Lasse Juhani Wallden

Kansainvälisten koulutusarvioiden vertailu koulutuksellisen tiedonlouhinnan keinoin

Tietotekniikan pro gradu -tutkielma

1. joulukuuta 2016

Jyväskylän yliopisto

Tietotekniikan laitos

Tekijä: Lasse Juhani Wallden

Yhteystiedot: lasse.wallden@gmail.com

Ohjaajat: Tommi Kärkkäinen ja Mirka Saarela

Työn nimi: Kansainvälisten koulutusarvioiden vertailu koulutuksellisen tiedonlouhinnan keinoin

Title in English: Comparing international study assessments using educational datamining methods

Työ: Pro gradu -tutkielma

Suuntautumisvaihtoehto: Koulutusteknologia

Sivumäärä: 71+14

Tiivistelmä: Koulutusta ja eri ikäisten lasten akateemista suorituskykyä mittaavat tutkimustulokset ovat kiinnostavaa tarkasteltavaa monien alojen työntekijöille ja tutkijoille. Nykyään monet organisaatiot, kuten OECD (Organisation for Economic Co-operation and Development) ja IEA (International Association for the Evaluation of Educational Achievement), järjestävät tietyin aikavälein kansainvälisiä mittauksia, joissa mitataan tietyn ikäisten lasten akateemisia kykyjä ja kysellään heidän elämästään koulussa ja kotona. Näistä mittauksista syntyvät tietokannat ovat suuria ja ne tarjoavat monipuolista tietoa koulutuksesta ja lasten oppimiseen vaikuttavista tekijöistä. Kaiken tämän lisäksi, nämä tietokannat ovat vapaassa käytössä tutkijoille, mikä puolestaan lisää tietokantojen käytettävyyttä tutkimuksen kentällä.

Mitä suuremmaksi tietokannat kasvavat, sitä suuremmalla todennäköisyydellä ne sisältävät tietoa, joka ei paljastu vain datan perinteisellä silmäilyllä tai listauksella. Tällaisten koulutuksellista tietoa sisältävien tietokantojen tutkimiseksi on viimeisten 20 vuoden aikana kehitetty monenlaisia menetelmiä ja työkaluja, joita yhdessä kutsutaan koulutukselliseksi tiedonlouhinnaksi. Koulutuksellisen tiedonlouhinnan tarkoituksena on yleensä löytää tietokannasta uutta tietoa tai tiivistää sen tulokset. Koulutuksellisen tiedonlouhinnan avulla tutkijat ovat onnistuneet löytämään koulutukseen liittyvistä tietokannoista monenlaista kiinnostavaa tietoa, jonka pohjalta ollaan pyritty mm. ennustamaan opiskelijoiden menestystä heidän ai-

kaisempien suoritustensa pohjalta ja etsimään vahvistusta maakohtaisille stereotyypeille.

Tässä tutkimuksessa sovellan K-means++ -klusterointialgoritmia vuoden 2012 PISA-aineistoon ja vuoden 2011 yhdistettyyn TIMSS ja PIRLS -aineistoon ja tarkastelen, löytyisikö niistä keskenään samanlaisia oppilasprofiileja. Pitääkseni tutkimuksen pro gradu -tutkielman rajoissa käytän tutkimuksessani vain suomalaisista oppilaista ja kouluista kerättyä dataa. Klusteroinnin tuloksena muodostuneet oppilasprofiilit olivat aineistojen välillä erilaisia, mutta jakoivat keskenään joitakin samoja piirteitä, paljastaen koko tutkimusalan näkökulmasta uutta tietämystä.

Avainsanat: tiedonlouhinta, koulutuksellinen tiedonlouhinta, PISA, TIMSS, PIRLS, klusterointi, k-means

Abstract: Research data about education and academic abilities of students of all ages has been an interesting subject for many people in different areas of work. These days there are many organizations, like OECD (Organisation for Economic Co-operation and Development) and IEA (International Association for the Evaluation of Educational Achievement), that organize international measurements of academic skills of children of certain ages on a certain time frame. The databases of these measurements are huge and they can offer versatile knowledge about the education systems and things that can affect student's ability to learn. On top of that, some of these databases are free to use for the public, increasing their usability in the field of research.

The bigger the databases get, the more likely it is for them to contain information that's not visible with just looking at the results or ranking them. During the past twenty years many methods and tools has been developed to analyze these kind of education oriented databases. All togehter these kind of methods and tools are called educational data mining. Usually the aim of educational data mining research is to either find new information from the used databases or to summarize the findings. Using educational data mining, researchers have managed to find out all kind of interesting information from the educational databases. Ranging from making predictions on students performance according their past results to figuring out if country stereotypes exists amongst PISA-data.

In this study I'm going to apply K-means++ clustering algorithm to the PISA 2012 database

and the unified TIMSS and PIRLS 2011 database in order to test if similar student profiles can be found from them. In order to keep the study in the frames of a master's thesis, I'm going to focus only on the Finnish student results. The student profiles formed with the clustering were different between datasets, but they shared some features between each other and revealed some new knowledge to the field of research.

Keywords: data mining, educational data mining, PISA, TIMSS, PIRLS, clustering, k-means

Kuviot

Taulukot

Sisältö

1 Johdanto

Koulutuksen ollessa suuressa arvossa ympäri maailmaa on luonnollista, että sitä halutaan myös tutkia ja mitata erilaisilla menetelmillä. Tämä on nähtävissä esimerkiksi siitä, kuinka kahden viime vuosikymmenen aikana erilaisten koulutusta mittaavien kyselyjen määrä on kasvanut huomattavasti. Tämänkaltaisia kansainvälisiä tutkimuksia kutsutaan englanniksi nimellä International Large-Scale Assessment (LSA) ja niistä suurimpia ovat TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study) ja PISA (Programme for International Student Assessment) (Rutkowski ym. 2010).

PISA on OECD:n alaisuudessa toimiva tutkimusohjelma, joka tutkii koulutusta mittaamalla 15-vuotiaiden oppilaiden tietoja ja taitoja. PISA-tutkimuksia järjestetään joka kolmas vuosi ja vuonna 2012 pidettyyn mittaukseen osallistui noin 510 000 oppilasta 65 eri maasta (OECD 2010). TIMSS-tutkimusohjelmassa mitataan neljäs- ja kahdeksasluokkalaisten kykyjä matematiikassa ja luonnontieteissä joka neljäs vuosi. PIRLS-tutkimusohjelma keskittyy puolestaan mittaamaan neljäsluokkalaisten lukemisen taitoja joka viides vuosi. Vuonna 2011 TIMSS- ja PIRLS-mittaukset sattuivat samalle vuodelle, jolloin TIMSS:iin osallistui noin 600 000 oppilasta 52 maasta ja PIRLS:iin noin 325 000 oppilasta (IEA 2015). Akateemisten taitojen lisäksi PISA-, TIMSS- ja PIRLS-mittauksissa kysellään myös muista oppilaan elämään ja mahdollisesti opiskeluun liittyvistä asioista, kuten kodista, perheestä ja koulutustaustoista (P. OECD 2012) ja (Michael O. Martin 2013). Edellä mainittujen tutkimusohjelmien tuottama data on vapaasti saatavilla internetistä ja niiden käyttämistä ja tulkitsemista varten on laadittu useita ohjeistuksia, raportteja ja oppaita tutkimusten järjestäjien toimesta (Foy 2013), (Kastberg ym. 2013) & (P. OECD 2014).

Kun mittauksiin osallistuvien oppilaiden määrä on satoja tuhansia, ovat mittauksista saatavat tietokannat myös erittäin suuria. Tällaisien suurien tietokantojen analysoimiseksi on kehitetty jo 1960-luvulta lähtien erilaisia metodeja ja algoritmeja, joiden avulla tietokannoista etsitään ennestään odottamattomia rakenteita ja säännönmukaisuuksia (Smyth 2000). Näiden metodien kokonaisuutta kutsutaan tiedonlouhinnaksi (engl. Data Mining) (Smyth 2000). Kun tiedonlouhinnan keinoja hyödynnetään koulutukseen liittyvien tietokantojen analysoi-

miseen, voidaan tästä käyttää nimitystä koulutuksellinen tiedonlouhinta (engl. Educational Data Mining). Vaikka koulutuksellista tiedonlouhintaa käyttävissä tutkimuksissa pyritään muun tiedonlouhinnan tavoin löytämään uutta tietoa tutkitusta datasta, eroavat joidenkin mielestä koulutuksellisessa tiedonlouhinnassa käytetyt metodit muun tiedonlouhinnan yleisimmistä metodeista, koska niissä on otettava huomioon käytetyn datan monikerroksinen hierarkkisuus ja riippuvaisuus (Baker, Corbett ja Koedinger 2004).

Koulutuksellista tiedonlouhintaa on käytetty aikaisemminkin PISA-, TIMSS- ja PIRLS-tietokantojen tarkasteluun. Esimerkiksi Mirka Saarela ja Tommi Kärkkäinen ovat etsineet vuoden 2012 PISA-aineistoista kansainvälisesti esiintyviä stereotypioita (Saarela ja Kärkkäinen 2015) ja suomalaisten lasten tuloksissa esiintyviä sukupuoliin liittyviä säännönmukaisuuksia (Saarela ja Kärkkäinen 2014). Liu ja Ruiz (2008) ovat puolestaan louhineet vuosien 1995, 1999 ja 2003 TIMSS-aineistoja, sekä NAEP-aineistoja ennustaakseen K-12 oppilaiden menestystä energiaa koskevissa koekysymyksissä.

Tutkimuksessani pyrin selvittämään, onko vuoden 2011 yhdistetyssä TIMSS ja PIRLS -aineistosta, sekä vuoden 2012 PISA-aineistosta löydettävissä samanlaisia oppilasprofiileja, kun niihin sovelletaan samaa koulutukselliseen tiedonlouhintaan kuuluvaa klusterointialgoritmia. Jaottelin aineistot niin, että vertailin samaa aihealuetta koskevaa dataa kummastakin aineistosta. Tämän lisäksi keskityin vain suomalaisia oppilaita koskevaan dataan pitääkseni tutkimuksen sopivan laajana pro gradu -tutkielmaani ajatellen. Vastaavanlaisia tutkimuksia, joissa useita eri LSA-aineistoja verrattaisiin näin koulutuksellisen tiedonlouhinnan avulla, on raportoitu vähän. Ainakin Skryabin ja muut (2015) ovat tutkineet valtion tietoteknisen kehityksen tason ja lasten tietotekniikan käytön vaikutuksia heidän matematiikan, luonnontieteiden ja lukemisen taitoihinsa. Tutkimuksessa todettiin valtiollisen tietoteknisen kehityksen tason ennustavan oppilaiden menestystä kaikissa kolmessa aineessa sekä neljännellä että kahdeksannella luokalla. Henkilökohtaisen käytön vaikutukset kuitenkin vaihtelivat mm. käyttötarkoituksen ja oppilasryhmän mukaan.

Työn kolmessa seuraavassa kappaleessa esittelen tutkimuksessani käytettyjen tietokantojen ja tutkimusmetodien kannalta tärkeää taustatietoa. Kappaleessa kaksi esittelen tietokannat luoneet tutkimusohjelmat, eli PISA:n, TIMSS:in ja PIRLS:in. Tarkastelen tekstissäni näitä tutkimusohjelmia niiden keräämän tiedon, historian ja toimintatapojen näkökulmista. Lisäk-

si esittelen lyhyesti kaikkien käsittelemieni tutkimusten viralliset tutkimustulokset. Kappaleessa kolme puolestaan esittelen tutkimuksessani käytetyn tiedonhankintamenetelmän, koulutuksellisen tiedonlouhinnan, taustoja ja tieteenalan nykytilaa. Kappaleessa neljä esittelen klusteroinnin perusperiaatteen, erilaisia klusterointimenetelmiä ja tutkimuksessani käyttämäni K-means++ -algoritmin.

Kappaleessa viisi aloitan varsinaisen tutkimukseni esittelyn. Aloitan selittämällä miten valitsin otannan käyttämistäni tietokannoista, kuinka käsittelin dataa ennen K-means++ -algoritmin soveltamista ja kuinka määrittelin etsittyjen klustereiden määrän. Tämän jälkeen kappaleessa kuusi käyn läpi tiedonlouhinnalla löytämäni tulokset. Aloitan esittelemällä klusteroinnin avulla löytämäni klusterit ja niitä karakterisoivat tekijät. Tämän jälkeen peilaan klusteroinnin pohjalta muodostuneita klusteriprototyyppejä niitä vastaaviin selittäviin muuttujiin ja pohdin mitä nämä tulokset yhdessä tarkoittavat. Kappaleen lopussa vertailen edellä mainittuja tuloksia keskenään PISA-aineiston ja yhdistetyn TIMSS ja PIRLS -aineiston välillä ja vastaan varsinaiseen tutkimuskysymykseeni samanlaisten klustereiden mahdollisesta löydettävyydestä.

Työn viimeisessä kappaleessa käyn tiivistelmän omaisesti läpi vielä kertaalleen tekemäni tutkimuksen ja sen tulokset, jonka lisäksi pohdin myös tulosten mahdollisia syitä. Lopuksi pohdin myös, millaista tutkimusta tulevaisuudessa tulisi tehdä, jotta tutkimukseni tai samanlaisten tutkimusten tuloksia voitaisiin ymmärtää paremmin.

2 PISA, TIMSS ja PIRLS

1990-luvun aikana kansainvälisten koulutusta ja lasten oppimistuloksia tutkivien tutkimusten määrä alkoi kasvaa merkittävästi (Rutkowski ym. 2010). PISA, TIMSS ja PIRLS ovat näistä tutkimuksista suurimpia ja niihin osallistuu satoja tuhansia oppilaita kymmenistä eri maista. Oppilaiden akateemisen osaamisen lisäksi näissä tutkimuksissa kartoitetaan myös oppilaiden taustoja sekä heidän koulujensa tietoja (P. OECD 2012) ja (Michael O. Martin 2013). Tutkimusten koosta ja datan korkeasta laadusta johtuen päätin käyttää juuri PISA:n, TIMSS:in ja PIRLS:in tarjoamaa dataa omassa tutkimuksessani. Tässä kappaleessa kuvailen lyhyesti kaikkien kolmen tutkimusohjelman toimintaa ja esittelen tarkemmin juuri niiden vuosien tutkimuksia ja tutkimustuloksia, joiden dataa käytin omassa tutkimuksessani.

2.1 PISA: Programme for International Student Assessment

PISA (Programme for International Student Assessment) on OECD:n (Organisation for Economic Co-operation and Development) alaisuudessa toimiva tutkimusohjelma, joka on mitannut 15-vuotiaiden lasten kykyjä höydyntää koulussa oppimiaan matematiikan, luonnontieteiden ja lukemisen taitoja joka kolmas vuosi aina vuodesta 2000 alkaen (P. OECD 2012) & (OECD 2010). Akateemisen osaamisen lisäksi PISA-tutkimuksissa ollaan kiinnostuneita myös lasten taustoista, eli kotioloista ja koulusta jossa hän opiskelee. PISA:n historian aikana tutkimuksiin on osallistuttu yli 70 eri maasta, jotka ovat joutuneet erikseen ilmoittautumaan jokaiseen tutkimukseen aina kaksi vuotta ennen sen pitämistä. Ilmoittautumisen lisäksi osallistuvien maiden on itse pystyttävä järjestämään tutkimukseen kuuluvat koetilaisuudet ja kyselyt, sekä kattamaan niistä koituvat kustannukset (OECD 2010).

PISA-tutkimuksessa käytettyjen kysymysten ja kyselylomakkeiden laatiminen on monivaiheinen ja työläs prosessi, jossa on otettava huomioon monia mahdollisesti tuloksiin vaikuttavia tekijöitä. Kaikki osallistuvat maat saavat lähettää ehdotuksia käytettävistä kysymyksistä (OECD 2010), jotka kuitenkin analysoidaan ja testataan tarkkaan ennen varsinaisen kokeen pitämistä erinäisten datan luotettavuutta heikentävien tekijöiden poistamiseksi (P. OECD 2012). Esimerkiksi yksi suurimmista huolenaiheista on ongelma kysymysten kult-

tuurillisesta yhteensopivuudesta. Varsinkin Likert-asteikkoa käyttävissä kyselyissä ja kysymyksissä on todettu huomattavaa kulttuurien välisten vastaustapojen erilaisuuden aiheuttamaa vääristymää (P. OECD 2012). Varsinaisessa testissä käytetyt kyselylomakkeet sisältävät sekä monivalinta- että avokysymyksiä.

Myös otannan valinnassa käytetään monivaiheista prosessia. Vuoden 2012 tutkimuksessa Venäjällä käytettiin jopa kolmiportaista menetelmää, siinä missä muissa osallistujamaissa otannan valinta suoritettiin kaksiportaisella menetelmällä. Yleisemmin käytetyn mallin ensimmäisessä vaiheessa kartoitettiin kaikki osallistujamaiden koulut, jotka pystyivät osallistumaan tutkimukseen. Tämän jälkeen koulut jaoteltiin niiden ominaisuuksien mukaan ryhmiin ja tämän ryhmittelyn avulla valittiin tutkimukseen osallistuvat koulut. Jokaisesta osallistujamaasta pyrittiin valitsemaan vähintään 150 koulua, mutta mikäli jossain maassa ei ollut niin montaa koulua, otettiin kyseisen maan kaikki koulut mukaan tutkimukseen. Osallistuvien koulujen valinnan jälkeen listattiin kaikki oppilaat, jotka pystyivät ja suostuivat osallistumaan tutkimukseen. PISA:n työntekijät valitsivat lopullisen otoksen kaikista listatuista oppilaista KeyQuest-nimisen ohjelmiston avulla. (P. OECD 2012)

Varsinainen testi tapahtui vielä vuonna 2012 osittain kynä ja paperi -kokeita hyödyntäen ja osittain tietokoneilla. Kaikki oppilaat eivät vastanneet jokaiseen tutkimuksessa mukana olleeseen kysymykseen, vaan kysymykset olivat jaoteltu usealle lomakkeelle. Tutkimukseen osallistuva oppilas sai täytettäväkseen tietyt lomakkeet, joihin hän vastasi kaksi tuntia kestävän testin aikana. Oppilaiden vastausten tarjoaman raakadatan lisäksi PISA-tutkimuksissa raportoidaan myös tästä raakadatasta johdettuja muuttujia. Nämä johdetut muuttujat muodostetaan antamalla kullekin kysymykselle ja vastaajalle tietty painoarvo joka sitten huomioidaan muuttujien muodostamiseen käytetyissä laskukaavoissa. (P. OECD 2012)

Vuoden 2012 PISA-tutkimukseen osallistui 65 maata ja noin 510 000 oppilasta. Tutkimuksen pääpaino oli tänä vuonna matematiikassa, mutta lukeminen ja luonnontieteet olivat myös mukana pienemmällä painotuksella. Kaksituntisen kynä ja paperi -kokeen lisäksi oppilaat vastasivat myös puolen tunnin mittaiseen taustakyselyyn, jossa kartoitettiin heidän olojaan kotona ja koulussa. Osallistuvat maat saivat myös päättää, mikäli oppilaat osallistuivat 40 minuuttiseen, tietokoneilla suoritettavaan tieto- ja viestintätekniikan käyttöön liittyvään kyselyyn tai kyselyyn oppilaan tähänastisen opetuksen sujumisesta ja tulevaisuuden suunnitel-

mista (PISA 2012). Suomesta tutkimukseen osallistui 8829 oppilasta 311 eri koulusta. (P. OECD 2012)

2.1.1 Vuoden 2012 PISA-tutkimuksen virallisista tuloksista

Vuosien 2012 ja 2013 aikana OECD julkaisi vuoden 2012 PISA-tutkimuksensa tulokset mm. viisiosaisessa julkaisujen sarjassa ja erillisessä tiivistelmässä. Seuraavaksi käyn lyhyesti läpi näiden virallisten tulosten pääkohdat ensin kansainvälisellä ja sitten Suomen tasolla. OECD:n julkaisuista keskityn seuraavassa kuvauksessa tiivistelmään ja laajempien raporttien osiin I, III ja IV.

Raporttisarjan osassa I keskitytään oppilaiden akateemisen osaamisen tarkasteluun, osassa III selvitetään oppilaiden taustojen, motivaation ja asenteiden vaikutusta heidän oppimiseensa ja osassa IV keskitytään tarkkailemaan toimivien koulusysteemien rakennetta (OECD 2014c), (Economic Co-operation ja Development 2014), (Economic Co-operation ja Development 2013) & (Economic Co-operation ja (OECD) 2013).

Kansainvälisesti tutkimuksen tulokset voidaan tiivistää seuraavasti:

- Matematiikassa viisi parhaiten pärjännyttä maata olivat Singapore, Hongkong, Kiinan Taipei, Korea ja Macaon Kiina.
- Lukemisessa viisi parhaiten pärjännyttä maata olivat Shanghai, Hongkong, Singapore, Japani ja Korea.
- Luonnontieteissä viisi parhaiten pärjännyttä maata olivat Shanghai, Hongkong, Singapore, Japani ja Suomi.
- Oppilaan sosioekonomisella asemalla ja esiopetukseen osallistumisella on vaikutusta tämän akateemisiin taitoihin.
- Matematiikkaan liittyvä ahdistus on suhteellisen yleistä, jopa 30% oppilaista kertoo kärsivänsä siitä.
- Valtiot, joissa resursseja jaetaan tasaisesti koulujen välille, pärjäävät muita paremmin. Resurssien tasainen jakautuminen nähdään yhtä tärkeänä kuin resurssien määrä.
- Koulut, joilla on enemmän autonomiaa opetuksen järjestämiseen ja kehitykseen, pärjäävät muita paremmin.

- Vanhempien odotuksilla lastensa akateemista menestystä ja tulevaisuutta kohtaan on vaikutusta lapsen motivaatioon ja itsevarmuuteen, joilla puolestaan on taas vahva yhteys hänen akateemiseen menestykseensä.
- Oppilaiden toistuvilla poissaoloilla on yhteys heikentyneeseen akateemiseen kyvykkyyteen.
- Kouluviihtyvyydellä ja kouluun kuuluvuuden tunteella on yhteys akateemisiin tuloksiin.

Suomen suoriutuminen vaihteli vuoden 2012 PISA-tutkimuksessa huomattavasti aineittain. Luonnontieteissä Suomi ylsi viiden parhaan maan joukkoon, kun taas lukemisessa sijoitus oli kuudes ja matematiikassa vasta 12. Verrattuna vuoden 2003 tutkimukseen, suomalaisten opiskelijoiden suoriutuminen oli heikentynyt kaikissa mitatuissa oppiaineissa. Tästä heikentymisestä huolimatta tulokset olivat vielä huomattavasti yli OECD-maiden keskiarvon. Kokonaisuudessaan tytöt näyttivät pärjäävän paremmin kaikissa mitatuissa oppiaineissa.

Suhteellisen hyvästä suoriutumisestaan huolimatta suomalaiset oppilaat kertoivat, että heidän kouluviihtyvyytensä, kouluun kuuluvuuden tunteensa ja itseluottamus omaan matemaattiseen kyvykkyyteen olivat alle OECD-maiden keskiarvon. Oppilaat käyttivät kotona vähemmän aikaa opiskeluun kuin oppilaat OECD-maissa keskimäärin. Kuitenkin sinnikkyyden vaikutus oppimistuloksiin oli Suomessa OECD-maiden vahvin, joka luultavasti selittää myös Suomen sijoittumista tuloksissa.

Suomalainen koulusysteemi otetaan raporteissa puolestaan esiin vähäisen varianssin takia. Suomalainen koulusysteemi on siis onnistunut tuomaan oppilaiden osaamisen ääripäät lähemmäs toisiaan, joka puolestaan on todennäköisesti vaikuttanut positiivisesti Suomen asemaan keskimääräistä osaamista mittaavissa testituloksissa. Suomalaiset opettajat ovat myös OECD-maiden keskiarvoa koulutetumpia ja heillä on keskiarvoa useammin pätevyys työhönsä.

