



**TURUN
YLIOPISTO**

PIILOMUUTTUAJAN JAKAUMAN VINOUDEN VAIKUTUS
MARGINAALISEN RELIABILITEETTIKERTOIMEN HARHAAN
KAKSIPARAMETRISSESSA LOGISTISESSA JA LUOKITELLUSSA
OSIOVASTEMALLISSA

Toni Rissanen

Pro gradu -tutkielma
Kesäkuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:

Apul.prof. Janne Kujala

FM Timi Niemensivu

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck-järjestelmällä

TURUN YLIOPISTO, Matematiikan ja tilastotieteen laitos

Pro gradu -tutkielma

Pääaine: Tilastotiede

Tekijä: Toni Rissanen

Otsikko: Piilomuuttujan jakauman vinouden vaikutus marginaalisen reliabiliteettikertoimen harhaan kaksiparametrisessa logistisessa ja luokitellussa osiovastemallissa

Ohjaaja: Apul.prof. Janne Kujala

Sivumäärä: 36 sivua + liitteet 7 sivua

Aika: Kesäkuu 2026

Psykologisessa testaamisessa mielenkiinnon kohteena ovat piirteitä tai kykyjä kuvaavat havaitsemattomat piilomuuttujat, joita mitataan epäsuorasti testitehtävien avulla. Piilomuuttujien tärkeä ominaisuus on niiden johdonmukaisuus eli reliabiliteetti.

Osiovasteteoriassa yksi keskeinen tapa tarkastella reliabiliteettia on testin tuottama informaatio, joka vaihtelee piilomuuttujan eri tasoilla. Marginaalinen reliabiliteettikerroin määritellään tämän pohjalta, ja sen estimoinnissa käytetään usein oletusta, että piilomuuttuja noudattaa standardinormaalijakaumaa. Empiiriset testipistejakaumat poikkeavat kuitenkin usein normaalijakaumasta esimerkiksi vinoutensa tai huipukkuutensa vuoksi. Tämän pro gradu -tutkielman tavoitteena on selvittää, miten piilomuuttujan jakauman vinous vaikuttaa marginaalisen reliabiliteettikertoimen estimaattorin harhaan, kun populaation todellinen jakauma on vino.

Tutkielman empiirinen osuus toteutettiin Monte Carlo -simulaationa, jossa vaihdeltiin osiovastemallia, vastauskategorioiden lukumäärää, piilomuuttujan jakauman vinoutta, otoskokoa ja testin pituutta.

Tulokset osoittivat, että estimaattori toimi harhattomasti, kun piilomuuttujan jakauma oli standardinormaali. Piilomuuttujan jakauman ollessa vino estimaattori aliarvioi reliabiliteettia, ja harha kasvoi vinouden kasvaessa. Useampi vastauskategoria ja pidempi testi vähensivät estimaattorin harhaa, kun taas suurempi otoskoko kasvatti sitä.

Asiasanat: reliabiliteetti, osiovasteteoria, marginaalinen reliabiliteettikerroin, Monte Carlo -simulaatio

