRC Press.

van der Linden, W. J. (toim.) 2016b. Handbook of item response theory. Volume two: statistical tools. Boca Raton: CRC Press.

Mislevy, R. J. 1985. Estimation of latent group effects. Journal of American Statistical Association, 80, 993–997.

OECD. 2009a. PISA Data Analysis Manual. SAS Second Edition. Paris: OECD Publishing.

OECD. 2009b. PISA Data Analysis Manual. SPSS Second Edition. Paris: OECD Publishing.
OECD. 2014. PISA 2012 technical report. http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm.
OECD. 2016. PISA 2015 results (Volume I): Excellence and equity in education. Paris: OECD Publishing.
OECD. 2017. PISA 2015 technical report. Julkaisematon käsikirjoitus.
Olsen, R. V. & Nilsen, T. 2017. Standard setting in PISA and TIMSS and how these procedures can be used nationally. Teoksessa S. Blömeke & J. E. Gustafsson (toim.) Standard setting in education: the Nordic countries in an international perspective. New York: Springer.
Rogers, A., Tang, C., Lin, M.-J. & Kandathil, M. 2006. DGROUP computer software. Princeton, NJ: Educational Testing Service.
Rubin, D. B. 1987. Multiple imputation for nonresponse in surveys. New York: Wiley.
Rutkowski, L., von Davier, M. & Rutkowski, D. (toim.) 2014. Handbook of international large-scale assessment. Background, technical issues and methods of data analysis. Boca Raton, FL: CRC Press.
Snijders, T. A. B. & Bosker, R. J. 2012. Multilevel analysis. 2nd edition. Thousand Oaks, CA: Sage.
Wu, M. 2005. The role of plausible values in large-scale surveys. Studies in Educational Evaluation, 31, 114–128.