2.2 TIMSS: Trends in International Mathematics and Science Study

TIMSS (Trends in International Mathematics and Science Study) on IEA:n (International Association for the Evaluation of Educational Achievement) alaisuudessa toimiva tutkimusohjelma, joka mittaa neljäs- ja kahdeksasluokkalaisten osaamista matematiikassa ja luonnontieteissä joka neljäs vuosi. Akateemisen osaamisen lisäksi TIMSS-tutkimuksissa ollaan kiinnostuttu myös oppilaan oppimiseen vaikuttavista tekijöistä tämän oppimisympäristössä. Ensimmäinen TIMSS-tutkimus järjestettiin vuonna 1995 ja tähän mennessä tutkimuksiin on osallistuttu yli 60 eri maasta. Osallistuvat maat saavat itse päättää, osallistutaanko maista molempia ikäluokkia tarkasteleviin tutkimuksiin vai pelkästään toiseen niistä. (Mullis ym. 2009)

TIMSS:issä käytettävät tehtävät ja kysymykset laaditaan yhteistyössä IEA:n ja osallistuvien maiden kanssa. PISA:n tavoin kaikki TIMSS:iin osallistuvat oppilaat eivät täytä samoja lomakkeita, vaan kaikista käytettävissä olevista kysymyksistä valmistetaan useita erilaisia yhdistelmiä, joista oppilaat saavat sitten testissä yhden täytettäväkseen. Pidetyn tutkimuksen jälkeen osa käytetyistä kysymyksistä annettaan julkiseen tarkasteluun ja nämä kysymykset on korvattava aina seuraavaa testiä varten. Osallistuvien maiden opetusjärjestelmien edustajat voivat lähettää kysymysehdotuksiaan IEA:lle, jonka alaisuudessa toimivat tarkastajat sitten tarkastavat ja mahdollisesti hyväksyvät ne. Kysymykset myös testataan osallistuvissa maissa ennen varsinaista tutkimusta. Testin aikana pystytään määrittämään kysymysten vaikeustaso, kyky erotella taitavia oppilaita heikommista ja muita kysymysten käytön kannalta tärkeitä asioita. (Kastberg ym. 2013).

Vuoden 2011 TIMSS:in otannan valinnassa hyödynnettiin kaksivaiheista prosessia, jossa tutkimukseen osallistuvat maat muodostivat koululaisistaan tutkimukseen sopivan otoksen yhdessä TIMSS:in kanssa työskentelevien asiantuntijoiden kanssa. Prosessin ensimmäisessä vaiheessa valittiin tutkimukseen osallistuvat koulut ja toisessa vaiheessa valituista kouluista valittiin kokonaisia luokkia tutkittavaksi. TIMSS:in tarkkuusvaatimusten täyttämiseksi suurimmassa osassa osallistujamaista riitti noin 150 koulua ja noin 4000 oppilasta kattava otos (Joncas ja Foy 2012). Otannan muodostamisen molempiin vaiheisiin kuuluu luonnollisesti tämän lisäksi moninaisia sääntöjä ja pienempiä vaiheita, mutta en käsittele niitä tämän työn puitteissa tämän tarkemmin.

Myös TIMSS-tutkimuksissa raakadatalle annetaan myöhempää käsittelyä ja laajempien muuttujien muodostamista varten erinäisiä painoarvoja (Rutkowski ym. 2010). Painoarvoja annetaan akateemista kyvykkyyttä mittaavien kysymysten lisäksi myös oppilaan taustoja ja kouluun liittyviä ominaisuuksia mittaaville kysymyksille.

Vuoden 2011 TIMSS-tutkimukseen osallistui yli 600 000 oppilasta 63 eri maasta. Suomesta tutkimukseen osallistui 4638 oppilasta 145 eri koulusta (Martin ja Mullis 2012). Neljäsluokkalaisille suunniteltu testi oli pituudeltaan noin 72 minuuttia pitkä ja kahdeksasluokkalaisille suunniteltu testi taas noin 90 minuuttia (Kastberg ym. 2013).

2.2.1 Vuoden 2011 TIMSS-tutkimuksen virallisista tuloksista

Vuonna 2012 IEA julkaisi viralliset raporttinsa vuoden 2011 TIMSS-tutkimuksen tuloksista. Matematiikan ja luonnontieteiden tulokset jaettiin omiin julkaisuihinsa, mutta alla olevassa katsauksessani käsittelen näitä kahta julkaisua yhdessä (Mullis, Martin, Foy ja Arora 2012) & (Martin ym. 2012). Raporteissa tutkimustuloksista esitettiin tiivistelmiä ja johtopäätöksiä kansainvälisellä tasolla, mutta myös yksittäisten valtioiden tuloksista mainittiin, mikäli ne olivat jollain tavalla poikkeavia. Käsittelen TIMSS-tutkimuksen tuloksia ensin kansainvälisellä ja sitten Suomen tasolla.

Kansainvälisellä tasolla vuoden 2011 TIMSS-tutkimuksen tulokset voidaan tiivistää seuraavasti:

- Matematiikassa tutkimuksen viisi parhaiten pärjännyttä maata olivat Singapore, Korea, Hongkong, Kiinan Taipei ja Japani. Nämä maat olivat viiden parhaan joukossa neljäs- ja kahdeksasluokkalaisten tuloksissa.
- Luonnontieteissä tutkimuksen viisi parhaiten pärjännyttä maata olivat neljännellä luokalla Korea, Singapore, Suomi, Japani ja Venäjä, kun taas kahdeksannella luokalla: Singapore, Kiinan Taipei, Korea, Japani ja Suomi.
- Oppilaiden matemaattinen ja luonnontieteellinen tietämys oli huomattavasti vahvempaa kuin soveltaminen ja päättely.
- Matemaattisten taitojen aikainen harjoittelu vaikuttaa olevan vahvasti yhteydessä taitojen kehitykseen ja myöhempään osaamiseen. Esiopetuksella on tärkeä tehtävä auttaa

lasta pääsemään irti luontaisista matemaattisista heikkouksistaan.

- Kodin oppimista tukevilla resursseilla on vahva yhteys lapsen matemaattiseen ja luonnontieteelliseen menestykseen.
- Oppilaat, joiden kouluilla on paljon resursseja, turvallinen ilmapiiri ja joissa tuetaan enemmän akateemista menestystä, pärjäävät muita paremmin sekä matematiikassa että luonnontieteissä.
- Opettajien valmistautuminen opetukseen ja tyytyväisyys omaan työuraansa ovat yhteydessä heidän oppilaidensa matemaattiseen ja luonnontieteelliseen suoriutumiseen.
- Oppilaat, jotka suhtautuvat positiivisesti matematiikkaan ja luonnontieteisiin, pärjäävät muita paremmin, mutta asenteet muuttuvat kahdeksanteen luokkaan mennessä. Suhde asenteiden ja suoriutumisen välillä on kaksisuuntainen.
- Ravinnon ja unen puutteen vaikutukset häiritsevät oppilaiden akateemista suoriutumista monissa maissa. Kaikista tutkimukseen osallistuneista neljäsluokkalaisista oppilaista noin 29% kärsii unen puutteen tuottamista ongelmista, kun taas ravinnon puutteen ongelmista kärsii noin 47%.

Suomen sijoittuminen matematiikan osaamisessa oli sekä neljäs- että kahdeksasluokkalaisilla sijalla kahdeksan. Luonnontieteiden puolella Suomi puolestaan sijoittui paremmin ja ylsi kummassakin ikäluokassa viiden parhaan maan joukkoon. Vuoden 1999 TIMSS-tutkimuksen tuloksiin verrattuna Suomen tulokset olivat heikentyneet kaikissa mitatuissa oppiaineissa.

Tutkimuksen tarkemmat tulokset antavat ymmärrystä suomalaisten lasten korkeisiin tuloksiin, mutta nostavat esiin myös joitakin kiinnostavia ristiriitoja. Korkeita tuloksia tukee mm. suomalaisten kotien suhteellisen korkea oppimista tukevien resurssien määrä, opettajien kansainvälistä keskiarvoa korkeampi koulutus, lasten oikeus esikouluun ja suomalaisten koulujen kansainvälistä keskiarvoa rauhallisempi työympäristö. Nämä kaikki tekijät ovat myös kansainvälisellä tasolla yhteydessä oppilaiden korkeampaan akateemiseen suoriutumiseen.

Ristiriitaisia huomioita datassa olivat mm. suomalaisten oppilaiden alhainen mielenkiinto matematiikkaan ja luonnontieteisiin, joka vähenee neljänneltä luokalta kahdeksannelle siirryttäessä, suomalaisten vähäinen ajankäyttö kotitehtävien tekemiseen ja kotona opiskeluun, suomalaisten opettajien vähäinen itsensä kehittäminen, tyytymättömyys työuraansa ja kansainvälistä keskiarvoa alhaisempi itseluottamus, koulujen resurssipula luonnontieteiden osal-

ta ja oppilaiden kansainvälistä keskiarvoa alhaisempi sitoutuminen luokkaan (engl. engagement in class). Näiden kaikkien seikkojen pitäisi heikentää oppilaiden akateemista kyvykkyyttä, mutta näistä huomattavistakin ongelmista riippumatta, Suomi onnistui sijoittumaan kymmenen parhaan joukkoon kaikissa mitatuissa oppiaineissa.

Pääpiirteittäin suomalainen tutkimusdata mukaili edellä mainittua listaa kansainvälisestä tiivistelmästä. Myös ravinnon ja unen puutteen vaikutuksia havaittiin suomalaisilla oppilailla. Ravintoon liittyvistä ongelmista kärsi noin 10% neljäsluokkalaisista ja 16% kahdeksasluokkalaisista, kun taas unen puutteen ongelmista kärsi noin 60% neljäsluokkalaisista ja noin 80% kahdeksasluokkalaisista.

Muita mielestäni huomionarvoisia havaintoja olivat:

- Neljännellä luokalla pojat suoriutuivat tyttöjä paremmin sekä matematiikassa että luonnontieteissä, mutta kahdeksanteen luokkaan mennessä tytöt olivat ohittaneet pojat kummassakin oppiainekokonaisuudessa.
- Mitä korkeammat odotukset vanhemmilla ja oppilaalla itsellään oli tämän akateemisesta tulevaisuudesta, sitä korkeammat tulokset hänellä oli.
- Oppilaiden aiheuttaman häiriön määrä lisääntyy neljänneltä luokalta kahdeksannelle siirryttäessä.

2.3 PIRLS: Progress in International Reading Literacy Study

TIMSS:in tavoin myös PIRLS (Progress in International Reading Literacy Study) toimii IEA:n alaisuudessa. Kuitenkin TIMSS:istä poiketen, PIRLS:issä tutkitaan neljäsluokkalaisten lukemisen taitoja joka viides vuosi. PIRLS-tutkimuksia on järjestetty vuodesta 2001 lähtien. Samoin kuin PISA:ssa, myös PIRLS:issä kysellään oppilaiden oloista kotona ja koulussa.

PIRLS:issä käytetään otannan valinnassa, testikysymysten laadinnassa ja datan muodostamisessa samoja menetelmiä kuin TIMSS:issä (Kastberg ym. 2013), jotka esittelin kappaleessa 2.2. Vuoden 2011 PIRLS:iin osallistui yli 325 000 oppilasta 48 eri maasta. Suomesta tutkimukseen osallistui 4640 oppilasta 145 koulusta (Martin ja Mullis 2012). Oppilaille teetetty

testi oli vastauspaperista riippumatta noin 80 minuuttia pitkä (Kastberg ym. 2013).

2.3.1 Vuoden 2011 PIRLS-tutkimuksen virallisista tuloksista

Myös vuoden 2011 PIRLS-tutkimuksen tuloksista laadittiin IEA:n toimesta raportti vuonna 2012 (Mullis, Martin, Foy ja Drucker 2012). Kansainvälisen tiivistelmän puolesta raportti vastasi pitkälti TIMSS:in tiivistelmää, varsinkin hyvän koulun määritelmään, aikaiseen opetukseen ja oppilaiden asenteisiin liittyvissä seikoissa. Nämä sivuuttaen voidaan kansainvälinen tiivistelmä kuvailla seuraavasti:

- Tulosten pohjalta neljä parhaiten pärjännyttä maata olivat Hongkong, Venäjä, Suomi ja Singapore.
- Vanhempien lukuinnokkuudella näyttää olevan yhteys oppilaiden lukemisen taitoihin.

Myös suomalaisia koskevat tulokset olivat PIRLS-tutkimuksessa hyvin samanlaisia TIMSS-tutkimuksen tuloksiin verrattuna, kun kyse oli oppilaan taustoista ja motivaatiosta, opettajan koulutuksesta ja tyytyväisyydestä, koulujen resursseista ja työympäristöstä. Tutkimuksen tuloksista oli kuitenkin huomattavissa seuraavia seikkoja suomalaisten oppilaiden äidinkielen opiskeluun liittyen:

- Tytöt pärjäsivät poikia paremmin lukemisessa.
- Suomalaiset vanhemmat lukevat kansainvälistä keskiarvoa enemmän.
- Suomalaiset oppilaat olivat vahvempia suorassa tiedonhaussa kuin tekstin ymmärtämisessä tai arvioimisessa.
- Suomalaiset oppilaat luottavat omiin lukemisen taitoihinsa enemmän kuin oppilaat maailmalla keskimäärin.

2.4 TIMSS ja PIRLS 2011, yhdistetty aineisto

Vuonna 2011 TIMSS- ja PIRLS-tutkimusten järjestäminen osui ensimmäistä kertaa samalle vuodelle. Tämän mahdollisuuden IEA käytti hyväkseen ja kokosi kaikista neljäsluokkalaisista, jotka osallistuivat kumpaankin tutkimukseen, yhden suuren aineiston (Michael O. Martin 2013).

Yhteensä 34 maata osallistui IEA:n tarjoamaan tilaisuuteen mitata neljäsluokkalaisten kykyjä sekä TIMSS:in, että PIRLS:in puitteissa (Michael O. Martin 2013). Kaikenkaikkiaan tämä aineisto sisältää yli 180 000 oppilaan tiedot (Foy 2013). Suomen osalta tähän aineistoon päätyi 4541 oppilasta 145 koulusta.

Tämä aineisto muistuttaa sisällöltään joissain määrin vuoden 2012 PISA-aineistoa, jonka takia päätin käyttää tätä aineistoa omassa tutkimuksessani.

2.4.1 Yhdistetyn TIMSS ja PIRLS -aineiston virallisista tuloksista

Vuonna 2013 IEA julkaisi raportin, jossa käsiteltiin yhdistetyn TIMSS ja PIRLS -aineiston pohjalta saatuja tuloksia ja johtopäätöksiä. Raportissa analysoitiin mm. matematiikan ja äidinkielen osaamisen yhteyttä toisiinsa, sekä koulun ja kodin vaikutusta kaikkien mitattujen oppiaineiden osaamiseen. Kuitenkin jo raportissa itsessään myönnettiin, että tällaisten kokonaisuuksien vertailu eri tavalla rakentuvien kielien välillä on hankalaa, joka puolestaan on saattanut vaikuttaa raportissa mainittuihin tuloksiin (Michael O. Martin 2013).

Kansainvälisesti tutkimustuloksista pystytään tekemään seuraavia havaintoja:

- Monilahjakkaita oppilaita on maailmanlaajuisesti suhteellisen vähän. Vain Singaporessa yli puolet oppilaista ylsi kaikissa kolmessa aineessa IEA:n asettamiin korkean luokan tuloksiin.
- Mitä korkeammat lukemisen taidot oppilaalla oli, sitä korkeammat hänen matemaattiset taitonsa olivat. Kuitenkin suurimmalla osalla oppilaista oli enemmän vaikeuksia matematiikan ja luonnontieteiden tehtävissä, jotka vaativat keskitason lukemisen taitoa, kuin korkeaa lukemisen taitoa vaativissa tehtävissä. Heikoilla lukijoilla suoritustaso laski tehtävän lukuvaatimusten kasvaessa.
- Oppilaan kehityksen kannalta on parasta, mikäli koulu, vanhemmat ja oppilas itse tavoittelevat oppilaan korkeaa akateemista menestystä.

Suomalaisista oppilaista noin puolet ylsivät yhdessä tai kahdessa oppiaineessa IEA:n määrittämään korkeaan taitoluokkaan, mutta kaikista heikoiten Suomi pärjäsi matematiikassa. Osa muistakin tutkimustuloksista erosi Suomen kohdalla jonkin verran kansainvälisestä kes-

kiarvosta. Verrattaessa oppilaiden lukemisen taitoja ja suoriutumista matematiikan tehtävistä suomalaisten oppilaiden suoriutuminen parani sitä mukaan mitä vaativampaa lukemisen taitoa tehtävissä vaadittiin. Luonnontieteiden tehtävissä suomalaisten tulokset vastasivat kansainvälistä keskiarvoa ja molemmissa aineissa paremmin lukevat pärjäsivät heikompia lukijoita paremmin.

Tuloksista oli myös havaittavissa suomalaisten vanhempien vaikutus lastensa opiskeluun. Suomessa oppilaiden vanhemmat olivat hyvin koulutettuja suhteessa kansainväliseen keskiarvoon. Tämän lisäksi lasten vanhempien koulutuksella oli yhteys heidän akateemisiin kykyihinsä.

Voimakkain yhteys koulun toiminnan ja oppilaiden testitulosten välillä oli koulun turvallisuudella ja työrauhalla, koulun panostamisella akateemiseen menestykseen ja oppilaiden mielenkiinnon pitämisellä oppitunneilla.

2.5 Tutkimusohjelmien yhtäläisyydet ja erot

Kuten Olsen (2005) ja Rutkowski ja muut (2010) toteavat töissään, LSA-tutkimuksissa on hyvin paljon toisiaan muistuttavia piirteitä. Taulukkoon 1 olen koonnut suurimmat yhtäläisyydet ja erot OECD:n PISA:n ja IEA:n TIMSS:in ja PIRLS:in välillä.

Samaa	Erilaista
- Kansainvälisiä tutkimuksia	- Tutkittavien oppilaiden ikä
- Tuhansia osallistujia kymmenistä maista	- Kysymykset eivät ole samoja
- Otannan ja kysymysten valinta monimutkainen prosessi	- Raakadatasta johdetut muuttujat eroavat toisistaan
- Data ja dokumentaatio saatavilla ilmaiseksi verkosta	- PISA-tutkimuksissa vaihdellaan tarkemmassa tarkastelussa olevaa oppiainetta
- Tutkitaan muutakin kuin akateemista kyvykkyyttä	- IEA:n tutkimuksissa otannat ovat luokkatasolta ja PISA:n otannat ovat koulutasolta
- Raakadatalle tehdään moninaista jälkikäsittelyä ja niille annetaan eri painoarvot	
- Mittausten suunnittelu ja toteuttaminen tapahtuu osallistuvien valtioiden ministeriöiden ja tutkimuslaitosten ammattilaisten yhteistyössä	

Taulukko 1: Yhtäläisyydet ja erot PISA-, TIMSS- ja PIRLS-tutkimusten välillä.

3 Koulutuksellinen tiedonlouhinta

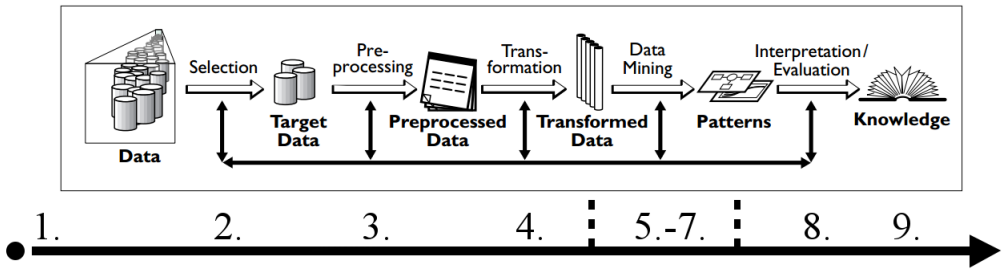
Tiedonlouhinnaksi kutsutaan tieteen kenttää, jossa suurista datajoukoista pyritään löytämään uutta ja hyödyllistä tietämystä. Tiedonlouhintaa aletaan kutsumaan koulutukselliseksi tiedonlouhinnaksi, kun louhittava data käsittelee koulutuksellisia aihepiirejä (Baker 2010).

Tässä kappaleessa esittelen tiivistetysti tiedonlouhinnan ja koulutuksellisen tiedonlouhinnan periaatteen, historiaa ja metodeja.

3.1 Tiedonlouhinnan toimintaperiaate

Tiedonlouhintaa (engl. data mining) ja sitä muistuttavaa toimintaa kutsutaan eri yhteisöissä hieman eri nimillä. Esimerkiksi Fayyad ja muut (1996) kutsuvat artikkelissaan tällaista toimintaa englanniksi termillä knowledge discovery in databases (KDD) ja sanovat data miningin (DM) olevan vain osa KDD-prosessia. Kuitenkin vertaamalla Fayyadin ja muiden kuvausta KDD:n vaiheista esimerkiksi Liun ja muiden (2008) kuvaukseen DM:än vaiheista, voidaan huomata selitysten välillä hyvin paljon samanlaisuutta.

Fayyad ja muut (1996) tiivistävät artikkelissaan tietämyksen etsimisen datasta yhdeksän askeleen prosessiksi. Kuviossa 1 olen sovittanut nämä yhdeksän askelta Fayyadin kuviossa kuvaamiin KDD:n päävaiheisiin.



Kuvio 1: Fayyadin käyttämä kaavio KDD prosessin päävaiheista (Fayyad, Piatetsky-Shapiro ja Smyth 1996). Olen muokannut kuvaa lisäämällä sen alle aikajanan, jonka tarkoitus on helpottaa KDD prosessin yhdeksän askeleen sijoittamista kuvan vaiheisiin.

1. Tutkimusalaan tutustuminen. Työkalujen käyttö ei ole välttämättä kovinkaan tehokasta, ellei käyttäjä tiedä mitä niillä on tarkoitus tehdä. Tiedonlouhintaan ryhtyvän on siis hyvä hieman tutustua aikaisemmin suoritettuihin tutkimuksiin ja teksteihin oppiakseen, miten tiedonlouhintaa on tarkoitus käyttää ja mihin sillä yleisesti ottaen pyritään.

2. Tutkittavan datan valitseminen. Tilanteesta riippuen tiedonlouhija voi koota datansa itse, tai käyttää jo valmiiksi olemassa olevaa dataa.

3. Datan puhdistaminen ja esivalmistelut. Kaikki arkistoitu data ei ole aina täydellistä, vaan raakadataan sisältyy usein myös tiedonlouhintaa häiritseviä muuttujia, puuttuvia arvoja tai muita tekijöitä, jotka voivat vaikuttaa tiedonlouhinnan tuloksiin. Tutkijan on tässä kohtaa valittava menetelmänsä näiden häiriötekijöiden poistamiseksi.

4. Datan vähentäminen ja projektio. Valittua dataa voidaan karsia vielä poistamalla tutkimuksen kannalta tarpeettomat muuttujat tai yhdistämällä useammasta muuttujasta uusia, laajempia muuttujia.

5. Tiedonlouhintamenetelmän valinta. Tutkijan on päätettävä datan louhinnan tarkoitus. Pyrkiikö tutkija esimerkiksi tiivistämään tietoa, luokittelemaan sitä vai klusteroimaan?

6. Tiedonlouhinta-algoritmin valinta. Tässä vaiheessa tutkija valitsee tutkimuksessaan käytettävän tiedonlouhinnan menetelmän toteuttavan algoritmin. Esimerkiksi klusterointia varten on olemassa useita erilaisia algoritmeja (katso kappale 3.2). Jotkin algoritmeista vaativat jo ennen käyttöään niiden toteutukseen liittyvien parametrien määrittelyä, joka myös tulisi toteuttaa tässä vaiheessa.

7. Tiedonlouhinnan toteuttaminen. Tutkija käyttää edellisessä vaiheessa valitsemaansa tiedonlouhinnan metodia ja algoritmia aikaisemmin valitsemaansa ja siistimäänsä dataan. Louhimisen tuloksena voi käytetystä metodista riippuen syntyä esimerkiksi luokittelupuu, klusterijakauma tai riippuvuusanalyysi.

8. Tulosten tulkitseminen. Tiedonluohinnan tuloksia tarkastellaan ja tulkitaan. Tarvittaessa tutkijan on palattava aikaisempiin vaiheisiin ja muutettava jotakin louhinnassa käytettyä osaa, kuten louhitun datan muuttujia tai tiedonlouhinnan metodia tai algoritmia, tulkittavien hahmojen ja representaatioiden selkeyttämiseksi.

9. Löydetyn tiedon hyödyntäminen. Tiedonlouhinnasta löydettyä tietoa voidaan mahdollisesti soveltaa suoraan käytäntöön tai vain dokumentoida ja raportoida myöhempää käyttöä varten.