Sisällys

1	Johdanto	1
2	Testiteoria	4
2.1	Klassinen testiteoria	4
2.2	Osiovasteteoria	5
2.2.1	Osiovasteteorian kehityksen historiaa	6
2.2.2	Osiovasteteoriamallit	8
2.2.3	Kaksi- ja kolmiparametrinen logistinen malli	9
2.2.4	Luokiteltu vastausmalli	10
3	Reliabiliteetti	13
3.1	Reliabiliteetin merkitys testaamisessa	13
3.2	Reliabiliteetti klassisessa testiteoriassa	15
3.3	Reliabiliteetti osiovasteteoriassa	16
3.3.1	Fisher-informaatio reliabiliteetin perustana	16
3.3.2	Marginaalinen reliabiliteetti	18
4	Menetelmät	20
4.1	Osiovastemallit	20
4.2	Piilomuuttujan jakauma	21
4.3	Otoskoot, replikaatioiden määrä ja testin pituus	22
4.4	Marginaalisen reliabiliteettikertoimen laskenta	23
4.5	Uudelleenpoimintaehto	23
4.6	Osioparametrien generointi	24
4.7	Simulaatiotulosten raportointi	25
5	Tulokset	26
5.1	Uudelleenpoiminta, konvergenssi ja Monte Carlo -keskivirhe	26
5.2	Harha vinousparametrin α suhteen	27
5.3	Muiden parametriyhdistelmien vaikutus estimaattorin harhaan	27
5.3.1	Vastauskategorioiden lukumäärän vaikutus	27
5.3.2	Otoskoon vaikutus	28
5.3.3	Testin pituuden vaikutus	28
6	Pohdinta	30
6.1	Estimaattorin harhan epäsymmetria kolmen vastauskategorian luokitellussa vastausmallissa	30
6.2	Otoskoon vaikutus estimaattorin harhaan	32
6.3	Testipituuden vaikutus estimaattorin harhaan	33
7	Yhteenveto	34

Liitteet

A Tulososiossa tarkasteltujen ehtojen harhataulukot

B Simulaatiossa käytetyt kiinteät osioparametrit

C Marginaalisen reliabiliteettikertoimen estimaattorin harha populaatiotasolla

1 Johdanto

Psykologinen testaaminen ja osaamisen arviointi edustavat kognitiivisten ja käyttäytymistieteiden soveltavaa aluetta, jonka tavoitteena on tuottaa luotettavaa ja hyödyllistä tietoa yksilöiden ja ryhmien ominaisuuksista. Testillä tarkoitetaan menetelmää, jossa kerätään otos tutkittavan käyttäytymisestä ja tämä aineisto arvioidaan ja pisteytetään standardoidun menettelyn mukaisesti. Arkikielessä käsitteitä arviointi ja testi käytetään usein toistensa synonyymeinä. Arviointi on kuitenkin testiä laajempi ja moniulotteisempi käsite, joka viittaa tyypillisesti prosessiin, jossa testien avulla tuotettua tietoa yhdistetään muihin tietolähteisiin, kuten haastatteluihin sekä yksilön sosiaalisesta, koulutuksellisesta ja psykologisesta taustasta saatuihin tietoihin.[1]

Psykologiseen testaamiseen liittyy kuitenkin perustavanlaatuinen menetelmällinen haaste, sillä tarkastelun kohteena olevia ominaisuuksia ei voida mitata suoraan. Koska tutkittavia kykyjä tai piirteitä ei voida havaita suoraan, niitä tarkastellaan epäsuorasti mittaamalla muita niihin teoreettisesti tai empiirisesti liittyviä suureita. Esimerkiksi yksilön matemaattista kyvykkyyttä ei voida mitata suoraan, vaan ainoastaan hänen suoriutumistaan joukosta matemaattisia tehtäviä[2]. Tällaisia teoreettisia konstruktioita kutsutaan piilomuuttujiksi (englanniksi *latent variables*). Ne ovat olemassa käsitteellisellä tasolla, ja ilmenevät havaittavina vain välillisesti niihin kytkeytyvien havaittavien suureiden kautta.

Koska psykologiset ja osaamisen arvioinnin testit mittaavat haluttua ominaisuutta epäsuorasti, testistä saatu pistearvo ei kuvaa tutkittavaa ominaisuutta täydellisesti. Tämän seurauksena kaksi erillistä, mutta yhtä pätevää testiä voivat tuottaa eri pisteet samalle henkilölle, vaikka hänen ominaisuutensa ei olisi muuttunut. Tällainen vaihtelu ei liity varsinaiseen mittaustarkoitukseen, vaan voidaan lukea mitausvirheen aiheuttamaksi.[2]