Tiedonlouhinnasta tulee koulutuksellista tiedonlouhintaa, kun louhittava data on peräisin koulutukseen liittyvästä lähteestä. Koulutuksellisen tiedonlouhinnan piireissä käytetyt tiedonlouhinnan metodit myös poikkeavat hieman muun tiedonlouhinnan metodeista. (Baker 2010)

3.2 Tiedonlouhinnan ja koulutuksellisen tiedonlouhinnan historiaa

Smyth kuvailee artikkelissaan (2000) lyhyesti tiedonlouhinnan historiaa. Smyth näkee tiedonlouhinnan kehityksen olleen vahvasti yhteydessä koneoppimisen ja tietokantojen kehittymiseen ja aloittaa tarinansa tiedonlouhinnan historiasta 1950-luvulta.

1950-luvulla koneoppimisen saralla pyrittiin kehittämään algoritmeja ja laitteita, jotka pystyisivät oppimaan ihmisen tavoin niille syötetyn datan pohjalta. Ajan myötä tutkimuksen pääpaino kuitenkin siirtyi ihmismäisen oppimisen jäljittelystä tietyissä tilanteissa hyvin suoriutuvien ja oppivien algoritmien kehitykseen. Tämä kehityssuunta johti osittaisiin päällekkäisyyksiin tilastotieteiden kanssa. Koneoppimisen juuret näkyvät tiedonlouhinnassa mm. puu- ja sääntöpohjaisten algoritmien käyttönä. Myös tilastotieteiden osallisuus on läsnä, sillä useissa tiedonlouhinnan artikkeleissa käytetään tilastotieteellisiä termejä.

Kun 1980- ja 1990-luvun vaihteessa digitaalisten tietokantojen kehitys alkoi olla siinä vaiheessa, että niiden käyttö oli mahdollista esimerkiksi pankeissa ja lentoyhtiöissä, syntyi markkinarako uudenlaiselle datan käsittelylle. Datan määrän kasvaessa syntyi tarve päästä analysoimaan tätä dataa sen tallentamisen ja noutamisen lisäksi. Tällaisia toimenpiteitä varten oli aloitettava uudenlaisten, kevyempien algoritmien kehitys. Tästä kehityksestä muodostunut tietokantapohjainen tiedon tutkiminen ja analyysi oli nykyisen tiedonlouhinnan peruskivi.

Koulutuksellinen tiedonlouhinta alkoi puolestaan nostaa päätään, kun 1990- ja 2000-luvulla suurten kansainvälisten koulutusta koskevien tutkimusten määrä alkoi kasvaa (Rutkowski ym. 2010). Koulutuksellista tiedonlouhintaa käsittelevien julkaisujen määrä nousi selvästi

vuonna 1999 (Romero ja Ventura 2010) ja vuoteen 2008 mennessä se oli kasvanut omaksi tutkimuksen kentäkseen, kun vuotuinen International Conference on Educational Data Mining ja Journal of Educational Data Mining perustettiin (Baker 2010).

3.3 Koulutuksellisen tiedonlouhinnan metodit

Koulutuksellisen tiedonlouhinnan termin sisään mahtuu useita erilaisia menetelmiä piilotetun tietämyksen etsimiseksi datasta, sekä sen tutkimiseksi ja hyödyntämiseksi. Baker on jakanut nämä metodit viiten eri ryhmään; ennustaminen, suhteiden louhinta, klusterointi, mallien hyödyntäminen ja datan tislaaminen ihmisten arvioitavaksi (Baker 2010).

Ennustaminen (engl. prediction). Ennustamisessa datasta pyritään löytämään muuttujia, joilla on tilastollinen yhteys toiseen muuttujaan. Tästä muuttujien välisestä yhteydestä yritetään muodostaa malli, jolla voitaisiin ennustaa tutkimusta vastaavien olosuhteiden lopputuloksia uusissa, ennennäkemättömissä tilanteissa. Tätä koulutuksellisen tiedonlouhinnan muotoa voidaan käyttää esimerkiksi oppilaiden oppimistulosten ennustamiseen tai mallien luomisessa tärkeiden elementtien kartoittamiseen.

Baker jakaa artikkelissaan ennustamisen vielä kolmeen alaluokkaan. Nämä alaluokat ovat *luokittelu* (engl. classification), *regressio* (engl. regression) ja *tiheysjakaumien arviointi* (engl. density estimation). Erona näiden välillä on ennustettujen muuttujien tyypit ja käytetyt metodit.

Suhteiden louhinta (engl. relationship mining). Suhteiden louhimisessa keskitytään ryhmän nimen mukaisesti etsimään muuttujien välisiä suhteita. Tutkimuksen keskipisteenä voi olla löytää muuttuja, joka on vahvimmassa suhteessa tutkittuun muuttujaan, tai etsiä kaksi muuttujaa, jotka ovat keskenään voimakkaimmassa suhteessa.

Myös suhteiden louhinta jaetaan Bakerin artikkelissa tarkempiin alaluokkiin, joista jokaisessa etsityt suhteet muuttujien välillä ovat erilaiset. *Assosiaatiosääntöjen louhinnassa* (engl. association rule mining) etsitään jos-niin tyylisiä suhteita muuttujien välillä, *korrelaatiolouhinnassa* (engl. correlation mining) etsitään positiivisia tai negatiivisia lineaarisia korrelaatioita muuttujien välillä, *peräkkäisten hahmojen louhinnassa* (engl. sequential pattern

mining) muuttujien väliltä etsitään väliaikaisia suhteita ja *kausaalisessa tiedonlouhinnassa* (engl. causal data mining) taas tutkitaan, johtaako tietty tilanne toiseen.

Muuttujien välisiä suhteita etsiessä tutkijoiden on pidettävä mielessään ehdot, jotka tieteellisesti hyväksyttävien suhteiden on täytettävä. Nämä ehdot ovat tilastollinen merkittävyys ja kiinnostavuus. Tilastollisella merkittävyydellä tarkoitetaan, että löydetyn suhteen aitous ja voimakkuus testataan tilastollisilla menetelmillä. Kiinnostavuudella puolestaan viitataan löydettyjen suhteiden vahvuuteen ja merkittävyyteen tutkimuksen kentällä ja arkielämässä.

Mallipohjainen etsintä (engl. discovery with models). Tutkimuksen keskipisteessä on tutkittua tapahtumaa koskeva malli, joka on muodostettu käyttäen ennustusta, klusterointia tai jotain muuta mallinnusmenetelmää. Tehtyä mallia käytetään toisen tutkimuksen osana, tai sitä hyödynnetään mallien muodostamisen edistämiseksi. Mallin käyttötarkoitus vaihtelee jatkotutkimuksen perusteella. Esimerkiksi ennustuksessa mallia voidaan käyttää määrittämään muuttujat, joiden avulla muita muuttujia pyritään ennustamaan.

Tiedon tislaus ihmisten arvioitavaksi (engl. distillation of data for human judgement). Mikäli dataa onnistutaan kuvaamaan oikein, voidaan siitä jo silmin havaita sellaisia ominaisuuksia, joita tiedonlouhinnan algoritmitkaan eivät välttämättä löydä helposti. Varsinaisesti tiedon tislauksessa keskitytään siis datan kuvaamiseen.

Tiedon tislaamisessa pyritään usein joko identifikaatioon tai luokitteluun. Identifikaatiossa pyritään kuvaamaan dataa siten, että ihmiset pystyvät omin silmin erottamaan datassa esiintyviä tekijöitä. Luokittelussa puolestaan pyritään nimeämään datassa esiintyviä ominaisuuksia, jotta niitä voitaisiin hyödyntää myöhemmässä tutkimuksessa.

3.4 Koulutuksellisen tiedonlouhinnan nykytilasta

Alejandro Peña-Ayala tarkasteli artikkelissaan (2014) koulutuksellisen tiedonlouhinnan kentän nykytilaa analysoimalla 240 koulutuksellisen tiedonlouhinnan työtä, jotka oli julkaistu aikaisintaan vuonna 2010 ja viimeistään vuoden 2013 ensimmäisen neljänneksen aikana. Peña-Ayala rajasi artikkeleidensa valintaa ottamalla mukaan vain aikakauslehdissä, kirjoissa tai koulutuksellisen tiedonlouhinnan konferensseissa tai työpajoissa julkaistuja artikkeleita.

Peña-Ayala analysoi valitsemansa otoksen käyttäen tilastollisia menetelmiä ja KDD:ta. Lopputuloksena Peña-Ayala sai poikkileikkauskuvan viime vuosina tehdyistä koulutuksellisen tiedonlouhinnan tutkimuksista, niissä käytetyistä lähestymistavoista, algoritmeista, metodeista ja osa-alueista, jonka lisäksi hän myös analysoi koulutuksellisen tiedonlouhinnan nykytilaa SWOT-analyysilla.

Taulukkoon 2 olen tiivistänyt Peña-Ayalan listauksen nykyaikaisessa koulutuksellisessa tiedonlouhinnassa käytetyimmistä perusmenetelmistä, metodeista, tekniikoista, algoritmeista, tehtävistä, osa-alueista, tietojärjestelmistä, malleista, sisällöistä ja alustoista. Taulukkoa lukiessa on kuitenkin hyvä pitää mielessä, että yksittäisessä tutkimuksessa on saatettu käyttää useampia metodeja ja algoritmeja.

Oppilaan käyttäytymismallit (engl. student behavior modeling) oli Peña-Ayalan analyysin mukaan suosituin koulutuksellisen tiedonlouhinnan osa-alue, jonka osuus analysoiduista tutkimuksista oli noin 22%. Siinä keskitytään tutkimaan oppilaiden käytöstä ja toimintaa tietynlaisissa tilanteissa, yleensä opetuksen tai jonkin oppimisalustan kehittämiseksi (Peña-Ayala 2014). Oppilaan käyttäytymismalleilla ollaan tutkittu esimerkiksi Tel Aviv Yliopiston Moodle-ympäristön käyttöä. Yli tuhannen oppilaan aineistoa analysoitiin tilastollisin menetelmin ja päätöspuu-algoritmilla, jonka tuloksena oppilaat saatiin jaettua viiteen ryhmään. Lopullisena tuloksenaan tutkijat totesivat, että 46% opiskelijoista vähensi aktiivisuuttaan oppimisympäristössä kurssin aikana tai lopettivat sen käytön kokonaan, kun taas 42% oppilaista lisäsi aktiivisuuttaan tai aloitti koko alustan käytön kunnolla vasta kurssin loppusuoralla (Hershkovitz ja Nachmias 2011). Eräässä toisessa oppilaan käyttäytymismallien tutkimuksessa selvitettiin oppilaiden valitsemien istumapaikkojen yhteyttä näiden testituloksiin ohjelmoinnin kurssilla. Oppilaiden suoriutuminen mitattiin sähköisellä testaussysteemillä. Datan käsittelyssä tutkijat onnistuivat muodostamaan oppilaista ryhmiä, joiden suoriutumisella vaikutti olevan yhteyttä valittujen istumapaikkojen kanssa. Oppilaiden suoritukset näyttivät olevan sitä parempia, mitä harvemmin he vaihtoivat istumapaikkaa kurssin aikana (Ivančević, Čeliković ja Luković 2010).

Viime vuosina koulutuksellisen tiedonlouhinnan tutkimuksessa suosituimmat...	
Perusmenetelmät (% osuus otannasta)	**Tehtävät** (% osuus otannasta)
1. Todennäköisyyslaskenta (37.27%) 2. Koneoppiminen (33.21%) 3. Tilastotiede (17.34%)	1. Luokittelu (42.15%) 2. Klusterointi (26.86%) 3. Regressio (15.29%)
Metodit (% osuus otannasta)	**Tekniikat** (% osuus otannasta)
1. Bayes-menetelmät (19,67%) 2. Päätöspuut (18.03%) 3. Instanssi-pohjaiset, laiskat menetelmät (9.02%)	1. Logistinen regressio (17.86%) 2. Lineaarinen regressio (11.61%) 3. Frekvenssit (8.93%)
Algoritmit (% osuus otannasta)	**Sisällöt ja alustat** (% osuus otannasta)
1. K-Means (6.93%) 2. EM-Algoritmi (5.47%) 3. J48 (5.47%)	1. Algebra (15.38%) 2. ASSISTments (oppimisen tuki) (14.62%) 3. Moodle (10.00%)
Osa-alueet (% osuus otannasta)	**Tietojärjestelmät** (% osuus otannasta)
1. Oppilaan käyttäytymismallit (21.62%) 2. Oppilaan suoriutumismallit (20.72%) 3. Arviointi (20.27%)	1. Älykkäät ohjausjärjestelmät (39.64%) 2. Opiskelun hallintajärjestelmä (9.01%) 3. Perinteinen opetus (9.01%)
Mallit (% osuus otannasta)	
1. Bayes-verkko (40%) 2. Dynaaminen Bayes-verkko (12.50%)	

Taulukko 2: Viime vuosina koulutuksellisessa tiedonlouhinnassa suosituimmat perusmenetelmät, tehtävät, metodit, tekniikat, algoritmit, osa-alueet, tietojärjestelmät, mallit, sisällöt ja alustat (Peña-Ayala 2014).

Oppilaan suoriutumismalleja (engl. student performance modeling) käytettiin Peña-Aylan analyysin perusteella toiseksi eniten ja tämän osa-alueen osuus oli noin 21% analysoiduis-

ta tutkimuksista. Oppilaan suoriutumismalleilla tutkitaan mm. oppilaan tarkkuutta, nopeutta ja tehokkuutta. Tarkoituksena on selvittää, miten nämä oppilaan ominaisuudet muuttuvat tietyissä tilanteissa ja kuinka hyvin oppilaat suoriutuvat heille annetuista tehtävistä (Peña-Ayala 2014). Tässä osa-alueessa on verrattu mm. erään ANN (Artificial Neural Network) pohjaisen adaptiivisen oppimisjärjestelmän avulla opiskelleiden oppilaiden oppimistuloksia ja tavallisella verkkokurssilla opiskelleiden oppimistuloksia. Oppilaiden oppimista mitattiin sanaston, kieliopin ja lukemisen saralla ja tutkijoiden esittämällä adaptiivisella järjestelmällä opiskeleminen tuotti parempia tuloksia (Wang ja Liao 2011). Osa-alueen tutkimuksissa ollaan myös yritetty löytää hahmoja (engl. pattern) yliopisto-opiskelijoiden luonteenpiirteiden ja yliopistoa edeltävien ominaisuuksien yhteyttä yliopistossa suoriutumisen välillä. Tavoitteena on ollut kehittää järjestelmä, jolla voitaisiin ennustaa oppilaiden suoriutumista taustatietojen perusteella. Tulosten perusteella päätöspuuluokittelija (J48) ennusti tuloksia parhaiten, mutta senkin tarkkuus jäi alle 70% (Kabakchieva, Stefanova ja Kisimov 2010).

Peña-Ayalan analyysin perusteella *arviointi* (engl. assessment) oli viimeaikaisen koulutuksellisen tiedonlouhinnan kolmanneksi suosituin osa-alue (noin 20% analysoiduista töistä). Tämän osa-alueen tutkimuksissa keskitytään oppilaiden suoriutumiseen ja taitojen kehittymiseen, paneutuen kuitenkin pintaa syvemmälle ja tutkaillen asiaa tarkemmalla tasolla (Peña-Ayala 2014). Tutkimuksessa ollaan pyritty mm. selvittämään MAT:n (Multidimensional Adaptive Testing) käyttömahdollisuuksia PISA:n lukemisen tulosten analysointiin käyttäen simuloituja aineistoja. Tutkijoilla oli käytössään 14 624 oppilaan aineisto, jota he hyödynsivät MAT:n simuloinnissa. Lopputuloksena tutkijat totesivat, että perinteiseen, PISA 2006 julkaisuihin pohjautuvaan testausmenetelmään verrattuna, MAT:n käyttö lisäsi testauksen tehokkuutta jopa 74% ja vähensi tarvittavien arviointiyksiköiden (engl. item) määrää huomattavasti (Frey ja Seitz 2011). Turkissa on yritetty koulutuksellisen tiedonlouhinnan keinoin muodostaa malleja, joiden avulla voitaisiin ennustaa oppilaiden suoriutumista toisen asteen pääsykokeissa ja samalla selvittää oppilaan tuloksiin vaikuttavia tekijöitä. Tutkimuksessa C5 päätöspuu todettiin parhaimmaksi ennustuksen työkaluksi 95% tarkkuudella ja vahvimmiksi tuloksiin vaikuttaviksi tekijöiksi todettiin aikaisempi kokemus testistä, oppilaan mahdollisesti saamat stipendit ja sisarusten määrä (Şen, Uçar ja Delen 2012).

Edellisten osa-alueiden lisäksi tutkijat ovat olleet huomattavan kiinnostuneita oppilaiden

mallintamisesta yleensä (engl. student modeling), sekä oppilaiden tukemisesta ja heidän antamastaan palautteesta (engl. student support and feedback) (Peña-Ayala 2014). Jälkimmäisen puitteissa on tutkittu mm. oppilaan ja opettajan välistä vuorovaikutusta tilanteissa, joissa opettaja on kaksin oppilaan kanssa. Tutkimusta varten videoitiin 50 tunnin edestä materiaalia, joka sitten analysoitiin ja josta eroteltiin dialogisia siirtoja (engl. move), jotka saattoivat olla puhetta, toimintoja tai oppilaan antamia kvalitatiivisia vastauksia. Näistä siirroista saatu data käsiteltiin koulutuksellisen tiedonlouhinnan keinoin ja siitä onnistuttiin muodostamaan kolme hahmoa, jotka kuvasivat oppilaan ja opettajan välistä yhteistyöhön pyrkivää vuorovaikutuksen tapaa (D'Mello, Olney ja Person 2010).

Perinteistä opetusta enemmän tutkijat olivat kiinnostuneita sähköisistä oppimisen keinoista, kuten älykkäistä opinto-ohjauksen järjestelmistä ja oppimisen hallinnan järjestelmistä. Tarkempaa opetuksen osa-aluetta hakiessa suosituimmiksi tutkimusaiheiksi nousivat Algebra, ASSISTments ja Moodle.

Metodeista suosituimpia olivat Bayesin Teoreema, Puudiagrammi ja IBL (Instance-based learning) ja tekniikoista suosituimmat olivat logistinen regressio, lineaarinen regressio ja frekvenssit. Tutkimuksissa käytetyistä algoritmeista suosituimpia olivat K-means, EM-algoritmi ja J48, kun taas malleista suosituimpia olivat Bayes-verkko ja dynaaminen Bayes-verkko.

Artikkelinsa lopussa Peña-Ayala antoi oman SWOT-analyysinsa koulutuksellisesta tiedonlouhinnasta tieteenä. Vahvuuksiksi hän mainitsi tieteenalan hyvät ja kypsyneet perusteet, avoimen ja aktiivisen yhteisön ja tiedonlouhinnan ja koulutuksen tutkimuksen kentiltä saatavan tuen. Koulutuksellisen tiedonlouhinnan mahdollisuudet ovat Peña-Ayalan mielestä suuret, sillä koulutus ja sen kehittäminen ovat korkealla prioriteetilla tämän hetkisessä maailmassa. Opetuksen kehityksen ongelmiin ja haasteisiin voidaan löytää ratkaisuja myös koulutuksellisen tiedonlouhinnan avulla.

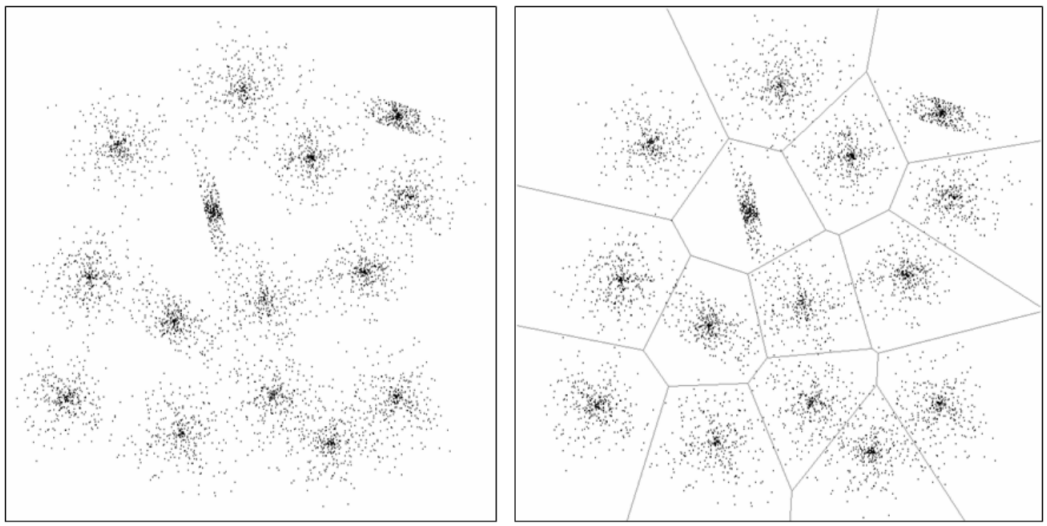
Koulutuksellisen tiedonlouhinnan heikkoudeksi Peña-Ayala puolestaan näkee sen kapean tutkimusalueen. Kun tutkimuksessa keskitytään vain yhteen tai kahteen tutkimusalueeseen, kuten esimerkiksi hänen analyysissaan yli 40% tutkimuksista käsitteli oppilaan mallinnusta, muiden tutkimusalueiden kehitys kärsii ja jää alikehittyneeksi. Koulutuksellisen tiedonlouhinnan julkaisut keskittyvät myös liikaa vain tutkimusalan omiin konferensseihin ja leh-

tiin, jolloin nämä saavat vähemmän huomiota tiedonlouhinnan piireissä. Peña-Ayala näkee myös, että koulutuksellinen tiedonlouhinta on liian keskittynyt vain omaan kehitykseensä, siinä missä tutkimuksesta voisi olla apua koko tiedonlouhinnan tieteenalan kehittämiseen. Tieteenalan uhkia ovat analyysin perusteella tyypilliset orastavan tieteenalan ongelmat, kuten teoriataustan puute ja terminologian kehittymättömyys.

4 Klusterointi tiedonlouhinnan menetelmänä

Klusterointi on datan käsittelyyn tarkoitettu menetelmä, jonka tarkoituksena on löytää datasta havainnot, jotka ryhmittyvät luonnollisesti keskenään. Näitä havaintojen ryhmiä kutsutaan klustereiksi. Klusteroinnin yleisenä tavoitteena on, että saman klusterin sisään saadut havainnot ovat enemmän samanlaisia keskenään kuin muissa klustereissa olevien havaintojen kanssa. Klusterointi on erityisen hyödyllistä, mikäli havainnoista ei ole ennalta muodostettu ryhmityksiä ja se voidaan suorittaa joko ilman alustavaa hypoteesia klustereiden määrästä tai valmiin hypoteesin ja etsittyjen klustereiden määrän kanssa (Baker 2010).

Kuviossa 2 on esimerkki 5000 havainnon aineiston klusteroinnista. Klusteroinnin tuloksena syntyneet 15 klusteria on rajattu viivoilla, jotka puolestaan ovat muodostettu piirtämällä Voronoin diagrammi klustereiden keskipisteiden suhteen (Aurenhammer 1991).



Kuvio 2: Esimerkki aineiston klusteroinnista (Tuononen 2005).

Klusterointia ollaan hyödynnetty vuosien varrella monissa empiirisissä tieteissä, kuten biologiassa, psykiatriassa, psykologiassa, lääketieteessä, arkeologiassa, geologiassa ja sosiologiassa. Tiedonkäsittelytieteiden yhteydessä klusterointia käytetään myös tiedonlouhinnan yhteydessä (kts. Tuononen, 2005,ja lähteet siellä) ja viimeaikaisessa koulutuksellisessa tiedonlouhinnassa klusterointi on yksi käytetyimmistä tutkimusmenetelmistä (Peña-Ayala 2014; katso tarkemmin luku 3.4).

Tässä kappaleessa esittelen lyhyesti erilaisten klusterointimenetelmien lajittelun ja K-means++ -algoritmin, jota käytin omassa tutkimuksessani tutkimusaineistoni klusteroimiseen.