Psykologinen mittaaminen poikkeaa fysikaalisista tieteistä myös kahdella muulla keskeisellä tavalla. Fysikaalisissa tieteissä sama mittaustapa voidaan tyypillisesti toistaa useita kertoja ilman, että toistaminen olennaisesti muuttaa mitattavan ilmiön luonnetta. Psykologisessa testauksessa toistomittaukset ovat sen sijaan käytännössä rajallisia. Jo muutaman toistokerran jälkeen testattavan vastaukset voivat muuttua esimerkiksi väsymyksen sekä oppimis- tai harjoitusvaikutusten seurauksena, jolloin mittaustulokset eivät enää kuvaa alkuperäistä piilomuuttujaa luotettavasti. Lisäksi fysikaalisissa tieteissä pyritään tavallisesti tekemään tilastollisia päätelmiä yhdestä kohteesta kerrallaan, kun taas psykologisessa testauksessa tavoitteena on useimmiten mitata joukkoja samanaikaisesti ja tehdä johtopäätöksiä sekä yksilötasolla että yksilöiden suhteesta koko ryhmään.[2]

Testejä on monenlaisia ja niiden ominaisuudet voivat vaihdella merkittävästi. Testimateriaalien esitystapa voi poiketa toisistaan esimerkiksi sen mukaan, tehdäänkö testi kynällä ja paperilla, suullisesti vai tietokoneavusteisesti. Myös ärsykemateriaali (englanniksi *stimulus material*), standardoinnin aste sekä vastausmuodon tyyppi voivat vaihdella: vastaus voi olla avovastaus tai monivalinta[1]. Näistä eroista huolimatta testit standardoivat prosessin, jolla testiin osallistuvien vastaukset arvioidaan ja pisteytetään.

Kaikki testit eivät kuitenkaan ole lähtökohtaisesti hyödyllisiä tai luotettavia,

vaan tällaisen hyödyn ja luotettavuuden saavuttaminen edellyttää asiantuntemusta testin kehittämis- ja tulkintavaiheissa. Laaja tutkimusnäyttö tukee hyvin suunniteltujen ja asianmukaisesti tulkittujen testien käyttökelpoisuutta: huolellisesti kehitetty ja aiottuun tarkoitukseensa pätevä testi voi tuottaa parempia päätöksentekotuloksia kuin tilanteessa, jossa vastaavaa testiä ei hyödynnetä.[1]. Testien virheellinen käyttö voi puolestaan aiheuttaa huomattavaa haittaa testin tekijöille tai muille testipohjaisten päätösten vaikutuksen alaisena oleville osapuolille. Jotta testit olisivat mahdollisimman tarkkoja ja tarkoituksenmukaisia, tarvitaan formaaleja matemaattisia viitekehyksiä, jotka kykenevät erottamaan piilevän ominaisuuden vaihtelun mittausvirheestä.

Testin tarkkuutta ja tarkoituksenmukaisuutta tarkastellaan ensisijaisesti reliabiliteetin ja validiteetin käsitteiden kautta. Reliabiliteetilla viitataan testipistemäärien johdonmukaisuuteen toistettujen mittausten välillä, joissa toisto voi tarkoittaa saman testin teettämistä uudelleen eri ajankohtana tai rinnakkaista testiversiota. Validiteetilla tarkoitetaan sitä, missä määrin testipisteistä tehdyt tulokset vastaavat tarkoitustaan[1]. Reliabiliteettia voidaan tarkastella numeerisesti reliabiliteettikertoimella, joka saa arvoja nollan ja yhden välillä ja kuvaa testin tuottamien pistemäärien johdonmukaisuuden astetta. Marginaalinen reliabiliteettikerroin (englanniksi *marginal reliability coefficient*, MRC) on eräs reliabiliteettikerroin, joka huomioi mittaustarkkuuden vaihtelun piilomuuttujan eri tasoilla.

Reliabiliteetin ja muiden testaamiseen liittyvien ominaisuuksien tarkasteluun käytetään kolmea keskeistä viitekehystä: klassista testiteoriaa (englanniksi *classical test theory*, CTT), yleistettävyysteoriaa (englanniksi *generalizability theory*) ja osiovasteteoriaa (englanniksi *item response theory*, IRT). Tässä työssä keskitytään osiovasteteoriaan, jonka lisäksi klassista testiteoriaa esitellään pohjana reliabiliteetin käsitteeseen ja tarkasteluun.