4.1 Klusterointimenetelmiä

Klusterointia varten on kehitetty lukemattomia menetelmiä ja algoritmeja, joiden luokittelemiseksi on myös yritetty tehdä useita linjauksia, mutta hyvin usein eri kirjoittajien esitykset aiheesta eroavat toisistaan. Karkeasti ottaen klusterointimenetelmät voidaan kuitenkin jakaa kahteen luokkaan: hierarkkisiin menetelmiin ja osittaviin menetelmiin. Tutkijat valitsevat tutkimuksessaan käytettävän menetelmän klusteroitavan aineiston ja klusteroinnin käyttötarkoituksen mukaan. Luonnollisesti yhdessä tutkimuksessa voidaan käyttää myös useampaa klusterointimenetelmää monipuolisempien tulosten saamiseksi (Tuononen 2005), (Filippone ym. 2008) & (Zaki ja Meira Jr 2014).

Käytän taulukossa 3 olevassa jaottelussa pohjana Tuonosen (2005), Zakin (2014) ja Filipponen ja muiden (2008) käyttämiä jaotteluja.

Klusterointimenetelmät	
Hierarkkiset menetelmät	Osittavat menetelmät
- Yhdistävät menetelmät	- Optimointimenetelmät
- Jakavat menetelmät	- Verkkoteoreettiset menetelmät
	- Tiheysperustaiset menetelmät
	- Ristikkoperustaiset menetelmät
	- Malliperustaiset menetelmät

Taulukko 3: Karkea luokittelu klusterointimenetelmistä.

4.1.1 Hierarkkiset menetelmät

Hierarkkisten klusterointimenetelmien tarkoituksena on löytää klusteroitavasta aineistosta klustereiden rakenne, joka voidaan esittää lopulta puurakenteena tai klustereiden hierarkkisuutta kuvaavana dendrogrammina. Hierarkkiset menetelmät voidaan jakaa vielä kahteen

pääluokkaan: Yhdistäviin menetelmiin ja jakaviin menetelmiin (Tuononen 2005) & (Zaki ja Meira Jr 2014).

Yhdistävät menetelmät toimivat ns. bottom-up periaatteella. Klusterien yhdistäminen alkaa vertaamalla jokaisen yksittäisen klusterin havaintoja toisiinsa, jonka jälkeen klusterit yhdistetään toisiinsa ennalta määritetyn kriteerin mukaisesti. Tämä kriteeri voi olla esimerkiksi klusterien lähimmät alkiot tai kauimmat alkiot. Klusterien varsinaisen yhdistämisen suoraviivaisuuden takia yhdistämisen kriteerin valinta on yhdistävien menetelmien käytössä erittäin tärkeää (Tuononen 2005) & (Zaki ja Meira Jr 2014).

Jakavat menetelmät toimivat päinvastoin verrattuna yhdistäviin menetelmiin. Menetelmä aloittaa tarkastelemalla kaikkia klusteriin kuuluvia pisteitä, jonka jälkeen klusteri jaetaan rekursiivisesti pienempiin klustereihin. Menetelmän käytössä on tärkeää päättää, miten jaettava klusteri valitaan ja kuinka klusterien jako pienempiin osiin toteutetaan. Jaettava klusteri voidaan valita esimerkiksi koon tai hajonnan perusteella, kun taas klusterien jakaminen voidaan suorittaa esimerkiksi pääakselia pitkin (Tuononen 2005) & (Zaki ja Meira Jr 2014).

4.1.2 Osittavat menetelmät

Osittavien klusterointimenetelmien tarkoituksena on löytää klusteroitavasta aineistosta vain yksittäinen klusterointi. Menetelmät perustuvat usein jonkin tietyn tavoitefunktion optimoimiseen. Osittavat menetelmät voidaan jakaa mm. seuraaviin alaluokkiin: Optimointimenetelmät, verkkoteoreettiset menetelmät, tiheysperustaiset menetelmät, ristikkoperustaiset menetelmät ja malliperustaiset menetelmät (Tuononen 2005).

Optimointimenetelmissä aineisto jaetaan ennalta määritettyyn määrään klustereita ja pyritään joko minimoimaan tai maksimoimaan annettu tavoitefunktio, joka kuvaa klusterointivirhettä (kts. Tuononen, 2005,ja lähteet siellä). Tutkijan täytyy siis päättää etsittyjen klustereiden määrä ennen menetelmän käyttöä, jonka helpottamiseksi hän voi käyttää esimerkiksi klusteri-indeksejä, joita käsittelen tarkemmin kappaleessa 4.3. Tutkimuksessani käytetty K-means++ -algoritmi kuuluu optimointimenetelmien luokkaan ja esittelen sen tarkemmin kappaleessa 4.2.

Verkkoteoreettiset menetelmät keskittyvät ongelman ratkaisemiseen luomalla klusteroitavasta aineistosta verkon, jolla ongelma voidaan ratkaista. Tekstissään Tuononen (2005) antaa esimerkin tällaisesta tilanteesta, joka on pienimmän virittävän puun muodostaminen, jossa aineiston havainnot toimivat solmuina ja kaaren painona näiden havaintojen väliset välimatkat. Poistamalla painavin kaari voidaan muodostaa klusterointeja eri klusterien lukumäärille (kts. Tuononen, 2005,ja lähteet siellä).

Tiheysperustaiset menetelmät keskittyvät etsimään klusteroitavasta aineistosta klusterit, joiden sisällä olevat havainnot ovat mahdollisimman tiheässä kasassa. Menetelmien avulla voidaan löytää mielivaltaisen kokoisia ja muotoisia klustereita, mutta niiden käyttö vaatii klusteroitavan aineiston tuntemista (kts. Tuononen, 2005,ja lähteet siellä).

Ristikkoperustaisissa menetelmissä klusteroitavan aineiston havaintojen muodostama taso jaetaan äärelliseen määrään neliöitä, jotka muodostavat yhdessä ristikkorakenteen. Menetelmän kaikki klusterointitoimenpiteet suoritetaan tämän ristikkorakenteen sisällä. (kts. Tuononen, 2005,ja lähteet siellä).

Malliperustaisilla menetelmillä pyritään hakemaan aineistolle mahdollisimman hyvä selitys jotakin matemaattista mallia käyttäen. Tämä matemaattinen malli voi olla esimerkiksi tilastotieteellinen tai neuroverkkoon pohjautuva (kts. Tuononen, 2005,ja lähteet siellä).

4.2 K-means ja K-means++

Monista klusterointia varten luoduista algoritmeista valitsin tutkimukseeni K-means++ -algoritmin. K-means on yksi suosituimmista klusterointialgoritmeista, myös koulutuksellisessa tiedonlouhinnassa (katso luku 3.4), ja sen juuret ulottuvat 1950-luvulle asti, jolloin se sai alkunsa Fisherin datan luokitteluun liittyvissä tutkimuksissa. 1960-luvulla Forgy ja MacQueen kehittivät omat versionsa K-meansista, jotka ovat edelleen käytetyimpien K-means -algoritmien joukossa. Suurin ero näiden kahden algoritmin välillä on siinä, milloin tutkitut havainnot lisätään osaksi klusteria ja milloin klustereiden keskukset päivitetään. MacQueenin algoritmi päivittää klusterin keskuksen jokaisen havainnon lisäyksen jälkeen ja vielä kerran kaikki keskukset, kun kaikki havainnot on liitetty klustereihinsa. Forgyn algoritmi puolestaan päivittää klusterien keskukset vasta, kun kaikki havainnot on lisätty niitä

lähimpänä oleviin keskuksiin (kts. Äyrämö, 2006,ja lähteet siellä).

K-meansia kutsutaan joissakin piireissä myös Loydin algoritmiksi, koska vuonna 1982 Stuart P.Lloyd esitteli työssään kvantisointimenetelmän (engl. quantization algorithm), joka muistuttaa Forgyn algoritmia ja siksi hänen nimeään käytetään myös K-means -klusteroinnin yhteydessä (kts. Äyrämö, 2006,ja lähteet siellä).

Tutkimuksessani käyttämässäni Matlab-ohjelmistossa K-means -klusterointi suoritetaan oletusarvoisesti käyttäen K-means++ -algoritmia ("k-means clustering - MATLAB kmeans - MathWorks Nordic" 2016). K-means++ on rakennettu Lloydin algoritmin pohjalta ja sen tarkoituksena on parantaa klusterointitulosten tarkkuutta algoritmin tehokkuutta vaarantamatta. Tätä klusterointitulosten tarkennusta tavoitellaan keskittymällä algoritmin alussa tapahtuvaan klusterien keskusten sijaintien valintaan. K-means++ -algoritmissa keskusten valinnassa hyödynnetään tiheysfunktiota, joka puolestaan nostaa todennäköisyyttä sille, että valitut keskusten sijainnit ovat selvästi erillään toisistaan. Luultavasti tästä ominaisuudesta johtuen K-means++ -algoritmista on tullut viimeaikoina suosituin K-means -klusteroinnin variantti (Arthur ja Vassilvitskii 2007) & (Hämäläinen ja Kärkkäinen 2016). Vaiheittan K-means++ -algoritmi etenee seuraavasti:

1. Valitaan ensimmäinen keskus c_1 satunnaisesti datajoukosta D

2. Valitaan seuraava keskus $c_i = x' \in D$ todennäköisyydellä $\frac{D(x')^2}{\sum_{x \in \chi} D(x)^2}$.

3. Toistetaan vaihetta 2 kunnes k-määrä keskuksia on valittu.

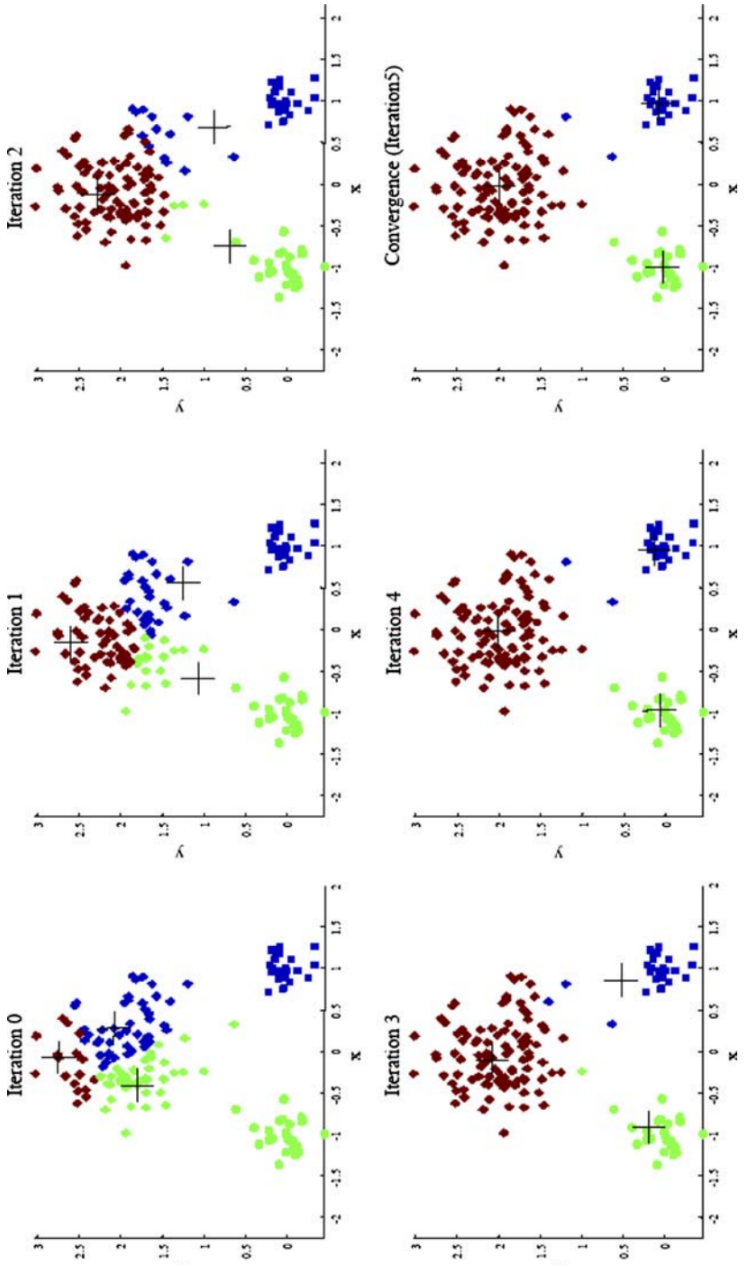
4. Asetetaan jokaisella $i \in \{1, ..., k\}$ klusteri C_i pistejoukoksi D:stä, jonka pisteet ovat lähempänä c_i kuin c_j kaikilla $j \neq i$ arvoilla.

5. Asetetaan jokaiselle $i \in \{i, ..., k\}$ c_i massakeskipisteeksi kaikille pisteille, jotka kuuluvat joukkoon C_i: $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.

6. Toistetaan kohtia 4 ja 5, kunnes mikään C_i ei muutu.

Tiivistetysti sanottuna K-means -klusterointi on siis iteratiivinen prosessi, jossa tutkittu data jaetaan $k$hon ryhmään. Äyrämö arvelee K-meansin suosion syiksi sen helppokäyttöisyyttä ja vähäisiä resurssivaatimuksia verrattuna esimerkiksi hierarkkisiin menetelmiin (Äyrämö

2006). Kuviossa 3 on kuvattu esimerkki havaintojen jakautumisesta klustereihin ja klusteri-prototyyppien sijainnin muutoksista K-means -klusteroinnin aikana.



Kuvio 3: Esimerkki klustereiden prototyyppien sijainneista (merkitty + -merkillä) ja datan jaottelusta klustereihin (merkitty värein) K-means -algoritmin suorituksen aikana. (Wu ym. 2008)

4.3 Klusteri-indeksit

Klusterointia varten on muodostettu useita erilaisia algoritmeja, joita tutkijat käyttävät etsimiensä klustereiden ja käyttämänsä datan ominaisuuksien mukaan. Jotta tutkijat voisivat olla varmoja klusterointinsa tarkkuudesta, on tämän tarkistamista varten kehitetty omat työkalunsa. Nämä työkalut keskittyvät kolmeen tehtävään: Klusteroinnin arviointi (engl. clustering evaluation), klustereiden olemassaolon ennakkoarviointi (engl. clustering tendency) ja klusteroinnin vakaus (engl. clustering stability) (Zaki ja Meira Jr 2014).

Klusteroinnin arvioinnin avulla tutkija pyrkii selvittämään klusterointinsa laadukkuutta ja tarkkuutta. Klustereiden olemassaolon ennakkoarvioinnilla tutkija pystyy tarkistamaan klusteroinnin soveltuvuuden hänen käyttämälleen aineistolle ja klusteroinnin vakaudella tutkija voi tarkastella klusterointialgoritmin suorittamiseen liittyvien parametrien vaikutusta klusteroinnin tuloksiin (Zaki ja Meira Jr 2014).

Arvioinnissa käytetyn datan pohjalta klusteroinnin arviot voidaan jakaa kolmeen pääryhmään. Ulkoisissa (engl. external) arvioissa arviointi suoritetaan käyttäen dataa, joka ei varsinaisesti sisälly klusteroinnin kohteena olevaan dataan. Sisäisessä (engl. internal) arvioinnissa käytetään klusteroitavaan dataan sisältyviä mittareita, kuten klustereiden rakenteita ja niiden välisiä etäisyyksiä. Suhteellisessa (engl. relative) arvioinnissa puolestaan hyödynnetään klusterointien välisiä tuloksia, jotka saadaan esimerkiksi ajamalla samaa klusterointialgoritmia erilaisilla parametreillä.

Tutkimukseni kannalta relevantteja klusteroinnin arviointimenetelmiä ovat Davies-Bouldin, Silhouette, Calinski-Harabasz ja Gap. Näistä kaksi ensimmäistä kuuluvat sisäisten arviointien ryhmään ja kaksi jälkimmäistä suhteellisten arviointien ryhmään.

Alla olevissa kaavoissa lyhenne SSW tarkoittaa klustereiden sisäistä virhettä eli neliövirhesummaa klustereiden sisällä (engl. sum of squares within clusters) ja lyhenne SSB tarkoittaa klustereiden välistä virhettä eli neliövirhesummaa klustereiden välillä (engl. sum of squares between clusters).

Davies-Bouldin. Olkoon $R_{ij} = \frac{(SSW_i + SSW_j)}{DC_{ij}}$, missä SSW_i on $i$nnen klusterin sisäinen neliövirhe ja DC_{ij} klustereiden i ja j prototyyppien välinen neliöetäisyys. Määritellään $R_i = \max_{j \neq i} R_{ij}$

ja Davies-Bouldin indeksi $R = \frac{1}{k}\sum_{i=1}^{k} R_i$, missä k on klustereiden lukumäärä. Alhaisimman Davies-Bouldin -arvon antama arvio osoittaa optimaalisen klustereiden määrän (Davies ja Bouldin 1979).

Calinski-Harabaz. $C = \frac{(N-k)BGSS}{(K+1)WGSS}$, jossa $BGSS = \sum_{k-1}^{K} |C_k||C_k - m||^2$ ja $WGSS = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - C_k||^2$. Kaavoissa oleva BGSS on ryhmien välinen hajonta ja WGSS on yhdistetty klusterin neliöiden summa. Tässä muodossa indeksin maksimi indikoi klustereiden määrän. (Desgraupes 2013) & (Caliński ja Harabasz 1974)

Silhouette. Silhouette arviossa jokaiselle havainnolle lasketaan arvo $S(i)$ kaavalla $S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$, jossa $a(i)$ on havainnon keskimääräinen etäisyys muista saman klusterin havainnoista ja $b(i)$ on havainnon pienin keskimääräinen etäisyys muissa klustereissa oleviin havaintoihin. Silhouette arvo vaihtelee välillä $[-1,1]$ ja mitä useammalla havainnolla on korkea arvo, sitä parempi arvion tulos on. (Rousseeuw 1987)

Gap. $Gap_n(k) = E_n^*\{log(W_k)\} - log(W_k)$, jossa n on otannan koko, k on arvioitavien klustereiden määrä ja W_k on yhdistetty klustereiden sisäinen virhe $W_k = sum_{r=1}^{k} \frac{1}{2n_r} SSW_r$, jossa n_r on havaintojen määrä klusterissa r ja SSW_r on neliövirhesumma klusterin r havainnoille. Korkeimman Gap-arvon antama arvio osoittaa optimaalisen klustereiden määrän (Tibshirani, Walther ja Hastie 2001).

5 Vuoden 2012 PISA- ja vuoden 2011 yhdistetyn TIMSS ja PIRLS -aineiston vertailu klusteroinnin avulla

Tässä kappaleessa esittelen varsinaisen tutkimukseni, jossa pyrin selvittämään, pystynkö löytämään K-means++ -klusterointialgoritmia käyttäen vuoden 2012 PISA-aineistosta ja vuoden 2011 yhdistetystä TIMSS ja PIRLS -aineistosta samanlaisia oppilasprofiileja, kun klusterointialgoritmin lisäksi myös käytetyt muuttujat ovat toisiaan vastaavat aineistojen välillä. Aloitan esittelemällä tutkimuksen otannan, klusteroinnissa käytetyt muuttujat ja klusteroinnin tuloksia selittävät muuttujat, sekä niiden valitsemiseen käyttämäni prosessit. Tämän jälkeen jatkan selittämällä datalle tekemäni esivalmistelut, jonka jälkeen kuvailen prosessin, jolla valitsin kummastakin aineistosta etsittävien klustereiden määrän.

5.1 Otannan ja klusteroitavien muuttujien valinta

Pitääkseni tutkimukseni sopivan kokoisena pro gradu -tutkielmaani varten päätin rajoittua vain suomalaisia oppilaita ja kouluja käsittelevää dataa. Vuoden 2012 PISA-aineistossa on 8829 suomalaista oppilasta ja 311 koulua. Vuoden 2011 yhdistetyssä TIMSS ja PIRLS -aineistossa on puolestaan 4541 suomalaista oppilasta ja 145 koulua. Yhdistetyssä TIMSS ja PIRLS -aineistossa käsiteltiin vain neljäsluokkalaisia. Taulukossa 4 olen esitellyt otantani koon ja sukupuolijakauman aineistoittain.

Aineisto	Oppilaiden määrä	Sukupuolijakauma
PISA 2012	8829	4370♀, 4459♂
TIMSS & PIRLS 2011	4541	2223♀, 2318♂
Yhteensä:	13 370	6593♀, 6777♂

Taulukko 4: Otannan koko ja sukupuolijakauma aineistoittain.

Muuttujien valitsemiseksi luin läpi tutkimusten aineistoja varten laaditut koodikirjat ja muuttujia koskevat tiedot tutkimusten teknisistä raporteista ja vertailin niitä keskenään (P. OECD 2014), (IEA 2014), (IEA 2016), (OECD 2015), (OECD 2014b) & (OECD 2014a). Klus-

terointiin käytettävien muuttujien valinnassa käytin kolmea kriteeriä: muuttujien tuli olla johdettuja muista muuttujista, niiden tuli olla skalaarisia ja samanlaisen muuttujan tuli löytyä kummastakin aineistosta. Nämä kriteerit huomioon ottaen, valitsin klusterointiani varten taulukkoon 5 merkityt muuttujat.

PISA 2012	TIMSS & PIRLS 2011	Mitä muuttujat kuvaavat?
INTMAT	ASBGSLM	Oppilaan kiinnostus ja innostus matematiikkaan.
SCMAT	ASBGSCM	Oppilaan itseluottamus omiin matemaattisiin taitoihinsa.
BELONG	ASBGSBS	Oppilaan tuntemukset kiusatuksi joutumisesta ja kouluun kuulumisesta.
HEDRES	ASBGHRL	Oppilaan kodin oppimista tukevien resurssien määrä.

Taulukko 5: Aineistoista valitut klusterointimuuttujat ja niiden lyhyet kuvaukset.

Muuttujien tarkemmat tiedot, kuten mahdolliset arvot, puuttuvan datan määrä yms. ovat nähtävissä liitteistä **E** ja **F**.

Muuttujat BELONG Sense of Belonging to School ja ASBGSBS *STUDENTS BULLIED AT SCHOOL/SCL* eivät mitanneet keskenään täysin samaa asiaa. Kuitenkin PISA:n muuttujassa BELONG käsiteltiin teemoja, joihin koulussa kiusatuksi tuleminen vaikuttaa suorasti, joten katsoin näiden muuttujien olevan tarpeeksi samanlaisia klusterointia varten.

5.2 Selittävien muuttujien valinta

Koska klusteroinnin tulokset itsessään eivät useinkaan anna mitään hyödyllistä tietoa, oli minun valittava tutkittavista aineistoista myös ns. selittäviä muuttujia tai metamuuttujia, joiden avulla klusteroinnista saatavat tulokset voitaisiin tulkita.

Muuttujien valintaprosessi muistutti edellä mainittua prosessia, jolla valitsin klusteroinnissa

käytettävät muuttujat. Valinnassa käytettävät kriteerit kuitenkin muuttuivat niin, että selittävien muuttujien ei tarvinnut olla skalaarisia tai johdettuja, mutta samanlaisen muuttujan oli löydyttävä myös toisesta aineistoista tai se oli voitava johtaa toisen aineiston muuttujista. Valintaprosessin jälkeen päädyin käyttämään selittävinä muuttujina taulukossa 6 olevia muuttujia.

PISA 2012	TIMSS & PIRLS 2011	Mitä muuttujat kuvaavat
ST04Q01	SEX	Oppilaan sukupuoli
AGE	ASDAGE	Oppilaan ikä
TestLang	ITLANG_P	Oppilaan testissä käyttämä kieli
hisced	ASDHEDUP	Oppilaan vanhempien korkein koulutustaso
ST57Q01	ASBH08	Oppilaan käyttämä aika kotitehtävien tekemiseen

Taulukko 6: Aineistoista valitut selittävät muuttujat ja niiden lyhyet kuvaukset.

Taulukossa 6 olevien muuttujien lisäksi muodostin taulukossa 7 olevat neljä uutta muuttujaa, joita käytin myös metamuuttujina. Eri muuttujien muodostukseen käytetyt toimenpiteet ja kaikkien selittävien muuttujien tarkemmat tiedot ovat nähtävissä liitteistä **G** ja **H**.

Muuttujan nimi	Mistä muuttujista muuttuja muodostettiin	Mitä muuttuja kuvaa
MATH	Oppilaan viiden matematiikan todennäköisen koetuloksen (engl. plausible value) keskiarvo	Oppilaan suoriutuminen matematiikan tehtävissä
SCIENCE	Oppilaan viiden luonnontieteiden todennäköisen koetuloksen keskiarvo	Oppilaan suoriutuminen luonnontieteiden tehtävissä
READING	Oppilaan viiden lukemisen todennäköisen koetuloksen keskiarvo	Oppilaan suoriutuminen lukemisen tehtävissä
SchlSize	PISA:n tapauksessa muuttujista SC07Q01 ja SC07Q02	Oppilaan koulun koko

Taulukko 7: Aineistojen muista muuttujista kootut selittävät muuttujat ja niiden lyhyet kuvaukset.

5.3 Datalle tehdyt esivalmistelut

Tiedonlouhinnan vaiheisiin luetaan myös datan transformaatio (Fayyad, Piatetsky-Shapiro ja Smyth 1996). Tällä transformaatiolla voidaan tarkoittaa puuttuvan datan korvaamista tai muuttujien muuttamista toiseen muotoon klusteroinnin tulosten parantamiseksi. Seuraavaksi esittelen, miten käsittelin käyttämäni aineiston klusterointia varten.

5.3.1 Puuttuvan datan korvaaminen

Mm. Kärkkäinen ja Saarela ovat todenneet tutkimuksissaan (2014), että vuoden 2012 PISA-aineisto sisältää paljon puuttuvaa dataa. Sama koskee myös vuoden 2011 yhdistettyä TIMSS ja PIRLS -aineistoa, joten päätin suorittaa aineistoille imputaation k-means-tyyppisen klusteroinnin mahdollistamiseksi. Kummassakin käsittelemässäni aineistossa käytettiin niin kutsuttuja STRATUM-muuttujia. Näiden muuttujien tarkoituksena on erotella tutkimukseen vastanneet koulut ja oppilaat erilaisiin ryhmiin, joiden sisällä olevat vastaajat olisivat taus-

toiltaan mahdollisimman samanlaisia (P. OECD 2012) ja (Foy 2013). Imputaatiossani käytin näitä STRATUM-muuttujia siten, että korvasin puuttuvan datan oppilaan oman STRATUM-ryhmän keskiarvolla.

Yhdistetyssä TIMSS ja PIRLS -aineistossa STRATUM-muuttujalla muodostettuja ryhmiä löytyi kymmenen, kun taas PISA-aineistossa ryhmiä oli 17. Liitteissä **I** ja **J** on tarkennettu aineistoista löytyneiden STRATUM-ryhmien koot ja sukupuolijakaumat.

5.3.2 Muuttujien arvojen muuntaminen samalle vaihteluvälille

Parantaakseni klusteroinnista saatavia tuloksia ja helpottaakseni tulosten lopullista esittämistä kaavioissa päätin normalisoida käyttämäni datan samalle vaihteluvälille Kantardzicin (2011) esittelemää min-max normalisointia käyttäen:

$$v'(i) = (v(i) - min(v(i)))/(max(v(i)) - min(v(i))).$$

Valitsin vaihteluvälikseni $[-1,1]$, jonka johdosta liitteinä olevissa datan kuvaajissa (liitteet **C** ja **D**) kaikki muuttujien arvot vaihtelevat tällä välillä. Normalisoinnin jälkeen kaikkien klusteroinnissa käytettävien muuttujien merkitys havaintojen ryhmittymiselle on yhtä vahva, koska muuttujien arvojen vaihteluväli on yhtä suuri.

5.4 Klustereiden optimaalisen määrän selvittäminen

K-means++:san käyttämiseksi Matlabissa on ohjelmalle annettava etsittävien klustereiden määrä klusteroinnin aloituskomennossa. Selvittääkseni klustereiden optimaalisen määrän käyttämilleni aineistoille hyödynsin Matlabin *evalclusters* metodia ja alaluvussa 4.3 esittelemiäni klusteri-indeksejä: Davies-Bouldin, Gap, Silhouette ja Calinski-Harabasz.

Pyrin tarkentamaan *evalclustersin* tuloksia tekemällä arviointia varten oman funktion, joka käski ohjelmaa tekemään klusteri-indeksin mukaisen arvion 100 kertaa ja palauttamaan näistä sadasta arviosta parhaimman.

```
myfunc = @(X,K)(kmeans(X,K,'emptyaction','singleton','replicate',100))
```

Itse klusteroinnin arviointi suoritettiin komennolla:

arvio = evalclusters(aineisto,myfunc,'silhouette','klist',[2:10])

Tässä esimerkissä klusteri-indeksinä toimi Silhouette ja tutkittavien klustereiden määrä oli kahdesta kymmeneen klusteria. Suoritin arvion kummallekin aineistolle neljään kertaan käyttäen aikaisemmin mainittuja klusteri-indeksejä.

Taulukossa 8 olen koonnut PISA-aineiston arvioinnin tulokset liitteessä **A** esitettyjen klusteri-indeksien perusteella.

Käytetty klusteri-indeksi	Etsitäänkö lokaalia minimiä vai maksimia	Matlabin ehdottama K:n arvo	Parhaimmat 3 K:n arvoa kuvasta
Davies-Bouldin	Minimi	6	6, 8, 10
Silhouette	Maksimi	6	6, 10, 9
Gap	Maksimi	7	7, 6, 9
Calinski-Harabasz	-	2	2, 3, 4

Taulukko 8: Matlabin ehdottamat klusterien määrät käytettyjen klusteri-indeksien pohjalta PISA-aineistossa.

Taulukossa 9 olen koonnut yhdistetyn TIMSS ja PIRLS -aineiston arvioinnin tulokset liitteessä **B** esitettyjen klusteri-indeksien perusteella.

Käytetty klusteri-indeksi	Etsitäänkö lokaalia minimiä vai maksimia	Matlabin ehdottama K:n arvo	Parhaimmat 3 K:n arvoa kuvasta
Davies-Bouldin	Minimi	2	2, 8, 4
Silhouette	Maksimi	2	4, 3, 2
Gap	Maksimi	4	2, 4, 5
Calinski-Harabasz	-	2	2, 3, 4

Taulukko 9: Matlabin ehdottamat klusterien määrät käytettyjen klusteri-indeksien pohjalta yhdistetyssä TIMSS ja PIRLS -aineistossa.

Tulosten perusteella päätin lukita klustereiden määrän PISA-aineiston kohdalla kuuteen ja yhdistetyn TIMSS ja PIRLS -aineiston kohdalla neljään. PISA-aineiston arvioinneissa $k = 6$ oli paras vaihtoehto kahdessa arviossa, yhdessä arviossa se ylsi kolmen parhaan joukkoon ja viimeisessä arviossa neljän parhaan joukkoon. Yhdistetyn TIMSS ja PIRLS -aineiston arvioinneissa puhtaasti tulosten perusteella paras vaihtoehto olisi ollut $k = 2$, mutta aineiston jakaminen vain kahteen klusteriin ei olisi ollut tutkimuksen kannalta kovinkaan hedelmällistä. Tästä syystä päädyin käyttämään aineiston klusterointiin $k = 4$, koska tämä oli yhdessä arviossa toisiksi paras vaihtoehto ja kolmessa muussa arviossa kolmen parhaimman vaihtoehdon joukossa.

Jo tässä vaiheessa tutkimusta oli havaittavissa, että aineistoista saatavat klusterointitulokset olivat erilaisia klusteroitavien aineistojen välillä.

6 Klusteroinnin tulokset ja tulkinnat

Tässä kappaleessa esittelen suorittamani klusteroinnin tulokset. Aloitan esittelemällä, minkälaisia löydetyt klusterit olivat ja mitkä muuttujat niitä karakterisoivat. Tämän jälkeen peilaan näitä tuloksia selittäviin muuttujiin, tutkailen niiden yhteyksiä toisiinsa ja erittelen eri aineistojen välisten tulosten yhtäläisyyksiä ja eroavaisuuksia.

6.1 Löydetyt klusterit

Olin tutkimuksen aikaisemmassa vaiheessa määrittänyt etsittyjen klustereiden määrän, jotka olivat PISA-aineistolle kuusi ja yhdistetylle TIMSS ja PIRLS -aineistolle neljä (katso kappale 5.4). Suoritin klusteroinnin Matlab-ohjelmistolla seuraavaa komentoa käyttäen:

[IDX, C] = kmeans(aineisto,k,'Replicates',100)

Ylläolevassa esimerkissä k tarkoittaa etsittyjen klustereiden määrää, joka vaihteli klusteroitavan aineiston mukaan.

Klusteroinnista sain ulos kaksi taulukkoa. IDX-taulukossa oli merkittynä kaikki indeksit, joista kuhunkin klusteriin kuuluvat vastaajat löytyivät. C-taulukossa oli puolestaan nähtävissä niin kutsutut klusteriprototyypit, jotka kuvaavat klustereihin kuuluvien vastaajien keskiarvoa. Nämä klusteriprototyypit ovat nähtävissä liitteistä **C** ja **D**. Klustereiden koot ja sukupuolijakaumat ovat nähtävissä taulukosta 10. Kuvaillessani klustereita karakterisoivia muuttujia käytän termiä prototyyppioppilas. Termin avulla pyrin havainnollistamaan paremmin, millainen kyseisen klusterin keskiverto-oppilas on.

Klusterin aineisto, numero ja koodi	Klusterin koko	Klusterin sukupuolijakauma
PISA 1 (P1)	2443	1234♀, 1209♂
PISA 2 (P2)	2637	1233♀, 1404♂
PISA 3 (P3)	934	494♀, 440♂
PISA 4 (P4)	869	360♀, 509♂
PISA 5 (P5)	902	607♀, 295♂
PISA 6 (P6)	1044	442♀, 602♂
TIMSS & PIRLS 1 (TP1)	990	611♀, 379♂
TIMSS & PIRLS 2 (TP2)	1294	495♀, 799♂
TIMSS & PIRLS 3 (TP3)	721	398♀, 323♂
TIMSS & PIRLS 4 (TP4)	1536	719♀, 817♂

Taulukko 10: Klusteroinnin tuloksena muodostuneiden klustereiden koot ja sukupuolijakaumat.

Taulukossa 11 olen kuvannut muuttujien jakautumisen klustereiden kesken. Klusterit on listattu suurimmasta pienimpään jokaisen klusterointimuuttujan mukaan.

Klusterointimuuttuja (PISA/TIMSS & PIRLS)	Suuruusjärjestys PISA-klustereissa	Suuruusjärjestys TIMSS & PIRLS -klustereissa
INTMAT/ASBGSLM	P6, P4, P2, P1, P5, P3	TP2, TP4, TP1, TP3
SCMAT/ASBGSCM	P4, P6, P2, P1, P3, P5	TP2, TP4, TP1, TP3
BELONG/ASBGSBS	P4, P6, P1, P3, P2, P5	TP1, TP2, TP3, TP4
HEDRES/ASBGHRL	P1, P6, P4, P5, P3, P2	TP2, TP1, TP4, TP3

Taulukko 11: Klusterointimuuttujien jakautuminen klustereiden kesken.

Katsoen taulukkoa 11 ja liitteinä olevia kaavioita (liitteet **C** ja **D**) voidaan määrittää jokaista klusteria karakterisoivat muuttujat.

6.1.1 PISA klustereiden karakterisoivat muuttujat

Taulukkoon 12 olen kuvannut PISA-klustereiden karakterisoivat muuttujat.

Klusterin koodi	Klusterin karakterisoivat muuttujat (+ korkea, - matala)
P1	HEDRES+
P2	HEDRES-
P3	INTMAT-, SCMAT-
P4	SCMAT+, BELONG+
P5	INTMA-, SCMAT-, BELONG-
P6	INTMAT+, HEDRES+, SCAMT+

Taulukko 12: PISA-aineistosta muodostettujen klustereiden karakterisoivat muuttujat.

Klusteri 1 (P1). Tämän klusterin voimakkain karakterisoiva muuttuja oli HEDRES, eli lapsen kotona olevien oppimista tukevien resurssien määrä. HEDRES:in lisäksi muut muuttujat eivät eronneet merkittävästi muista klustereista. Klusterin prototyyppioppilaan kotona on huomattavasti muita enemmän oppimista tukevia resursseja, mutta hän ei ilmoita olevansa kiinnostunut matematiikasta ja hänen itseluottamuksensa omiin matemaattisiin kykyihinsä on keskivertoa.

Klusteri 2 (P2). PISA-aineiston toista klusteria kuvasi kaikista alhaisin HEDRES-muuttuja. Tämän muuttujan lisäksi muut muuttujat eivät eronneet kriittisesti muista klustereista. Klusterin prototyyppioppilas tulee kodista, jossa on huomattavasti heikommin oppilaan oppimista tukevia resursseja. Tämän lisäksi hän ei ilmoita olevansa kiinnostunut matematiikasta, luottaa matemaattisiin kykyihinsä kohtalaisesti ja tuntee kuuluvansa kouluun hieman heikommin kuin muut oppilaat.

Klusteri 3 (P3). Kolmatta PISA-klusteria voimakkaimmin karakterisoivat muuttujat olivat alhaisin INTMAT-muuttuja, eli kiinnostus matematiikkaan ja toiseksi alhaisin SCMAT, eli oppilaan itseluottamus tämän matemaattiseen kyvykkyyteen. Klusterin prototyyppioppilas

ei siis ilmoita olevansa kiinnostunut matematiikasta, eikä hän myöskään luota omiin matemaattisiin kykyihinsä. Oppilas kuitenkin tuntee kuuluvansa kouluun siinä missä muutkin ja tämän kotona on kohtalaisesti tämän oppimista tukevia resursseja.

Klusteri 4 (P4). Klusteria karakterisoi selvästi muita korkeammat SCMAT- ja BELONG-muuttujat. Prototyyppioppilas on huomattavasti muita oppilaita luottavaisempi omiin matemaattisiin kykyihinsä ja tuntee myös kuuluvansa kouluun muita lapsia vahvemmin. Korkeasta itseluottamuksestaan huolimatta oppilas ei ilmoita olevansa kovinkaan kiinnostunut matematiikasta. Oppilaan kotona on myös keskimääräistä enemmän tämän oppimista tukevia resursseja.

Klusteri 5 (P5). Klusterin prototyyppioppilas suhtautuu kaikista negatiivisimmin koulunkäyntiin yleensä. Tästä viestii toiseksi alhaisin INTMAT-muuttuja, sekä toiseksi alhaisimmat SCMAT- ja BELONG-muuttujat. Klusterin prototyyppioppilas ei siis ilmoita olevansa kiinnostunut matematiikasta ja hän luottaa erittäin heikosti omiin matemaattisiin kykyihinsä. Oppilas ei myöskään tunne kuuluvansa kouluun kovinkaan vahvasti, mutta tämän kodissa on jonkin verran hänen oppimistaan tukevia resursseja.

Klusteri 6 (P6). PISA-aineiston kuudennetta klusteria karakterisoivat parhaiten korkein INTMAT-muuttuja ja toiseksi korkeimmat HEDRES- ja SCMAT-muuttujat. HEDRES oli toiseksi korkein, mutta BELONG ei eronnut huomattavasti muista klustereista. Klusterin prototyyppioppilas ilmoitti olevansa kiinnostunut matematiikasta ja hän myös luotti keskivertoa paremmin omiin matemaattisiin kykyihinsä. Oppilas tuntee myös kuuluvansa kouluun kohtalaisen hyvin ja tämän kotona on paljon tämän oppimista tukevia resursseja.

6.1.2 TIMSS ja PIRLS -klustereiden karakterisoivat muuttujat.

Taulukkoon 13 olen kuvannut TIMSS ja PIRLS -klustereiden karakterisoivat muuttujat.

Klusterin koodi	Klusterin karakterisoivat muuttujat (+ korkea, - matala)
TP1	ASGSBS+, ASBGSLM-
TP2	ASBGSLM+, ASBGSCM+, ASBGSBS+
TP3	ASBGSLM-, ASBGSCM-, ASBGSBS-
TP4	Ei karakterisoivia muuttujia.

Taulukko 13: Yhdistetystä TIMSS ja PIRLS -aineistosta muodostettujen klustereiden karakterisoivat muuttujat.

Yhdistetyn TIMSS ja PIRLS -aineiston klustereissa oppilaan kotona olevien resurssien määrä ei eronnut kriittisesti muodostettujen klusteriprototyyppien välillä, joten en mainitse muuttujaa ASBGHRL eritellessäni klustereita karakterisoivia muuttujia tai kuvaillessani klusterien prototyyppioppilaita.

Klusteri 1 (TP1). Klusteria voimakkaimmin karakterisoiva muuttuja oli ASBGSBS, eli oppilaan kouluun kuuluvuuden tunnetta kuvaava muuttuja. Klusteriprototyypin ASBGSBS-muuttuja oli huomattavasti korkeampi kuin missään muussa klusterissa. Tämän lisäksi klusteria karakterisoi myös toisiksi alhaisin ASBGSLM-muuttuja, joka kuvaa oppilaan kiinnostusta matematiikkaa kohtaan. Klusterin prototyyppioppilas ei tunne olevansa kiusattu, on hieman kiinnostunut matematiikasta ja luottaa omiin kykyihinsä kohtalaisesti.

Klusteri 2 (TP2). Klusterin selvimmät karakterisoivat muuttujat olivat muita selvästi korkeammat ASBGSLM- ja ASBGSCM-muuttuja, joka kuvaa oppilaan itseluottamusta tämän matemaattisiin kykyihin. Näiden lisäksi klusteriprototyypin ASBGSBS-muuttujan arvo oli toiseksi korkein. Prototyyppioppilas ilmoittaa olevansa hyvinkin kiinnostunut matematiikasta ja kertoo luottavansa omiin matemaattisiin kykyihinsä. Oppilas tuntee itsensä myös kiusatuksi, mutta vain erittäin harvoin.

Klusteri 3 (TP3). Kolmannella TIMSS ja PIRLS -klusterilla on ainoat negatiiviset ASBGSLM

ja ASBGSCM arvot. ASBGSBS-muuttuja oli puolestaan toisiksi alhaisin. Klusterin prototyyppioppilas ei ole kertomansa mukaan kiinnostunut matematiikasta lähes ollenkaan, eikä myöskään luota omiin matemaattisiin kykyihinsä. Oppilas tuntee itsensä myös kiusatuksi useammin kuin TP1- ja TP2-prototyyppioppilaat.

Klusteri 4 (TP4). Neljännessä TIMSS ja PIRLS -klusterissa mikään muuttuja ei erottunut muista merkittävästi. Viimeinen TIMSS ja PIRLS -prototyyppioppilas ilmoittaa olevansa vähän kiinnostunut matematiikasta ja uskoo kohtalaisesti omiin matemaattisiin kykyihinsä. Tämän lisäksi oppilas tuntee itsensä kiusatuksi useammin kuin muut prototyyppioppilaat.

6.2 Klusteriprototyyppien peilaaminen metadataan

Yksinään kappaleessa 6.1 kuvaillut klusterit ja klusteriprototyypit eivät tarjoa meille merkittäviä määriä hyödyllistä tietoa. Tästä syystä peilaan seuraavaksi klusteriprototyyppejä niille kuuluviin selittäviin muuttujiin, joita kutsutaan myös metamuuttujiksi. Prosessin tarkoituksena on tutkailla mahdollisia syitä klustereiden muodostumiselle ja yhtäläisyyksiä aineistojen välisten klusteriprototyyppien kesken.

Muodostin ensin klusteroinnin tuloksena saamani indeksitaulukon avulla selittävistä muuttujista ns. metaprototyypit, jotka ovat nähtävissä liitteistä **C** ja **D**. Käytän alla olevassa tekstissä klustereista edellisessä kappaleessa muodostamiani lyhenteitä (katso kappale 6.1, taulukko 10).

Taulukossa 14 kuvataan metamuuttujien jakautuminen klustereiden kesken. Klusterit on listattu suurimmasta pienimpään jokaisen metamuuttujan mukaan.

Selittävä muuttuja (PISA/TIMSS & PIRLS)	Suuruusjärjestys PISA-klustereissa	Suuruusjärjestys TIMSS & PIRLS -klustereissa
ST04Q01/SEX	P4, P6, P2, P1, P3, P5	TP2, TP4, TP3, TP1
AGE/ASDAGE	P5, P4, P1, P2, P6, P3	TP2, TP4, TP1, TP3
TestLang/ITLANG_P	P1, P6, P4, P2, P5, P3	TP1, TP4, TP2, TP3
hisced/ASDHEDUP	P4, P1, P6, P3, P5, P2	TP3, TP2, TP4, TP1
ST57Q01/ASBH08	P6, P1, P4, P5, P2, P3	TP1, TP3, TP4, TP2
MATH	P4, P6, P1, P2, P3, P5	TP2, TP1, TP4, TP3
SCIENCE	P4, P6, P1, P2, P3, P5	TP2, TP1, TP4, TP3
READ	P4, P6, P1, P3, P2, P5	TP1, TP2, TP4, TP3
SchlSize	P5, P3, P4, P6, P2, P1	TP3, TP2, TP4, TP1

Taulukko 14: Selittävien muuttujien jakautuminen klustereiden kesken.

Taulukossa 15 kuvaan lyhyesti tutkimieni aineistojen klusteri- ja metaprototyyppien vertailun avulla löytyneet havainnot.

PISA-aineistosta löytyneet havainnot	Yhdistetystä TIMSS ja PIRLS -aineistota löytyneet havainnot
- Korkeammin koulutetut vanhemmat panostavat enemmän jälkikasvunsa oppimiseen ja hankkivat koteihinsa enemmän oppimista tukevia resursseja. - Oppilaan kotona olevilla oppimista tukevilla resursseilla ei näytä olevan suoraa yhteyttä oppilaan testituloksiin - Oppilaan itseluottamuksella omiin matemaattisiin kykyihinsä on selvä yhteys matemaattiseen suoriutumiseen. - Oppilaan kiinnostuksella matematiikkaa kohtaan ei näytä olevan suoraa yhteyttä matematiikan testituloksiin. - Tytöillä on vahvempi taipumus olla pitämättä matematiikasta ja he näyttäisivät pärjäävän siinä myös heikommin.	- Oppilaan matemaattisella itseluottamuksella ja kiinnostuksella matematiikkaan näyttäisi olevan yhteys oppilaan matemaattiseen suoriutumiseen. - Oppilaan kokemalla kiusaamisella ei ole suoraa yhteyttä tämän testituloksiin.

Taulukko 15: Vuoden 2012 PISA- ja vuoden 2011 yhdistetystä TIMSS ja PIRLS -aineistosta klusteroinnin avulla löytyneet havainnot aineistoittain.

6.2.1 PISA-aineisto

PISA-aineistosta muodostuneiden metaprototyyppien välillä muuttujat ST57Q01 ja Schlsize eivät vaihdelleet kriittisesti. Toisin sanoen oppilaiden käyttämä aika kotitehtävien tekemiseen tai oppilaan koulun koko eivät vaikuttaneet muihin muuttujiin.

Korkeammin koulutetut vanhemmat panostavat enemmän jälkikasvunsa oppimiseen. Tarkastelemalla PISA-aineiston klusteri- ja metaprototyyppejä on havaittavissa, että mitä korkeammin prototyyppioppilaan vanhemmat olivat kouluttautuneet, sitä enemmän tämän kotona oli oppimista tukevia resursseja. Kuitenkin P1-klusterissa HEDRES:in arvo oli huomattavas-

ti korkeampi, vaikka P4 klusterissa vanhempien koulutuksen taso olikin sama. Taustoiltaan P1- ja P4-klustereita erotti selkeiten ruotsinkielisten vastaajien määrä, joita P1-klusterissa oli enemmän. Tämän perusteella voidaan siis tulkita, että vanhempien korkeampi koulutustaso on yhteydessä tämän tarjoamiin oppimista tukeviin resursseihin lapselleen ja ruotsinkielisyydellä voi olla tätä vahvistava vaikutus.

Oppilaan kotona olevilla oppimista tukevilla resursseilla ei näytä olevan suoraa yhteyttä oppilaan testituloksiin. Mikäli vertailemme muuttujien HEDRES, MATH, SCIENCE ja READ arvojen jakautumista klustereiden välillä, voimme huomata, että järjestykset eivät vastaa toisiaan ollenkaan. Joten tulosten perusteella, oppilaan kotona olevien oppimista tukevien resurssien määrällä ole suoraa yhteyttä testituloksiin.