Klassinen testiteoria on pitkään ollut vakiintunut viitekehys psykologisessa testaamisessa ja sen pohjalta on kehitetty lukuisia tilastollisia menetelmiä ja tunnuslukuja. Klassisen testiteorian käyttöä rajoittavat kuitenkin muun muassa sen mallien parametrien estimoinnin otosriippuvuus sekä oletus mittausrvirheen yhtäsuuruudesta kaikille vastaajille piilomuuttujan tasosta riippumatta. Osiovasteteoria puolestaan mallintaa testattavan vastustodennäköisyyttä piilomuuttujan funktiona ja mahdollistaa mittaustarkkuuden tarkastelun erikseen kullakin piilomuuttujan tasolla[3]. Osiovasteteoria kattaa useita yksittäisiä malleja. Tässä työssä tarkastellaan kaksiparametrinen logistista mallia (englanniksi *two-parameter logistic model*, 2PL) sekä luokiteltua vastausmallia (englanniksi *graded response model*, GRM), joka on 2PL-mallin laajennus useamman kuin kahden vastauskategorian tilanteisiin.

Vinon piilomuuttujien jakaumien tutkiminen osiovastemallien kontekstissa on perusteltua, koska psykologisten testien tuottamat empiiriset testipistejakaumat ovat harvoin normaaleja. Micceri[4] tutki 440 testin ja kyselyn pistejakaumaa ja totesi niiden kaikkien olevan tilastollisesti merkitsevästi ei-normaaleja käyttäen Kolmogorov-Smirnovin testiä luottamustasolla $\alpha = 0,01$. Ho ja Yu[5] laajensivat Miccerin työtä osaamisen arvioimiseen tutkimalla 504 testipistejakaumaa Yhdysvaltojen osavaltioiden testiohjelmista. Heidän työnsä vahvisti Miccerin havainnon ei-normaalien jakaumien yleisyydestä.

Ho ja Yu esittävät artikkelissaan[5] monta mahdollista tekijää testipistemäärien

jakaumien vinoudelle ja huipukkuudelle. Yhdeksi tekijäksi he esittivät osassa testeissä käytettyä pisteytysmenetelmää, jota kutsutaan testin käänteiseksi karakteristiseksi käyräksi (englanniksi *inverted test characteristic curve*, TCC). Tässä menetelmässä korkeat ja matalat testipistemäärät venytetään pisteasteikon ääripäihin, mikä kasvattaa näiden jakaumien huipukkuutta ja vinoutta, kun äärimmäisten pisteiden saajien suhteellista osuutta kasvatetaan. Toisena tekijänä Ho ja Yu esittävät, että monissa pisteytysmenetelmissä ei ole määritelty suurinta ja pienintä mahdollista testipistemäärää. Tällöin matalimmat ja korkeimmat testipistemäärät valitaan harkinnanvaraisesti tai johdetaan jollain vakiintuneella menetelmällä. Tämä vaikuttaa testipistemäärän jakaumaan erityisesti silloin, kun suuri osuus vastaajista saa äärimmäisiä testipistemääriä. Kolmantena tekijänä Ho ja Yu esittävät testipistemäärien linkittämisen vuosien tai luokka-asteiden välillä.