Oppilaan itseluottamuksella omiin matemaattisiin kykyihinsä on selvä yhteys matemaattiseen suoriutumiseen. Vertaamalla klustereiden SCMAT- ja MATH-muuttujia voidaan huomata näiden kahden välillä oleva yhteys. Klustereissa P4 ja P6 ovat ainoat positiiviset SCMAT-arvot ja näiden klustereiden MATH-arvot ovat myös muita korkeampia. Klusterien P1 ja P2 SCMAT ovat lähes yhtä suuret, eikä matematiikan testituloksissa ole suuria eroja. Klustereissa P3 ja P5 SCMAT-arvot ja matematiikan testitulokset olivat alhaisimmat. Klusterilla P4, jolla oli kaikista korkein SCMAT, oli myös kaiken kaikkiaan paremmat testitulokset kuin muilla klustereilla.

Oppilaan kiinnostuksella matematiikkaa kohtaan ei näytä olevan suoraa yhteyttä matemaattiseen suoriutumiseen. Huomattavasti korkein INTMAT-arvo on klusterilla P6. Kuitenkaan tämän klusterin matematiikan testitulokset eivät ole läheskään yhtä korkeat kuin klusterilla P4, jonka INTMAT on huomattavasti alhaisempi. Sama pätee myös alhaisissa testituloksissa. Klusterilla P5 on alhaisimmat matemaattiset testitulokset ja vasta toisiksi alhaisin INTMAT, kun taas klusterilla P3 INTMAT on huomattavasti muita alhaisempi, mutta klusterin matemaattiset testitulokset ovat toiseksi huonoimmat.

Tytöillä on vahvempi taipumus olla pitämättä matematiikasta ja he näyttäisivät pärjäävän siinä myös heikommin. Vertailemalla muuttujien INTMAT ja ST04Q01 jakautumista klustereiden kesken taulukoissa 11 ja 14 voidaan huomata, että klustereissa P5, P3 ja P1 joissa oli eniten tyttöjä, suhtauduttiin myös negatiivisimmin matematiikkaan. Ainoa klusteri, jon-

ka prototyyppioppilas suhtautui matematiikkaan selvästi positiivisesti oli P6, jossa taas oli huomattavasti enemmän poikia.

6.2.2 Yhdistetty TIMSS ja PIRLS -aineisto

Yhdistetystä TIMSS ja PIRLS -aineistosta muodostetut metaprototyypit ovat nähtävissä liitteestä **D**. Muuttujat ASDAGE, ASBH08, ITLANGP ja SchlSize eivät vaihdelleet huomattavasti klustereiden välillä. Tulosten perusteella voidaan siis sanoa, että oppilaan iällä, kotona opiskeluun käytetyllä ajalla, oppilaan äidinkielellä tai oppilaan koulun koolla ei ollut vaikutusta muihin muuttujiin.

Oppilaan matemaattisella itseluottamuksella ja kiinnostuksella matematiikkaa kohtaan näyttäisi olevan yhteys oppilaan matemaattiseen suoriutumiseen. Mikäli tarkastelemme aineiston relevantteja klustereita laskevassa järjestyksessä ASBGSCM ja ASBGSLM-muuttujien perusteella, saamme järjestykseksi TP2, TP1 ja TP3. Metaprototyyppejä katsomalla voidaan huomata varsinkin matematiikan testitulosten jakautuvan samaan järjestykseen. Myös SCIENCE-muuttuja näyttäisi jakautuvan samalla tavalla, mutta muuttujien erojen suuruus ei ole yhtä huomattava kuin MATH-muuttujalla. Tulosten perusteella oppilaan itseluottamuksella tämän matemaattisiin kykyihin ja kiinnostuksella matematiikkaa kohtaan näyttäisi olevan yhteys tämän matematiikan ja luonnontieteiden testituloksiin. Vertailemalla muuttujien ASBGSCM ja ASBGSLM vaihteluvälejä voimme myös huomata, että itseluottamuksen yhteys testituloksiin on kiinnostusta vahvempi.

Oppilaan kokemalla kiusaamisella ei ole suoraa yhteyttä testituloksiin. Aineiston ASBGSBS muuttuja on sitä suurempi, mitä harvemmin oppilas kokee tulleensa kiusatuksi. Vertaamalla muuttujien ASBGSBS, MATH, SCIENCE ja READ jakautumista taulukoista 11 ja 14 huomaamme, ettei useammin tapahtuva kiusaaminen vaikuta suoraan oppilaan testituloksiin.

6.3 Yhtäläisyydet ja eroavaisuudet aineistoista löydettyjen havaintojen välillä

Taulukossa 16 kuvaan lyhyesti tutkimieni aineistojen klusteri- ja metaprototyyppien vertailun avulla löytyneiden havaintojen yhteläisyydet ja eroavaisuudet aineistojen välillä.

Tutkimistani aineistoista löytyneiden havaintojen...	
Yhtäläisyydet	Eroavaisuudet
- Oppilaat, joilla on korkea itseluottamus, pärjäsivät hyvin testeissä	- PISA-aineistossa kodin oppimista tukevat resurssit vaikuttivat oppilaiden testituloksiin enemmän kuin yhdistetyssä TIMSS ja PIRLS -aineistossa
- Kotitehtävien tekemisen määrällä tai koulun koolla ei ollut vaikutusta muihin muuttujiin	- Yhdistetystä TIMSS ja PIRLS -aineistosta erottui klusteri, jolla ei ollut karakterisoivia muuttujia.
- Kummastakin aineistosta erottui huonosti matematiikassa suoriutuneiden ja huonosti koulussa viihtyvien tyttöjen ryhmä	- Yhdistetyssä TIMSS ja PIRLS -aineistossa oppilaat käyttivät keskimäärin enemmän aikaa kotitehtävien tekemiseen
- Kummassakin aineistossa tytöillä oli taipumus suhtautua poikia negatiivisemmin matematiikkaan	- Yhdistetyssä TIMSS ja PIRLS -aineistossa oppilaan kiinnostuksella matematiikkaa kohtaan oli yhteys tämän testituloksiin

Taulukko 16: Vuoden 2012 PISA- ja vuoden 2011 yhdistetystä TIMSS ja PIRLS -aineistosta klusteroinnin avulla löytyneiden havaintojen yhtäläisyydet ja eroavaisuudet.

6.3.1 Yhtäläisyydet

Oppilaat joilla on korkea itseluottamus pärjäsivät vertaisiaan paremmin testeissä. Tarkastelemalla klustereita P4 ja TP2 voidaan havaita, että kummassakin klusterissa on oman aineistonsa korkeimmat arvot itseluottamuksessa omiin matemaattisiin kykyihin ja parhaimmat testitulokset matematiikassa ja luonnontieteissä. PISA-aineistossa ero muiden klustereiden

oppimistuloksiin oli huomattavasti suurempi ja klusterin testitulokset olivat paremmat myös lukemisen saralla. Tulosten pohjalta ei voida kuitenkaan arvioida, onko oppilaan itseluottamus syntynyt tämän vahvoista akateemisista kyvyistä vai onko lapsen luontainen itseluottamus vahvistanut tämän oppimista ja mahdollistanut vertaisiaan paremmat testitulokset.

Kotitehtävien tekemisen määrällä tai koulun koolla ei ollut vaikutusta oppilaiden suoriutumiseen. Kummassakaan aineistossa kotona käytettävä opiskelun määrä tai koulun koko eivät näyttäneet vaikuttavan oppilaan testituloksiin. Nämä kaksi muuttujaa eivät vaihdelleet klustereiden välillä huomattavasti, vaikka selvää vaihtelua testituloksissa kuitenkin oli. Koulun koon merkitys ei sinänsä yllätä, sillä koulun koko ei suoranaisesti vaikuta päivittäisiin, luokkatiloissa tapahtuviin oppimistilanteisiin. Kotona käytetyn opiskeluajan olematon vaikutus on hämmentävää. Kotitehtävien tekemisen vaikutusta akateemiseen suoriutumiseen on kuitenkin tutkittu jo vuosikymmeniä ja monet tutkimukset ovat osoittaneet kotitehtävien tekemisen hyödyllisyyden (Cooper, Robinson ja Patall 2006). Kuitenkin joissakin tutkimuksissa ollaan selvitetty, että esimerkiksi oikein tahditettu kotitehtävien kanssa työskentely tehostaa oppimista, kun taas liian raskas työskentely voi jopa olla haitaksi (Trautwein 2007) & (Trautwein ym. 2002). Voi siis olla, että klustereiden välillä lapset käyttävät kotona aikaansa opiskeluun lähes yhtä paljon, mutta työtavat ja rytmitys tuovat vaihtelua työskentelyn vaikutuksiin. Toisaalta on myös todettu, että suomalaisilla koululaisilla on ylivoimaisesti vähiten aikaa kotitehtävien tekoon yleisesti (Saarela ja Kärkkäinen 2016). Kun kaikkien oppilaiden kotitehtäviin käytetty aika on pieni, ne eivät erotu toisistaan tällaisessa analyysissa.

Matematiikassa huonosti suoriutuneiden ja huonosti koulussa viihtyvien tyttöjen ryhmät. Klustereissa P5 ja TP3 molemmissa korostui tyttöjen suurempi määrä, alhainen itseluottamus matemaattiseen osaamiseen, alhainen kiinnostus matematiikkaa kohtaan, heikko kouluun kuuluvuuden tunne ja heikoimmat matemaattiset testitulokset omissa aineistossaan. P5 klusterissa nämä tekijät erosivat vahvemmin muista PISA-aineiston klustereista, mutta myös TP3 klusterissa nämä ominaisuudet olivat selvästi erotettavissa suhteessa muihin TIMSS ja PIRLS -klustereihin. Ottaen huomioon käytetyn aineiston koon, joudumme pohtimaan ilmiön laajuutta ja vakavuutta suomalaisessa koulumaailmassa. Tämä löydös muistuttaa Saarelan ja Kärkkäisen (2014) PISA-aineistosta löytämiä assosiaatiosääntöjä, joissa löydettiin yhteys sosiaalisesti heikossa asemassa olevien tyttöjen ja alhaisen matemaattisen suoriutu-

misen välillä.

Tytöt suhtautuivat poikia negatiivisemmin matematiikkaan. Kummankin aineiston klustereissa, joissa suhtautuminen matematiikkaan oli negatiivisinta suhteessa aineiston muihin klustereihin, oli myös enemmän tyttöjä suhteessa aineistojen muihin klustereihin. Nämä klusterit olivat: TP1, TP3, P3 ja P5. Myös tämä löydös tukee Saarelan ja Kärkkäisen (2014) tuloksia, joissa todettiin yhteys tyttöjen ja matematiikkaan kohdistuvan negatiivisen ajattelun välillä.

6.3.2 Eroavaisuudet

PISA-aineistossa kodin tarjoamat oppimisresurssit vaikuttivat oppilaiden testituloksiin enemmän kuin yhdistetyssä TIMSS ja PIRLS -aineistossa. Vertailemalla liitteissä **C** ja **D** olevia klusteri- ja metaprototyyppejä voidaan huomata kuinka yhdistetyn TIMSS ja PIRLS -aineiston klusteriprototyyppien välillä muuttuja ASBGHRL ei vaihtele huomattavasti, kun taas PISA-aineiston klusteriprotyyppien välillä muuttuja HEDRES vaihtelee. Voimakkaamman vaihtelun lisäksi PISA-aineistossa ne kolme klusteriprototyyppiä, joilla oli suurimmat HEDRES-arvot, olivat myös kolmen parhaimman joukossa testituloksissa. Vastaavanlaista yhteyttä kotona olevien oppimista tukevien resurssien ja testitulosten välillä ei voitu havaita yhdistetyssä TIMSS ja PIRLS -aineistossa.

Yhdistetyn TIMSS ja PIRLS -aineiston neljäs klusteri. Klusteroinnin tuloksena yhdistetystä TIMSS ja PIRLS -aineistosta koostui yksi klusteri, jota ei karakterisoinut yksikään klusterointimuuttujista. Klusteriprototyypin sisällä muuttujien arvot olivat siis kutakuinkin samat. Tällaista klusteria ei PISA-aineiston klusteroinnin yhteydessä muodostunut.

Yhdistetyssä TIMSS ja PIRLS -aineistossa oppilaat käyttivät enemmän aikaa kotona opiskeluun kuin PISA-aineistossa. Katsoessamme kummankin aineiston metaprototyyppien kuvaajia (liitteet **C** ja **D**) voidaan huomata, että jokaisessa TIMSS ja PIRLS -klusterissa muuttujan ASBH08 arvot ovat kaikki positiivisella puolella, kun taas PISA-klustereissa muuttujan ST57Q01 arvot ovat kaikki kaukana negatiivisella puolella. Tulosten perusteella voidaan siis sanoa, että TIMSS ja PIRLS -aineiston oppilaat käyttivät keskimäärin opiskeluun enemmän aikaa kotona kuin PISA-aineiston oppilaat. Tutkimustulosten avulla ei voida määrittää suoraa syytä tälle havainnolle, mutta yksi selittävä tekijä saattaa olla eri aineistossa olevien op-

pilaiden ikäero. Yhdistettyyn TIMSS ja PIRLS -aineistoon otettiin mukaan vain neljäsluokkalaiset, kun taas PISA-tutkimukseen osallistuvat yläkoulun kahdeksasluokkalaiset. Havainto voi olla myös yhteydessä tutkimusten virallisiin tuloksiin, joissa on huomattu oppilaiden asenteiden matematiikkaa kohtaan muuttuvan negatiivisemmiksi iän karttuessa.

Yhdistetyssä TIMSS ja PIRLS -aineistossa oppilaan kiinnostuksella matematiikkaa kohtaan oli yhteys tämän matematiikan testituloksiin. Liitteestä **D** voidaan huomata, että muuttujalla ASBGSLM näyttäisi olevan yhteys oppilaiden matematiikan testituloksiin. Kuitenkaan samanlaista yhteyttä muuttujan INTMAT ja MATH välillä ei voida havaita PISA-klustereissa.

7 Yhteenveto

Lähdin tutkimuksessani selvittämään, mikäli vuoden 2012 PISA-aineistosta ja vuoden 2011 yhdistetystä TIMSS ja PIRLS -aineistosta voitaisiin löytää samanlaisia oppilasprofiileja, jos molemmat aineistot klusteroitaisiin samalla klusterointialgoritmilla. Mikäli peilaamme tutkimuksessani käyttämiä metodeita ja algoritmia Peña-Ayalan analyysiin (2014), huomaamme tutkimukseni kulkevan vahvasti koulutuksellisen tiedonlouhinnan nykyaikaisessa valtavirrassa. Klusterointi tiedonlouhinnan tehtävänä ja K-means -klusterointimenetelmä ovat Peña-Ayalan analyysin mukaan molemmat hyvin suosittuja viimeaikaisessa koulutuksellisessa tiedonlouhinnassa. Klusterien mallintamiseen käyttämäni prototyyppioppilaiden muodostaminen kuuluu puolestaan oppilaiden mallintamisen piiriin, joka on yksi koulutuksellisen tiedonlouhinnan suosituimmista osa-alueista (Peña-Ayala 2014). Myös käyttämäni K-Means++ -klusterointialgoritmi on ollut suuressa suosiossa tieteen kentällä (Hämäläinen ja Kärkkäinen 2016).

Aloitin tutkimukseni rajaamalla otokseni kattamaan vain käyttämissäni aineistoissa mukana olleet suomalaiset oppilaat, jonka jälkeen otokseni koko oli 13 370 oppilasta. 8829 PISA-aineistosta ja 4541 yhdistetystä TIMSS ja PIRLS -aineistosta. Otoksen valitsemisen jälkeen jatkoin valitsemalla klusteroitavat ja selittävät muuttujat. Kummankin aineiston analysoimisen jälkeen päädyin käyttämään neljää klusterointimuuttujaa ja yhdeksää selittävää muuttujaa. Nämä muuttujat muistuttivat toisiaan aineistojen välillä ja neljä näistä muuttujista oli muodostettu omasta toimestani tätä tutkimusta varten.

Klusteroinnin luotettavuuden varmistamiseksi oli minun täytettävä aineistoissa ollut puuttuva data ja muunnettava data samalle vaihteluvälille. Puuttuvan datan korvaamiseksi käytin hyväkseni kummastakin aineistosta löytyneitä STRATUM-muuttujia. Näiden muuttujien tarkoitus oli ryhmittää vastaajat keskenään mahdollisimman samanlaisiin ryhmiin heidän ominaisuuksiensa ja taustatekijöidensä perusteella. Korvattuani puuttuvan datan oppilaan oman STRATUM-ryhmän keskiarvoilla muunsin kaiken datan samalle vaihteluvälille käyttäen ns. min-max normalisointia.

Kun otanta ja muuttujat olivat valittu ja aineisto valmisteltu, siirryin varsinaisen klusteroinnin suorittamiseen. Koska käyttämäni K-means++ -algoritmi vaati ennen toteuttamistaan etsittyjen klustereiden määrän, oli minun päätettävä tämä määrä. Käytin etsittävien klustereiden määrän arvioimiseksi neljää klusterointi-indeksiä: Davies-Bouldin, Gap, Calinski-Harabasz ja Silhouette. Arvioituani optimaalisen klustereiden määrän Matlabin avulla päädyin käyttämään PISA-aineiston klusterointiin $k = 6$ ja yhdistetyn TIMSS ja PIRLS -aineiston klusterointiin $k = 4$. Jo tässä vaiheessa voitiin huomata, että aineistojen klusteroinnin tulokset eivät muistuta toisiaan täydellisesti.

Kaksi selkeintä yhtäläisyyttä klusteroitujen aineistojen klusteri- ja metaprototyyppien välillä olivat oppilaan itseluottamuksen yhteys tämän oppimistuloksiin ja tyttöjen huono suhde matematiikkaan. Kummankin aineiston testeissä parhaiten menestynyt prototyyppioppilas omasi myös muita selkeästi korkeamman itseluottamuksen omiin matemaattisiin kykyihinsä. PISA-aineiston kohdalla tämä prototyyppioppilas pärjäsi myös muissa aineissa selvästi muita paremmin, kun taas yhdistetyssä TIMSS ja PIRLS -aineistossa ero näkyi selkeiten matematiikassa ja heikommin luonnontieteissä. Tyttöjen negatiivinen suhtautuminen matematiikkaan on noussut esille vuoden 2012 PISA-aineistosta aikaisemminkin, esimerkiksi Kärkkäisen ja Saarelan tutkimuksessa (2014) ja tässä tutkimuksessa tämä ilmiön olemassaolo saa lisää tukea myös yhdistetyn TIMSS ja PIRLS -aineiston puolelta. Kummankin aineiston tyttövaltaisimmissa klustereissa asenteet matematiikkaan olivat negatiivisimmat ja matemaattiset testitulokset heikoimpia. Tämän lisäksi kummankin aineiston heikoiten matematiikassa pärjännyt prototyyppioppilas oli myös kouluun kuuluvuuden tunteessa heikoimpien joukossa.

Mistä tuloksissa näkyvät eroavaisuudet klusteroitujen aineistojen välillä voisivat sitten johtua? Ensinnäkin molempien tutkimusten kohderyhmät ovat erilaiset. Siinä missä PISA-tutkimuksessa tutkittiin kahdeksasluokkalaisia, yhdistetyssä TIMSS ja PIRLS -aineistossa käsiteltiin neljä vuotta nuorempia oppilaita. Voisiko olla, että osa tutkimuksessani paljastuneista ilmiöistä ovat riippuvaisia oppilaiden iästä, eivätkä siksi esiinny kummassakin aineistossa? Osa klusterointitulosten eroavaisuuksista voisi siis selittyä yksinkertaisesti kohderyhmien erilaisuudella. Vikaa voi myös olla toisen, tai kummankin raakadatan tuottaneen tutkimuksen otannassa. Mikäli otannat eivät ole keskenään yhtä kattavia tai tarkkoja, on luonnollista,

että niistä saatavat klusterointituloksetkin eroavat.

Myös tutkimuksissa käytetyt mittarit ovat voineet olla keskenään liian erilaisia, jotta aineistojen klusterointitulokset olisivat voineet olla tarpeeksi samanlaisia. Karkeana esimerkkinä voidaan mainita PISA-aineiston BELONG-muuttuja ja yhdistetyn TIMSS ja PIRLS -aineiston ASBGSBS-muuttuja, jotka mittasivat keskenään hieman eri asioita. BELONG-muuttujassa oppilaalta udeltiin tämän kouluviihtyvyyttä hieman laajemmalla skaalalla, kun taas ASBGSBS-muuttujassa keskityttiin yksinomaan oppilaan tuntemuksiin kiusatuksi joutumisesta. Tietenkin mittareiden väliset erot aineistojen välillä voivat olla huomaamattomampiakin. Esimerkiksi mittareiden teemat voivat olla samat, mutta toisessa tutkimuksessa asiaa lähestytään hieman eri näkökulmasta kuin toisessa.

Osa klusteroinnistani paljastuneista tuloksista kulki käsi kädessä käytettyjen aineistojen virallisten tutkimustulosten kanssa, kun taas osa tuloksista oli vähintään hieman ristiriitaisia. Tämä voi viestiä klusteroinnissa tai sen suunnittelun aikana tapahtuneista virheistä. On mahdollista, että käyttämäni klusterointialgoritmi tai etsittyjen klustereiden määrä ei ollutkaan tarkin mahdollinen, jotta löydettyjen klustereiden paljastamat tulokset olisivat olleet yhtäläisiä tutkimusten virallisten tulosten kanssa. Mikäli tutkimukseni tuloksia verrataan myös muuhun koulutukseen ja oppimiseen keskittyneeseen tutkimukseen, voidaan tulosteni todeta olevan osittain samalla linjalla ja osittain ristiriidassa aikaisempien tutkimustulosten kanssa. Esimerkiksi koulukiusaamisen ja kotitehtävien tekemisen heikko vaikutus oppimistuloksiin ovat aikaisemman tutkimuksen kannalta hämmentäviä löydöksiä.

Juuri tutkielmani valmistumisen alla julkaistu KTL:n (koulutuksen tutkmislaitos) raportti vuoden 2015 TIMSS-tutkimuksen suomalaisista tuloksista vahvistavat aikaisemmissa TIMSS-tutkimuksissa ja omassa tutkimuksessani havaittuja ilmiöitä. Suomalaisten oppilaiden osaamistaso matematiikassa ja luonnontieteissä on edelleen yli kansainvälisen keskiarvon, mutta tulokset ovat selvästi heikentyneet suhteessa vuoden 2011 tutkimukseen. Suomalaiset oppilaat sijoittuvat yhdessä Hongkongin, Taiwanin ja Kazakstanin oppilaiden kanssa jaetulle viidennelle sijalle, mutta matematiikan sijoitus jäi kymmenen parhaimman maan joukon ulkopuolelle. Suomalaisten oppilaiden osaaminen on siis edelleen laskussa, joka on todettu jo vuoden 2011 tutkimuksessa. Suomalaisten oppilaiden asenteet ovat pysyneet matematiikassa ja luonnontieteissä vuoden 2011 tapaan kansainvälisessä häntäpäässä. (Vettenranta ym.

2016)

Oman tutkimukseni tuloksia KTL:n uudessa raportissa tukevat havainnot oppilaiden asenteiden ja itseluottamuksen yhteydestä heidän akateemiseen suoriutumiseensa. KTL:än raportissa mainitaan, kuinka oppilaiden asenteiden yhteys testituloksiin on huomattavasti heikompi kuin itseluottamuksen ja kuinka itseluottamuksen yhteys testituloksiin on huomattavasti vahvempaa matematiikassa kuin luonnontieteissä. Nämä havainnot kulkevat käsi kädessä kappaleessa 6.2.2 kuvaamieni havaintojen kanssa. Eroavaisuutena voidaan raporttia tarkastellessa kuitenkin huomata esimerkiksi koulussa tapahtuvan kiusaamisen selvä yhteys testituloksiin, jota ei omissa tutkimustuloksissani kuitenkaan näkynyt. (Vettenranta ym. 2016)

Vaikka tutkimukseni pääosin mukaili koulutuksellisen tiedonlouhinnan valtavirtaa, ovat siitä saadut tulokset tutkimuksen kentälle uusia. Tutkimuksia, joissa useampaa LSA-aineistoa vertailtaisiin näin koulutuksellisen tiedonlouhinnan tai klusteroinnin avulla, ei olla raportoitu paljoakaan, puhumattakaan juuri suomalaisten oppilaiden tutkimisesta. Mikäli vastaavanlaista tutkimusta jatketaan klusteroimalla PISA-, TIMSS- ja PIRLS-aineistoja keskenään käyttäen klusteroinnissa eri muuttujia ja eri menetelmiä kuin tässä tutkimuksessa, voidaan mahdollisesti löytää lisää kaikissa aineistoissa piilevää tietämystä ja ymmärtää paremmin, miksi tässä tutkimuksessa klusteroinnin tulokset olivat niinkin erilaiset.

Mielestäni olisi myös tärkeää, että tutkimusta suunnattaisiin selvittämään aineistosta löytynyttä ilmiötä, jonka keskiössä ovat heikosti koulussa viihtyvät ja huonosti matematiikassa pärjäävät tytöt. Jo aikaisemmissa tutkimuksissa ollaan havaittu yhteyksiä tyttöjen matematiikkaan kohdistuvan negatiivisen suuntautumisen, heikon matemaattisen suoriutumisen ja heikon sosiaalisen aseman välillä, ja tämä tutkimus vahvistaa näitä löydöksiä kahden aineiston pohjalta (Saarela ja Kärkkäinen 2014). Olisi perin kummallista, että näinkin useassa yhteydessä esiintyneellä ilmiöllä ei olisi mitään totuuspohjaa. Tutkimustulosten perusteella voidaan sanoa, että ilmiö saattaisi saada alkunsa jo alakoulun puolella. Voisi siis olla hedelmällistä tutkia enemmän alakoulun matematiikan opetusta ja oppilaiden välistä kanssakäymistä, jotta saisimme uutta tietoa ilmiön alkuperästä ja pystyisimme myös puuttumaan siihen.

Lähteet

Arthur, David, ja Sergei Vassilvitskii. 2007. “k-means++: The advantages of careful seeding”. Teoksessa *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms,* 1027–1035. Society for Industrial ja Applied Mathematics.

Aurenhammer, Franz. 1991. “Voronoi diagrams—a survey of a fundamental geometric data structure”. *ACM Computing Surveys (CSUR)* 23 (3): 345–405.

Baker, RSJD, ym. 2010. “Data mining for education”. *International encyclopedia of education* 7:112–118.

Baker, Ryan Shaun, Albert T Corbett ja Kenneth R Koedinger. 2004. “Detecting student misuse of intelligent tutoring systems”. Teoksessa *Intelligent tutoring systems,* 531–540. Springer.

Caliński, Tadeusz, ja Jerzy Harabasz. 1974. “A dendrite method for cluster analysis”. *Communications in Statistics-theory and Methods* 3 (1): 1–27.

Cooper, Harris, Jorgianne Civey Robinson ja Erika A Patall. 2006. “Does homework improve academic achievement? A synthesis of research, 1987–2003”. *Review of educational research* 76 (1): 1–62.

Davies, David L, ja Donald W Bouldin. 1979. “A cluster separation measure”. *IEEE transactions on pattern analysis and machine intelligence,* numero 2: 224–227.

Desgraupes, Bernard. 2013. “Clustering indices”. *University of Paris Ouest-Lab Modal’X* 1:34.

D’Mello, Sidney, Andrew Olney ja Natalie Person. 2010. “Mining collaborative patterns in tutorial dialogues”. *JEDM-Journal of Educational Data Mining* 2 (1): 2–37.

Economic Co-operation, Organisation for, ja Development. 2013. *PISA 2012 Results: Ready to Learn (Volume III): Students’ Engagement, Drive and Self-Beliefs.* OECD Publishing.

Economic Co-operation, Organisation for, ja Development. 2014. *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised Edition, February 2014): Student Performance in Mathematics, Reading and Science.* OECD Publishing.

Economic Co-operation, Organisation for, ja Development (OECD). 2013. *PISA 2012 results: what makes schools successful?: resources, policies and practices (volume IV).* OECD, Paris, France.

Fayyad, Usama, Gregory Piatetsky-Shapiro ja Padhraic Smyth. 1996. "The KDD process for extracting useful knowledge from volumes of data". *Communications of the ACM* 39 (11): 27–34.

Filippone, Maurizio, Francesco Camastra, Francesco Masulli ja Stefano Rovetta. 2008. "A survey of kernel and spectral methods for clustering". *Pattern recognition* 41 (1): 176–190.

Foy, Pierre. 2013. "TIMSS and PIRLS 2011 user guide for the fourth grade combined international database". *TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).*

Frey, Andreas, ja Nicki-Nils Seitz. 2011. "Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the programme for international student assessment". *Educational and Psychological Measurement* 71 (3): 503–522.

Hershkovitz, Arnon, ja Rafi Nachmias. 2011. "Online persistence in higher education web-supported courses". *The Internet and Higher Education* 14 (2): 98–106.

Hämäläinen, Joonas, ja Tommi Kärkkäinen. 2016. "Initialization of Big Data Clustering using Distributionally Balanced Folding". *ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning:* 587–592.

IEA. 2014. "TIMSS and PIRLS 2011 Fourth Grade Combined International Database Downloads - Codebooks". Viitattu 17. marraskuuta 2016. `http://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Codebooks.zip`.

———. 2015. "About TIMSS and PIRLS". Viitattu 24. syyskuuta. `timssandpirls.bc.edu/home/pdf/TP_About.pdf`.

IEA. 2016. “The PIRLS 2011 Students Bullied at School Scale”. Viitattu 17. marraskuuta. http://timssandpirls.bc.edu/methods/pdf/P11_R_Scales_SBS.pdf.

Ivančević, Vladimir, Milan Čeliković ja Ivan Luković. 2010. “Analyzing student spatial deployment in a computer laboratory”. Teoksessa *Proceedings of the 4th international conference on educational data mining,* 265–270.

Joncas, Marc, ja Pierre Foy. 2012. “Sample design in TIMSS and PIRLS”. *Methods and procedures. TIMSS and PIRLS International Study Center: Lynch School of Education, Boston College. Available at http://timssandpirls. bc. edu/methods/pdf/TP_Sampling_ Design. pdf.*

Kabakchieva, Dorina, Kamelia Stefanova ja Valentin Kisimov. 2010. “Analyzing university data for determining student profiles and predicting performance”. Teoksessa *Proceedings of the 4th international conference on educational data mining,* 347–348.

Kantardzic, Mehmed. 2011. *Data mining: concepts, models, methods, and algorithms.* John Wiley & Sons.

Kastberg, David, Stephen Roey, David Ferraro, Nita Lemanski ja Ebru Erberber. 2013. “US TIMSS and PIRLS 2011 Technical Report and User’s Guide. NCES 2013-046.” *National Center for Education Statistics.*

“k-means clustering - MATLAB kmeans - MathWorks Nordic”. 2016. Viitattu 18. marraskuuta. https://se.mathworks.com/help/stats/kmeans.html.

Liu, Xiufeng, ja Miguel E Ruiz. 2008. “Using data mining to predict K–12 students’ performance on large-scale assessment items related to energy”. *Journal of Research in Science Teaching* 45 (5): 554–573.

Martin, Michael O, ja Ina VS Mullis. 2012. “Methods and procedures in TIMSS and PIRLS 2011”. *Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.*

Martin, Michael O, Ina VS Mullis, Pierre Foy ja Gabrielle M Stanco. 2012. *TIMSS 2011 International Results in Science.* ERIC.

Michael O. Martin, Ina V.S. Mullis. 2013. “TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning”.

Mullis, Ina VS, Michael O Martin, Pierre Foy ja Alka Arora. 2012. *TIMSS 2011 international results in mathematics.* ERIC.

Mullis, Ina VS, Michael O Martin, Pierre Foy ja Kathleen T Drucker. 2012. *PIRLS 2011 International Results in Reading.* ERIC.

Mullis, Ina VS, Michael O Martin, Graham J Ruddock, Christine Y O’Sullivan ja Corinna Preuschoff. 2009. *TIMSS 2011 Assessment Frameworks.* ERIC.

OECD. 2010. “About PISA - OECD”. Viitattu 24. syyskuuta 2015. `www.oecd.org/pisa/aboutpisa/`.

———. 2014a. “Codebook for PISA 2012 Main Study Parent Questionnaire - MAIN DATABASE”. Viitattu 17. marraskuuta 2016. `https://www.oecd.org/pisa/pisaproducts/PISA12_par_codebook.pdf`.

———. 2014b. “Codebook for PISA 2012 Main Study School Questionnaire - MAIN DATABASE”. Viitattu 17. marraskuuta 2016. `https://www.oecd.org/pisa/pisaproducts/PISA12_sch_codebook.pdf`.

———. 2014c. “PISA 2012 Results in Focus: What 15-year-olds know and what they can do with ehat they know”.

———. 2015. “Codebook for PISA 2012 Main Study Student Questionnaire - MAIN DATABASE”. Viitattu 17. marraskuuta 2016. `https://www.oecd.org/pisa/pisaproducts/PISA12_stu_codebook.pdf`.

OECD, PISA. 2012. *PISA 2009 Technical Report.*

———. 2014. *PISA 2012 Technical Report.*

Olsen, Rolf Vegar. 2005. “Achievement tests from an item perspective: An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students’ knowledge and thinking in science”. Tohtorinväitöskirja, University of Oslo.

Peña-Ayala, Alejandro. 2014. “Educational data mining: A survey and a data mining-based analysis of recent works”. *Expert systems with applications* 41 (4): 1432–1462.

PISA, OECD. 2012. *Results in Focus: What 15-year-olds know and what they can do with what they know*.

Romero, Cristóbal, ja Sebastián Ventura. 2010. “Educational data mining: a review of the state of the art”. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 40 (6): 601–618.

Rousseeuw, Peter J. 1987. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. *Journal of computational and applied mathematics* 20:53–65.

Rutkowski, Leslie, Eugenio Gonzalez, Marc Joncas ja Matthias von Davier. 2010. “International large-scale assessment data issues in secondary analysis and reporting”. *Educational Researcher* 39 (2): 142–151.

Saarela, Mirka, ja Tommi Kärkkäinen. 2014. “Discovering gender-specific knowledge from Finnish basic education using PISA scale indices”. Teoksessa *Proceedings of the 7th International Conference on Educational Data Mining,* 60–68.

———. 2015. “Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data.”: 156–163.

———. 2016. “Knowledge Discovery from the Programme for International Student Assessment”. Luku 8 teoksessa *To appear in Learning analytics: Fundaments, applications, and trends: A view of the current state of the art.* Toimittanut Alejandro Peña-Ayala, 1–39. Springer.

Şen, Baha, Emine Uçar ja Dursun Delen. 2012. “Predicting and analyzing secondary education placement-test scores: A data mining approach”. *Expert Systems with Applications* 39 (10): 9468–9476.

Smyth, Padhraic. 2000. “Data mining: data analysis on a grand scale?” *Statistical methods in medical research* 9 (4): 309–327.

Tibshirani, Robert, Guenther Walther ja Trevor Hastie. 2001. “Estimating the number of clusters in a data set via the gap statistic”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2): 411–423.

Trautwein, Ulrich. 2007. “The homework–achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort”. *Learning and Instruction* 17 (3): 372–388.

Trautwein, Ulrich, Olaf Köller, Bernhard Schmitz ja Jürgen Baumert. 2002. “Do homework assignments enhance achievement? A multilevel analysis in 7th-grade mathematics”. *Contemporary Educational Psychology* 27 (1): 26–50.

Tuononen, Marko. 2005. *Klusterointimenetelmät.* Viitattu 17. marraskuuta 2015. `cs.joensuu.fi/~mtuonon/Klusterointimenetelmat.pdf`.

Wang, Ya-huei, ja Hung-Chang Liao. 2011. “Data mining for adaptive learning in a TESL-based e-learning system”. *Expert Systems with Applications* 38 (6): 6480–6485.

Vettenranta, Jouni, Jenna Hiltunen, Kari Nissinen, Eija Puhakka ja Juhani Rautopuro. 2016. *Lapsuudesta eväät oppimiseen. Neljännen luokan oppilaiden matematiikan ja luonnontieteiden osaaminen.* Koulutuksen tutkimuslaitos.

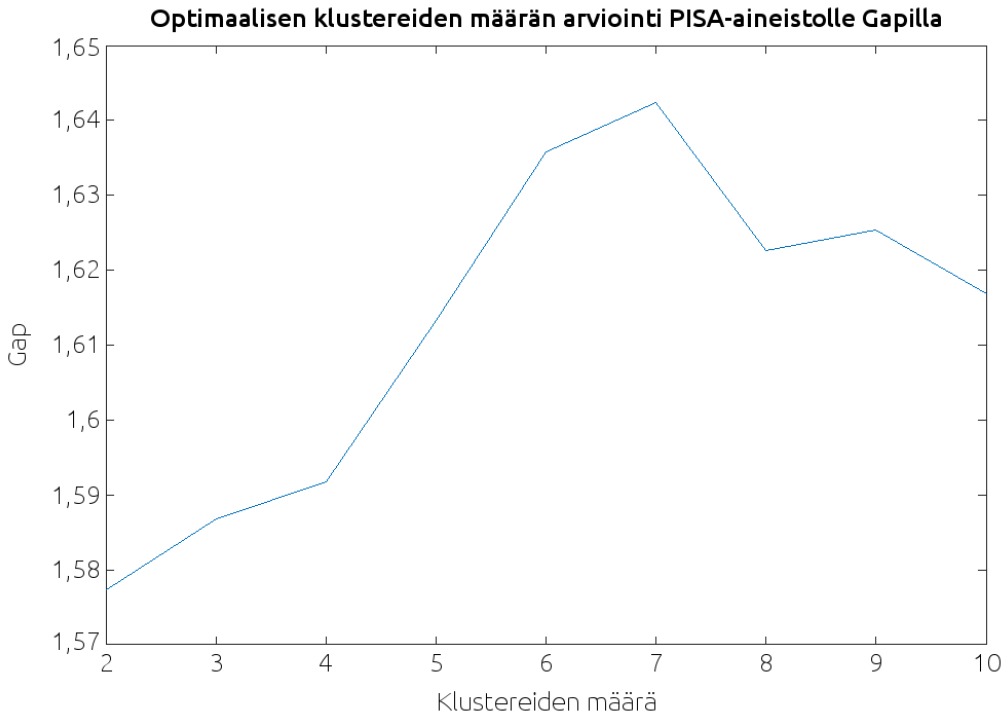
Wu, Xindong, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip ym. 2008. “Top 10 algorithms in data mining”. *Knowledge and information systems* 14 (1): 1–37.

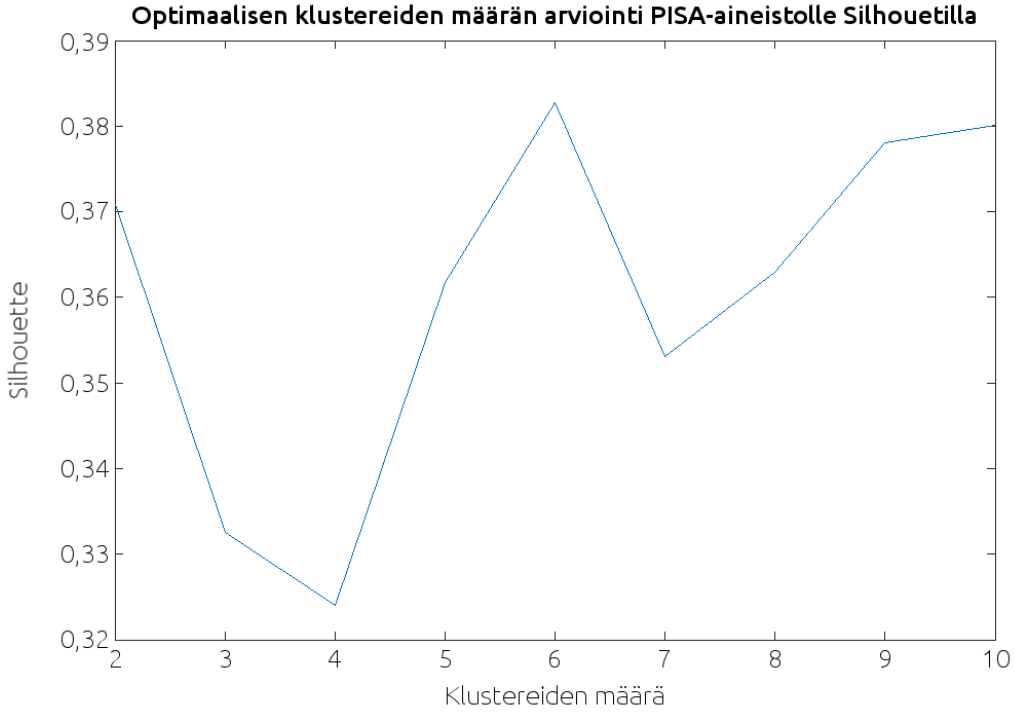
Zaki, Mohammed J, ja Wagner Meira Jr. 2014. *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press.

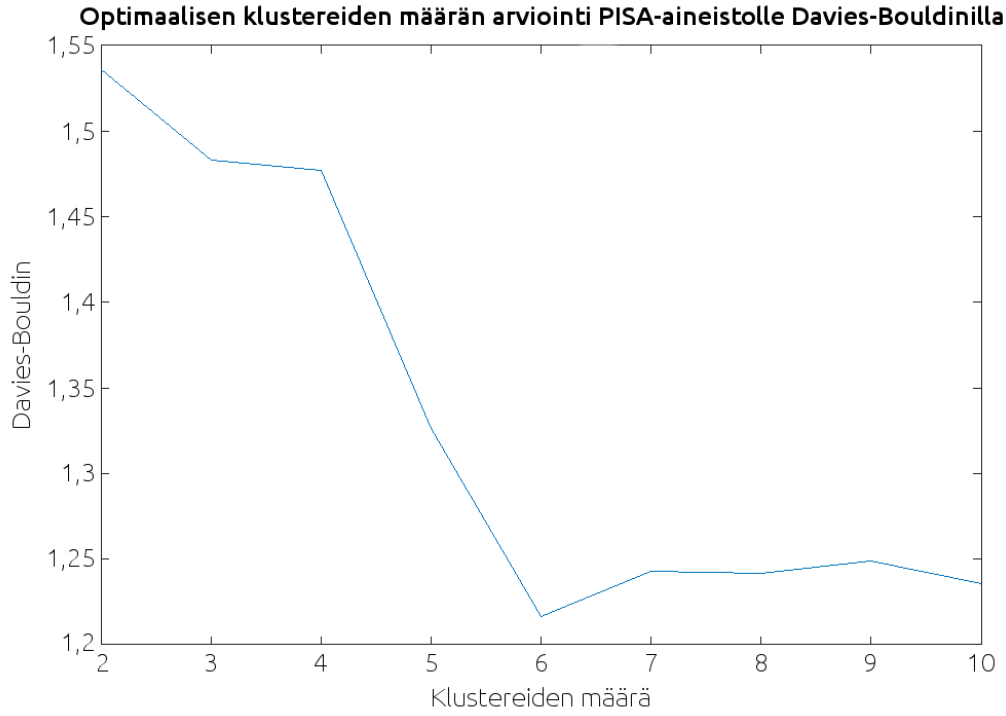
Äyrämö, Sami. 2006. *Knowledge mining using robust clustering*. University of Jyväskylä.