Tämän työn tavoitteena on selvittää, miten piilomuuttujan jakauman vinous vaikuttaa marginaalisen reliabiliteettikertoimen estimaattorin harhaan osiovastemalleissa, kun estimaattori olettaa standardinormaalijakauman, mutta piilomuuttujan jakauma on vinonormaali. Ho ja Yu[5] toteavat, että vinonormaalijakauman[6] kaltaiset jakaumat mahdollistavat normaaliusoletusten rikkoutumisen seurausten tutkimisen simulaatiotutkimuksissa. Tutkimus toteutetaan Monte Carlo -simulaationa, ja se laajentaa Anderssonin ja Xinin[7] työtä, jossa marginaalisen reliabiliteettikertoimen käyttäytymistä tarkasteltiin suurilla otoskoilla normaalijakaumaoletuksen pätiessä. Tutkielmassa on hyödynnetty Anthropicin avustavia kielimalleja Opus 4.6 ja Opus 4.7 kielenhuollon tukena ja ohjelmointiongelmien ratkaisemisessa. Tutkimuksellinen sisältö, analyysit ja johtopäätökset on tuotettu itsenäisesti.

2 Testiteoria

Testiteoriat muodostavat yleisen viitekehyksen, jossa havaittavat muuttujat, kuten osio- ja testipisteet, yhdistetään piilomuuttujiin, kuten todellisiin pistemääriin ja taitoparametreihin. Teorioihin pohjautuvat mallit määrittelevät käsitteiden väliset suhteet sekä niitä koskevat oletukset. On kuitenkin huomattava, että kaikki mallit ovat väistämättä epätäydellisiä kuvauksia aineistosta. Riittävän suurella testiaineistolla malli voi osoittautua sopimattomaksi, joten ei ole tarkoituksenmukaista kysyä, onko malli oikea vai ei, vaan arvioida, kuinka hyvin se palvelee mittausprosessin tarpeita[8]. Lisäksi Wainer[9] on huomauttanut, että mallin merkitys aineiston tulkinnaassa riippuu aineiston koosta. Pieniä otoksia tarkasteltaessa malli on usein välttämätön luotettavien päätelmien tekemiseen, mutta riittävän suurilla otoksilla mallin lisäarvo aineiston tulkintaan pienenee.

Testiteoriat ja niihin liittyvät mallit ovat tärkeitä mittaamisessa, koska ne tarjoavat viitekehyksen ongelmien tarkasteluun ja teknisten ongelmien ratkaisemiseen, joista yksi tärkeimmistä on mittausvirheiden käsittely. Hyvin valittu testiteoria tai malli voi auttaa ymmärtämään mittausvirheen roolia tutkittavan taitotason arvioinnissa ja miten virheiden vaikutusta tähän voidaan minimoida. Lisäksi testiteoria voi auttaa ymmärtämään muuttujien välistä korrelaatiota ja auttaa raportoimaan testattavan estimoidut testipisteet ja taitotason luottamusväleineen. Eri teoriat ja mallit käsittelevät virheitä eri tavoin. Esimerkiksi jokin malli voi olettaa virheiden olevan normaalisti jakautuneita, kun taas toinen malli ei tee oletusta virheiden jakaumasta. Tapa, jolla virheet on määritetty, vaikuttaa merkittävästi siihen, miten virheet estimoidaan ja raportoidaan.[8]

2.1 Klassinen testiteoria

Klassinen testiteoria muodostuu kolmesta peruskäsitteestä: havaitusta pistemäärästä, todellisesta pistemäärästä ja virhepistemäärästä. Lord ja Novick[2] esittävät klassisen testiteorian mallin seuraavasti. Olkoon X populaatiossa P määritelty satunnaismuuttuja, jonka realisaatio x on satunnaisesti poimitun henkilön havaittu pistemäärä. Olkoon vastaavasti T todellisen pistemäärän satunnaismuuttuja, jonka realisaatio τ on saman henkilön todellinen pistemäärä. Virhepistemäärän satunnaismuuttuja E määritellään havaitun ja todellisen pistemäärän erotuksena $E = X - T$. Klassisen testiteorian mallin vakiintunut esitysmuoto on

$$X = T + E. \tag{1}$$

Lord ja Novick[2] täydentävät tätä klassista testiteoriamallia neljällä perusoletuksella:

$$\mathbb{E}(E) = 0, \tag{KTT.1}$$

$$\rho_{T,E} = 0, \tag{KTT.2}$$

$$\rho(E_1, E_2) = 0, \tag{KTT.3}$$

$$\rho(E_1, T_2) = 0. \tag{KTT.4}$$

Oletuksen KTT.1 mukaan virhepisteiden odotusarvo on nolla kaikissa populaation osajoukossa. Oletuksen KTT.2 mukaan todelliset pistemäärät ja virhepistemäärät

ovat korreloimattomia. Oletukset KTT.3 ja KTT.4 koskevat kahta erillistä mittaus-
ta 1 ja 2. Kahdesta erillisestä mittauksesta saatujen virheiden välinen korrelaatio on
nolla, ja ensimmäisen mittauksen virheen ja toisen mittauksen todellisen pistemää-
rän välinen korrelaatio on nolla. Lisäksi oletetaan, että varianssit ovat äärellisiä ja
positiivisia: $0 < \sigma_X^2 < \infty$, $0 < \sigma_T^2 < \infty$ ja $0 < \sigma_E^2 < \infty$.

Oletuksesta KTT.1 ja yhtälöstä (1) seuraa, että todellisen pistemäärän odotusar-
vo on yhtä suuri kuin havaitun pistemäärän odotusarvo:

$$\mathbb{E}[T] = \mathbb{E}(X - E) = \mathbb{E}(X) - \mathbb{E}(E) = \mathbb{E}(X). \quad (2)$$

Lisäksi koska $\mathbb{E}(E \mid \tau) = 0$ kaikilla τ , pätee

$$\mathbb{E}(X \mid \tau) = \mathbb{E}(T + E \mid \tau) = \mathbb{E}(T \mid \tau) + \mathbb{E}(E \mid \tau) = \tau. \quad (3)$$

Tämä tarkoittaa, että todellinen pistemäärä on hypoteettinen odotusarvo, jonka
ympäriällä havaittu pistemäärä vaihtelee mittauskertojen välillä.[2]

Klassisen testiteorian mallin oletuksista voidaan johtaa havaitun pistemäärän va-
rianssin hajotelma[2]. Koska $X = T + E$, pätee

$$\begin{aligned} \sigma_X^2 &= \sigma^2(T + E) \\ &= \sigma_T^2 + 2\sigma_{T,E} + \sigma_E^2. \end{aligned} \quad (4)$$

Oletuksen (KTT.2) nojalla $\rho_{T,E} = 0$, jolloin kovarianssi $\sigma_{T,E} = \rho_{T,E} \sigma_T \sigma_E = 0$ ja
hajotelma yksinkertaistuu muotoon[2]

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \quad (5)$$

Havaitun pistemäärän varianssi on yhtä suuri kuin todellisen pistemäärän ja vir-
hepistemäärän varianssin summa. Klassisen testiteorian oletukset eivät edellytä ja-
kaumaoletuksia virheille tai todellisille pistemäärille, vaan sen oletukset koskevat
ainoastaan korrelaatorakennetta. Tämä tekee klassisesta testiteoriasta laajasti so-
vellettavan, mutta aiheuttaa myös rajoituksia.

Modernissa kirjallisuudessa todellinen pistemäärä T määritellään havaitun pis-
temäärän X ehdollisena odotusarvona kuvauksen f virittämän sigma-algebran $\sigma(f)$
suhteen

$$T = \mathbb{E}[X \mid \sigma(f)], \quad (6)$$

jossa f on vastaajan kaikkiin mahdollisiin lopputulemiin yhdistävä kuvaus. Tämän ja
edellä esitetyn virhepistemäärän määritelmän seurauksena saadaan tulokset KTT.1
ja KTT.2.[10]

2.2 Osiovasteteoria

Aiemmin esitelty klassinen testiteoria toimii testipistemäärän tasolla: se ei eritte-
le, miten yksittäiset tehtävät ja niiden ominaisuudet vaikuttavat havaittuun piste-
määrään, mikä on sen keskeinen rajoitus. Osiovasteteoria sen sijaan lähtee liikkeelle
tehtävätasolta: se mallintaa suoraan todennäköisyyden, jolla vastaaja tietyllä piilo-
muuttujan tasolla θ antaa tietyn vastauksen tiettyyn tehtävään.