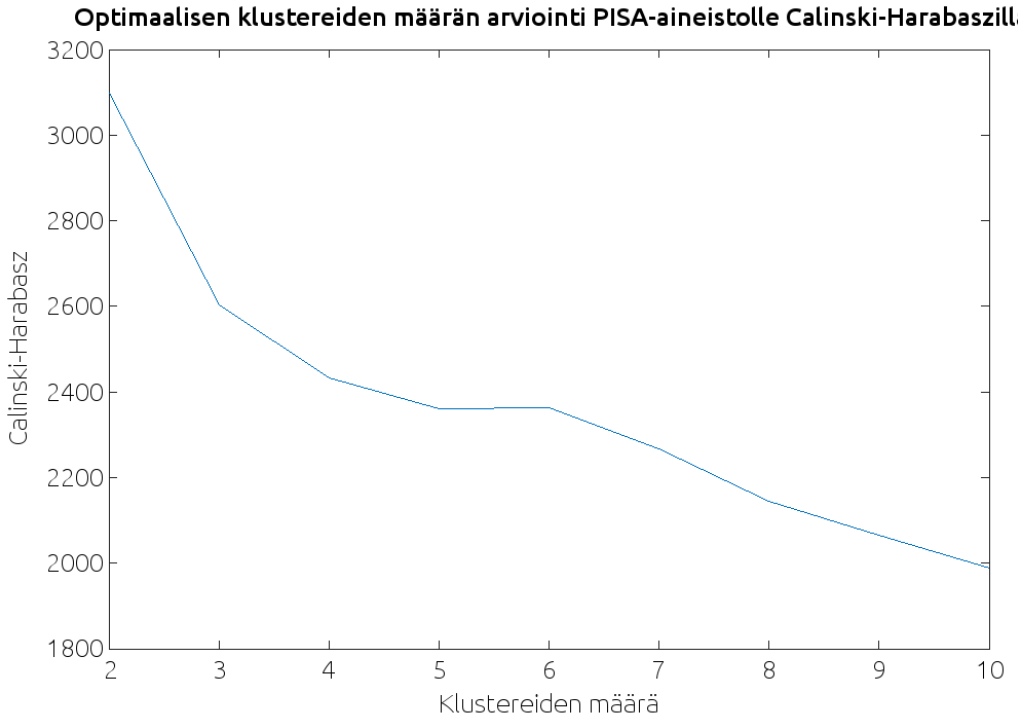
Liitteet

A Klusterien määrän arviointi PISA-aineistossa

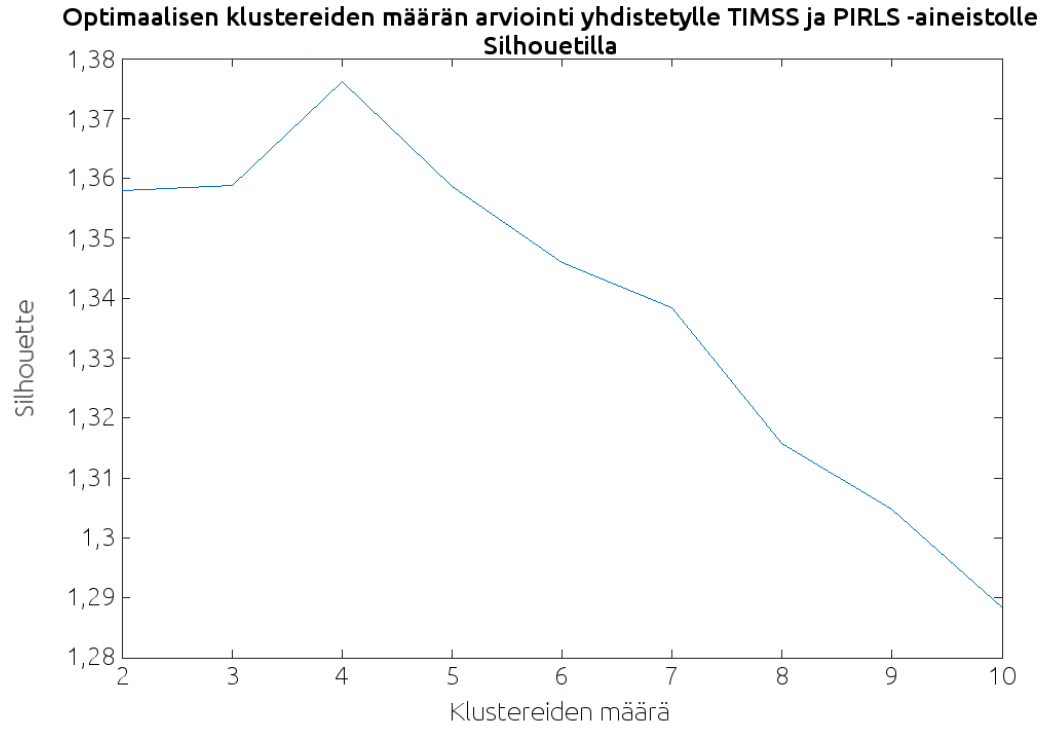


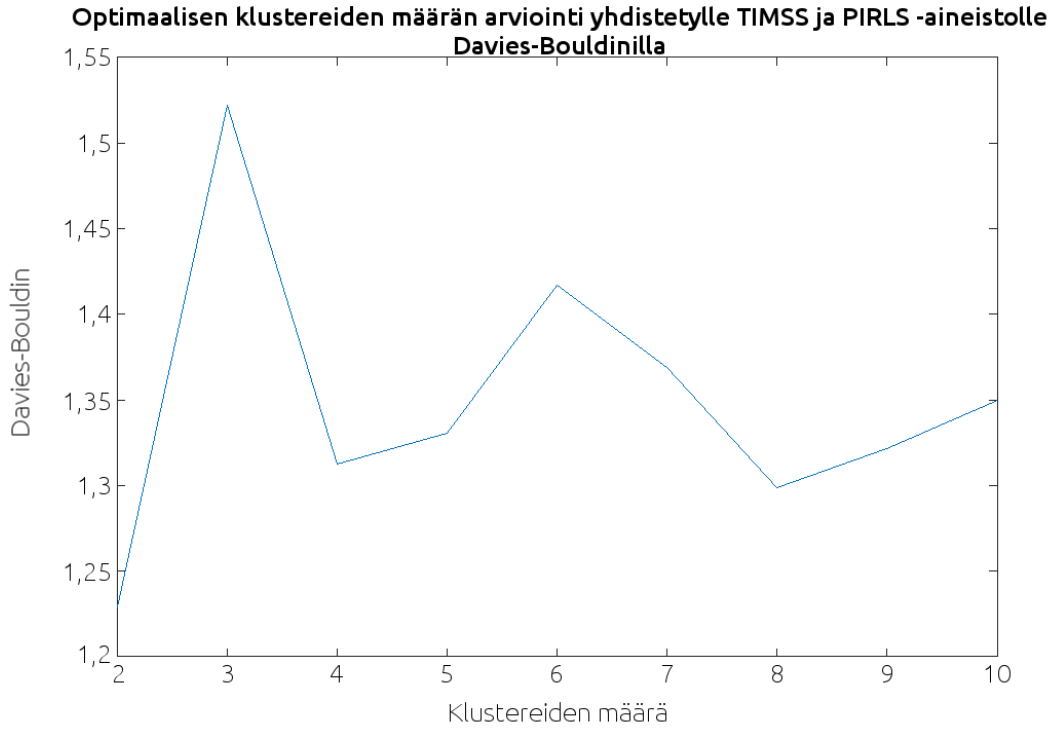


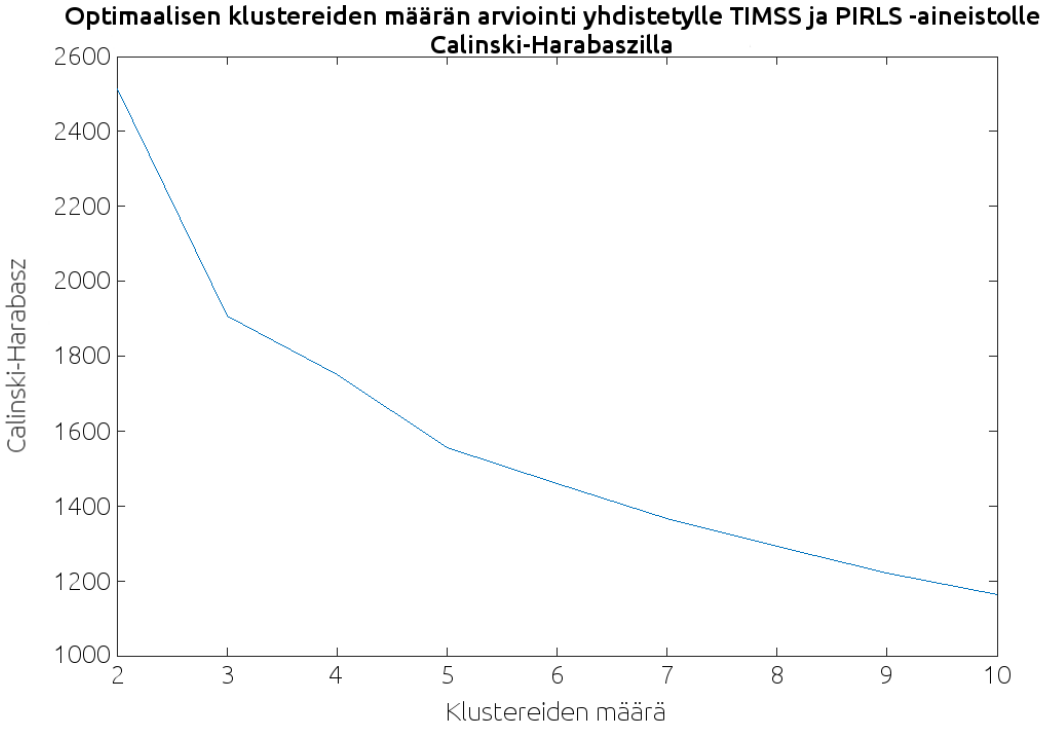

Optimaalisen klustereiden määrän arviointi PISA-aineistolle Davies-Bouldinilla
Davies-Bouldin
1,55
1,5
1,45
1,4
1,35
1,3
1,25
1,2
2
3
4
5
6
7
8
9
10
Klustereiden määrä


Optimaalisen klustereiden määrän arviointi PISA-aineistolle Calinski-Harabaszilla
Calinski-Harabasz
3200
3000
2800
2600
2400
2200
2000
1800
2
3
4
5
6
7
8
9
10
Klustereiden määrä

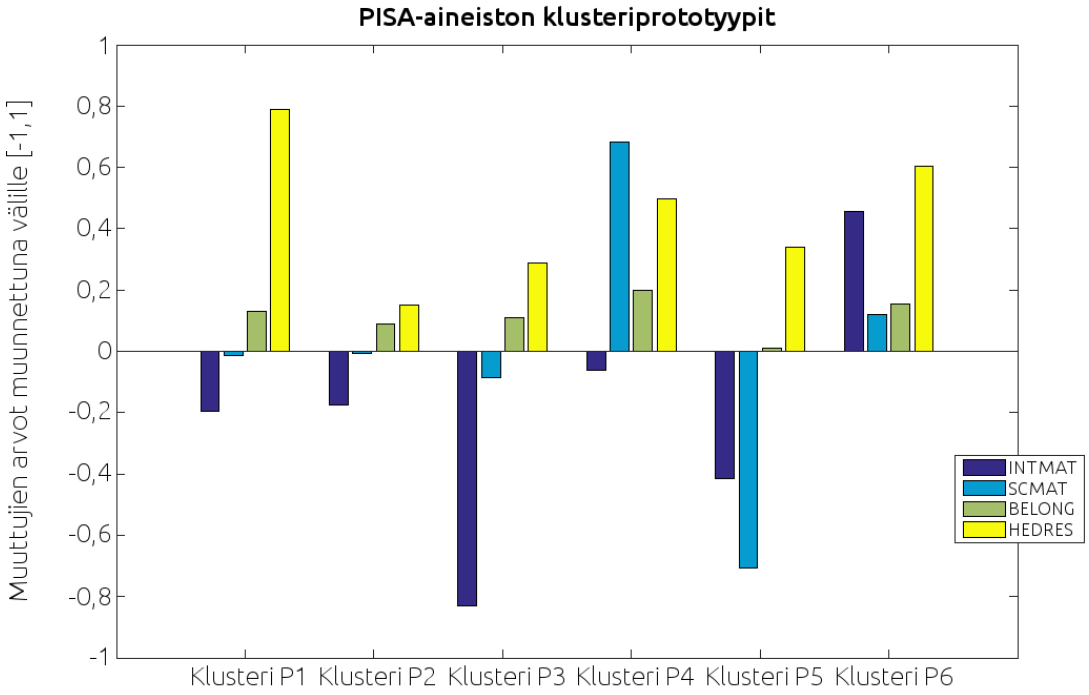
B Klusterien määrän arviointi yhdistetyssä TIMSS ja PIRLS-aineistossa

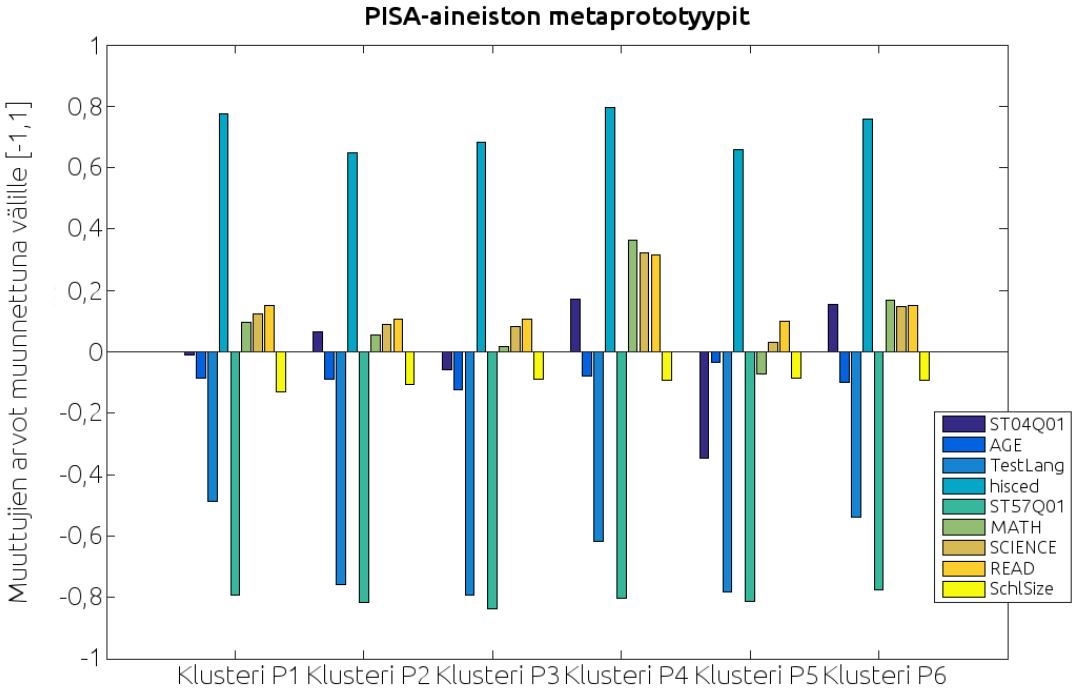



Optimaalisen klustereiden määrän arviointi yhdistetylle TIMSS ja PIRLS -aineistolle
Davies-Bouldinilla
Davies-Bouldin
1,55
1,5
1,45
1,4
1,35
1,3
1,25
1,2
2
3
4
5
6
7
8
9
10
Klustereiden määrä

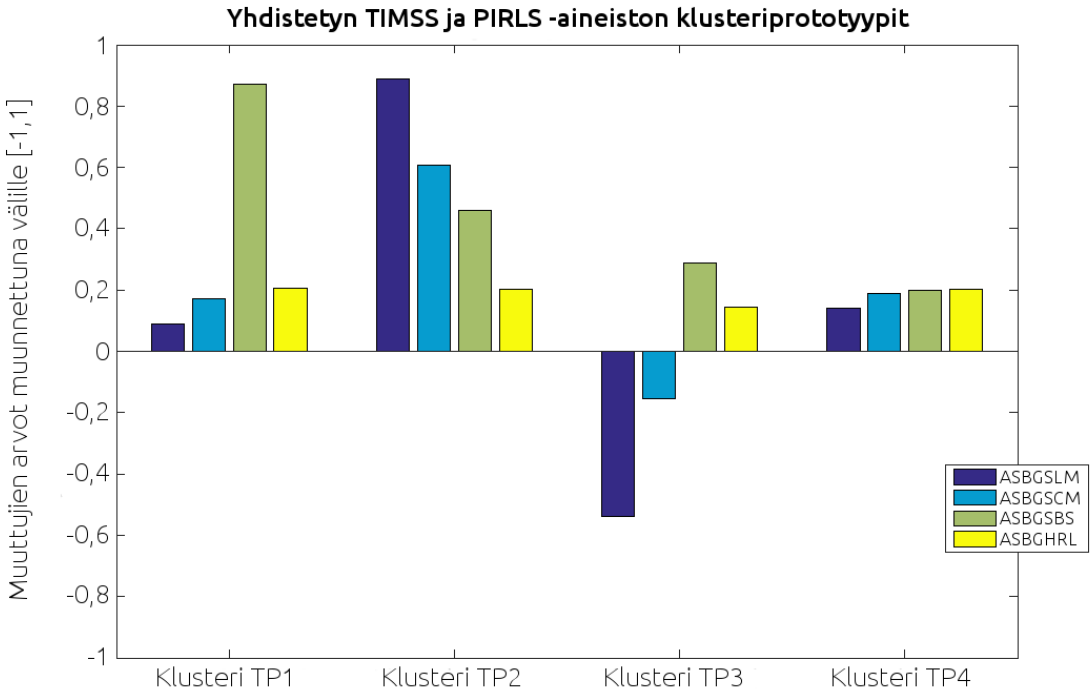

Optimaalisen klustereiden määrän arviointi yhdistetylle TIMSS ja PIRLS -aineistolle
Calinski-Harabaszilla
Calinski-Harabasz
2600
2400
2200
2000
1800
1600
1400
1200
1000
2
3
4
5
6
7
8
9
10
Klustereiden määrä

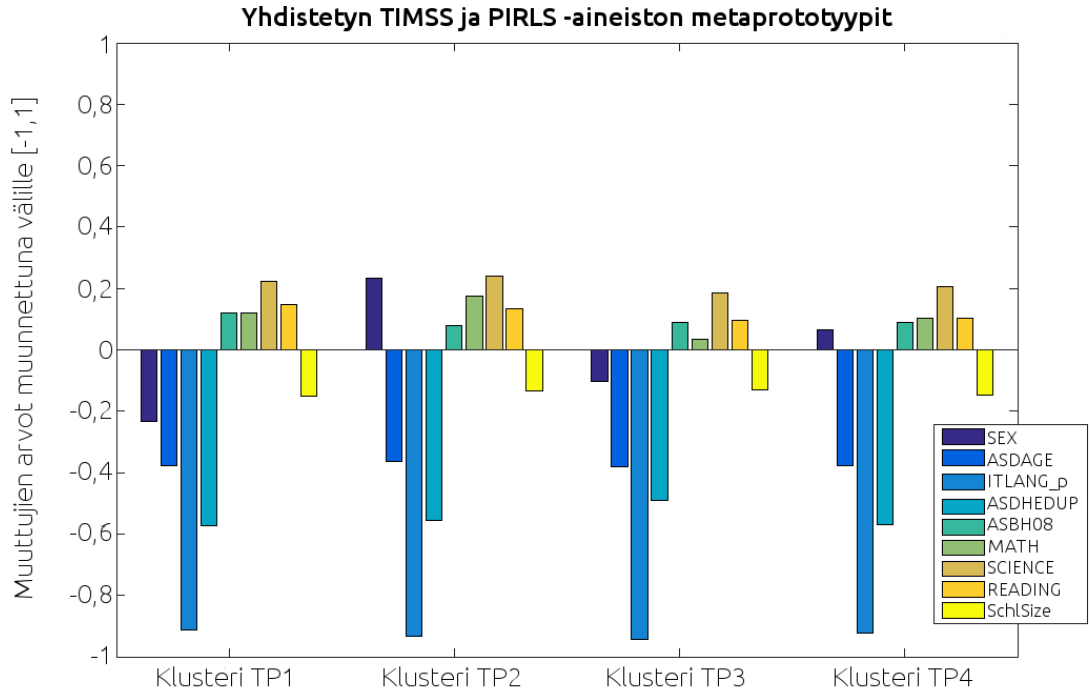
C PISA-aineiston klusteri- ja metaprototyypit





D Yhdistetyn TIMSS ja PIRLS-aineiston klusteri- ja metaprototyypit





E Klusteroidut muuttujat - PISA

Muuttujan koodi	Muuttujan otsikko	Mitä muuttuja mittaa	Mahdolliset arvot alkuperäisessä aineistossa	Mahdolliset arvot käsittelyn jälkeen	Puuttuvien arvojen määrä	Miten muuttujaa käsiteltiin
INTMAT	Mathematics Interest	Oppilaan kiinnostus matematiikkaan	-1,78-9999	-1,78-2,29	3070 (~35%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
SCMAT	Mathematics Self-Concept	Oppilaan käsitys omasta matematiikan osaamisesta	-2,18-9999	-2,18-2,26	3139 (~36%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
BELONG	Sense of Belonging to School	Oppilaan kouluun kuuluvuuden tunne	-3,69-9999	-3,69-2,63	3146 (~36%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
HEDRES	Home educational resources	Kodin oppimista tukevien resurssien määrä	-3,93-9999	-3,93-1,12	122 (~1%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla

F Klusteroidut muuttujat - TIMSS ja PIRLS

Muuttujan koodi	Muuttujan otsikko	Mitä muuttuja mittaa	Mahdolliset arvot alkuperäisessä aineistossa	Mahdolliset arvot käsittelyn jälkeen	Puuttuvien arvojen määrä	Miten muuttujaa käsiteltiin
ASBGSLM	STUDENTS LIKE LEARNING MATHEMATICS/SCL	Oppilaan kiinnostus ja innostus matematiikkaan	4,22728-999999	4,22728-12,30871	31 (~0,7%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
ASBGSCM	STUDENT CONFIDENCE WITH MATHEMATICS/SCL	Oppilaan itseluottamus koskien matematiikan osaamista	3,45971-999999	3,45971-13,67797	45 (~1%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
ASBGSBS	STUDENTS BULLIED AT SCHOOL/SCL	Oppilaan tuntemus kiusatuksi joutumisen yleisyydestä	3,66115-999999	3,66115-13,18339	16 (~0,6%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
ASBGHRL	HOME RESOURCES FOR LEARNING/SCL	Kodin oppimista tukevien resurssien määrä	5,52408-999999	5,52408-15,28708	225 (~5%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla

G selittävät muuttujat - PISA

Muuttujan koodi	Muuttujan otsikko alkuperäisessä aineistossa	Mitä muuttuja mittaa	Mahdolliset arvot alkuperäisessä aineistossa	Mahdolliset arvot käsittelyn jälkeen	Puuttuvien arvojen määrä	Miten muuttujaa käsiteltiin
ST04Q01	Gender	Oppilaan sukupuoli	1 - Nainen 2 - Mies	1 - Nainen 2 - Mies	0 (0%)	-
AGE	Age of student	Oppilaan ikä	15,17-9999	15,17-16,25	0 (0%)	-
TestLang	Language of the test	Oppilaan testissä käyttämä kieli	420 - Suomi 494 - Ruotsi 997 - N/A	1 - Suomi 2 - Ruotsi	27 (~0,3%)	Muutettu vastaamaan TIMSS & PIRLS -aineistoa
hisced	Highest educational level of parents	Oppilaan vanhempien korkein koulutustaso	0 - None 1 - Primary education 2 - Lower secondary 3 - Vocational/pre-vocational upper secondary 4- General upper secondary and/or non-tertiary post-secondary 5 - Vocational tertiary 6 - Theoretically oriented tertiary and post-graduate	0 - None 1 - Primary education 2 - Lower secondary 3 - Vocational/pre-vocational upper secondary 4- General upper secondary and/or non-tertiary post-secondary 5 - Vocational tertiary 6 - Theoretically oriented tertiary and post-graduate	255 (~3%)	-

ST57Q01	Out-of-School Study Time - Homework	Oppilaan käyttämä aika kotitehtävien tekemiseen	0,00-9999	0-30	3186 (36%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
MATH	-	Oppilaan suoriutuminen matematiikan tehtävissä	-	-	0 (0%)	Muodostettiin oppilaan viiden todennäköisen koetuloksen keskiarvosta
SCIENCE	-	Oppilaan suoriutuminen luonnontieteiden tehtävissä	-	-	0 (0%)	Muodostettiin oppilaan viiden todennäköisen koetuloksen keskiarvosta
READ	-	Oppilaan suoriutuminen lukemisen tehtävissä	-	-	0 (0%)	Muodostettiin oppilaan viiden todennäköisen koetuloksen keskiarvosta
SchlSize	-	Oppilaan koulun koko	-	10-946	11 koulua 307 oppilasta (~3,5%)	Yhdistetty muuttujista SC07Q01 ja SC07Q02

H selittävät muuttujat - TIMSS ja PIRLS

Muuttujan koodi	Muuttujan otsikko	Mitä muuttuja mittaa	Mahdolliset arvot alkuperäisessä aineistossa	Mahdolliset arvot käsittelyn jälkeen	Puuttuvien arvojen määrä	Miten muuttujaa käsiteltiin
SEX	SEX OF STUDENTS	Oppilaan sukupuoli	1 - Nainen 2 - Mies 9 - N/A	1 - Nainen 2 - Mies	0 (0%)	-
ASDAGE	STUDENTS AGE	Oppilaan ikä	9,67-99	9,67-13,25	0 (0%)	-
ITLANG_P	LANGUAGE OF TESTING/PIRLS	Oppilaan testissä käyttämä kieli	1-99	1-2	0 (0%)	-
ASDHEDUP	PARENTS' HIGHEST EDUCATION LEVEL	Oppilaan vanhempien korkein koulutustaso	1 - University or higher 2 - Post-secondary but not university 3 - Upper secondary 4 - Lower secondary 5 - Some Primary, lower secondary or no school 6 - Not applicable 99 - omitted or invalid 98 - not admin.	1 - University or higher 2 - Post-secondary but not university 3 - Upper secondary 4 - Lower secondary 5 - Some Primary, lower secondary or no school	346 (~8%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla Luokkaan 6 kuuluneet vastaajat siirretty luokkaan 5

ASBH08	GEN\SCHWORK\TIME SPEND HWORK A DAY	Oppilaan käyttämä aika kotitehtävien tekemiseen	1 My child does not have homework 2 - 15 minutes or less 3 - 16-30 minutes 4 - 31-60 minutes 5 - More than 60 minutes 9 - omitted or invalid 8 - not admin.	1 My child does not have homework 2 - 15 minutes or less 3 - 16-30 minutes 4 - 31-60 minutes 5 - More than 60 minutes	217 (~5%)	Puuttuvat arvot korvattiin oppilaan STRATUM-ryhmän keskiarvolla
MATH	-	Oppilaan suoriutuminen matematiikan tehtävissä	-	-	0 (0%)	Muodostettiin oppilaan viiden todennäköisen koetuloksen keskiarvosta
SCIENCE	-	Oppilaan suoriutuminen luonnontieteiden tehtävissä	-	-	0 (0%)	Muodostettiin oppilaan viiden todennäköisen koetuloksen keskiarvosta
READING	-	Oppilaan suoriutuminen lukemisen tehtävissä	-	-	0 (0%)	Muodostettiin oppilaan viiden todennäköisen koetuloksen keskiarvosta
ACBG01	GEN\TOTAL ENROLLMENT OF STUDENTS	Oppilaan koulun koko	21-99999	21-797	4 koulua 132 oppilasta (~3%)	Puuttuvat arvot korvattiin koulun STRATUM-ryhmän keskiarvolla

I PISA-aineiston STRATUM-ryhmät taulukossa

Ryhmän numero	Ryhmän koko	Ryhmän sukupuolijakauma
PISA 1	2542	1289♀, 1253♂
PISA 2	1227	585♀, 642♂
PISA 3	543	272♀, 271♂
PISA 4	169	84♀, 85♂
PISA 5	489	245♀, 244♂
PISA 6	734	356♀, 378♂
PISA 7	218	115♀, 103♂
PISA 8	137	67♀, 70♂
PISA 9	311	161♀, 150♂
PISA 10	163	73♀, 90♂
PISA 11	106	52♀, 54♂
PISA 12	461	226♀, 235♂
PISA 13	152	65♀, 87♂
PISA 14	75	43♀, 32♂
PISA 15	77	34♀, 43♂
PISA 16	1086	526♀, 560♂
PISA 17	339	177♀, 162♂

J Yhdistetyn TIMSS ja PIRLS -aineiston STRATUM-ryhmät taulukossa

Ryhmän numero	Ryhmän koko	Ryhmän sukupuolijakauma
TIMSS & PIRLS 1	391	191♀, 200♂
TIMSS & PIRLS 2	116	71♀, 45♂
TIMSS & PIRLS 3	505	242♀, 263♂
TIMSS & PIRLS 4	89	50♀, 39♂
TIMSS & PIRLS 5	2153	1056♀, 1097♂
TIMSS & PIRLS 6	131	65♀, 66♂
TIMSS & PIRLS 7	110	52♀, 58♂
TIMSS & PIRLS 8	57	24♀, 33♂
TIMSS & PIRLS 9	864	407♀, 457♂
TIMSS & PIRLS 10	125	65♀, 60♂