Tehtävätason mallinnuksesta seuraa kolme osiovastemallinnuksen etua. Ensinnäkin mallien osioparametrit, esimerkiksi osion vaikeus tai erottelukyky, eivät riipu siitä millaisesta vastaajien otoksesta ne on estimoitu, kunhan parametrit ilmaistaan samalla skaalalla. Toiseksi samalla skaalalla olevat piilomuuttujan estimaatit eivät riipu siitä, mitkä osiot vastaaja saa vastattavakseen kalibroidusta osiojoukosta. Kolmanneksi osiovastemalleissa mittaustarkkuus voidaan ilmaista jokaisen vastaajan piilomuuttujan tason θ funktiona. Edellytyksenä näiden ominaisuuksien toteutumiselle ovat osiovastemallin sopivuus aineistoon, piilomuuttujan skaalan kiinnitys (esimerkiksi $\theta \sim N(0, 1)$) sekä riittävän suuri osiojoukko ja vastaajaotos estimointiongelmien välttämiseksi. Lisäksi eri otoksissa estimoidut parametrit eivät ole suoraan vertailukelpoisia, vaan vertailu edellyttää skaalojen yhdenmukaistamista.[3]

Osiovasteteoria tarkastelee mittaamista tehtävä- ja testitasolla ja mallintaa tehtävävasteiden yhteyden vastaajan taitotasoon. Sen kolme keskeistä periaatetta voidaan tiivistää seuraavasti:

1. huomion kohteena ovat yksittäisten vastausten todennäköisyydet, ei testin kokonaispistemäärä,
2. vastaukset mallinnetaan satunnaismuuttujina, joiden selittämiseksi tarvitaan todennäköisyysmalleja, sekä
3. malleissa erotetaan toisistaan vastaajan piirrettä kuvaavat parametrit ja testitehtävän ominaisuuksia kuvaavat parametrit.

Tämä parametrien erottelu tekee osiovasteteorin perustavalla tavalla erilaisen lähestymistavan kuin klassisen testiteorian, jossa testipistemäärä esitetään todellisen pistemäärän ja virheen summana ilman vastaajakohtaisten tai tehtäväkohtaisten vaikutusten erittelyä.[11]

Osiovastemallit voivat käsitellä sekä dikotomisias että moniluokkaisia vasteita, ja testit voivat olla yksi- tai moniulotteisia. Osiovastemalleissa oletetaan yleensä yksiulotteinen mitattava taitotaso sekä tehtävän vastekäyrälle tietty funktiomuoto, kuten logistinen funktio.[11]

Osiovasteteorian keskeinen vahvuus on sen tarjoama yhteinen asteikko, jolla sekä tehtävien ominaisuudet että vastaajien taitotasot voidaan ilmaista. Tämä mahdollistaa osioiden, malliparametrien ja piilomuuttujan estimaattien keskinäisen vertailun riippumatta otoksesta tai käytetystä testiversiosta. Lisäksi mallien avulla voidaan määrittää tarkasti, millä piilomuuttujan tason θ välillä yksittäinen tehtävä tuottaa eniten informaatiota, ja siten suunnitella tarkoituksenmukaisia testejä, esimerkiksi adaptiivisia testejä. Osiovasteteorian sovelluksissa hyödynnetään osio- ja testi-informaatiofunktioita, joiden avulla voidaan arvioida mittauksen tarkkuutta eri taitotasoilla.[11]

Seuraavaksi esitellään osiovasteteorian kehityksen historiaa, minkä jälkeen esitellään tässä työssä käytettävien osiovastemallien ominaisuuksia ja reliabiliteettia osiovasteteoriassa tarkemmin.

2.2.1 Osiovasteteorian kehityksen historiaa

Osiovasteteorian ja klassisen testiteorian perustat luotiin 1900-luvun alussa. Charles Spearman esitteli CTT:n keskeiset ideat artikkelissaan *The proof and measurement*

of two things (1904), kun taas Alfred Binet julkaisi vuonna 1905 työnsä *Méthodes nouvelles pour le diagnostic du intellectuel des anormaux*, jota voidaan pitää IRT-ajattelun varhaisena edeltäjänä.[11, 12, 13]

Binet'n artikkeli jäi osin hänen tunnetuksi tekemänsä ensimmäisen standardoidun psykologisen testin varjoon. Spearmanin työ sen sijaan inspiroi aikalaistutkijoita, ja sen pohjalta kehitettiin muun muassa keskivirheen määritelmä, testin luotettavuuden arviointimenetelmät sekä tulokset testin pidentämisen vaikutuksesta luotettavuuteen. Binet'n ajatuksia jatkoi merkittävämmiin vasta 1920-luvulla Louis Thurstone.[11]

Binet sai Pariisin kaupungilta toimeksiannon kehittää testi, jolla voitaisiin erottaa erityisopetusta tarvitsevat oppilaat niistä, jotka olivat vain motivoitumattomia. Kehittäessään tätä testiä Binet ymmärsi, että testi edellytti sellaisen muuttujan mittaamista, jota ei ollut mahdollista havainnoida suoraan. Tällaista piilevää muuttujaa kutsutaan nykytermein piilomuuttujaksi.[11]

Binet kehitti toimeksiantoa varten standardoidun testin, jossa samat oppilaat saivat samat tulokset riippumatta testin toteuttajasta. Hän pyrki myös luomaan asteikon tehtävilleen, mutta havaitsi, ettei älykkyydelle ollut luonnollista mittaa-asteikkoa. Testissään hän antoi tehtäville asteikkoarvon sen perusteella, minkä ikäryhmän (3–11 vuotta) oppilaista noin 75 % ratkaisi tehtävän oikein. Näiden arvojen avulla arvioitiin oppilaiden henkistä ikää. Myöhemmin William Stern ehdotti henkisen iän ja kronologisen iän suhdetta älykkyyden mittariksi, ja Lewis Terman muunsi tämän suhteen älykkyyssosamääräksi kertomalla sen sadalla.[11]

Binet'n merkittävimpiin panoksiin kuuluu ajatus piilomuuttujan skaalauksesta. Hänen menetelmänsä muistutti yllättävän paljon nykyisiä IRT-käytäntöjä: hän asetti yhteisen asteikon sekä tehtäville että oppilaille, estimoi sen arvot empiirisestä datasta ja pisteytti oppilaat näiden perusteella. Lisäksi Binet ehdotti adaptiivisen testauksen periaatetta: kaikki oppilaat eivät välttämättä tarvitsisi samoja tehtäviä, jos tehtävien skaalaukseen voitiin luottaa.[11]

Thurstone kehitti Binet'n ideoita edelleen ja esitteli vuonna 1925 menetelmän, joka poisti tarpeen käyttää ikää piilomuuttujan korvikkeena[14]. Hän oletti piilomuuttujalle tuntemattoman asteikon ja määritteli tehtävien vastekäyrät normaali-jakauman kertymäfunktion avulla. Näin älykkyyden mittaaminen ei enää ollut sidoksissa ikään.[11]

Osiovasteteorian kehitys jatkui 1900-luvun puolivälissä. Frederic Lord esitteli vuonna 1952 kaksiparametrin normaali-jakaumaan pohjautuvan mallin, joka kuvasi oikean vastauksen todennäköisyyttä piilomuuttujan tasolla θ [15]. Georg Rasch puolestaan julkaisi vuonna 1960 mallin, jossa sekä henkilön taitoa että tehtävän vaikeutta käsiteltiin rinnakkaisina parametreina[16]. Rasch korosti mallinsa yksilötason tulkintaa ja loi tilastollisen perustan parametrien estimoinnille sekä mallin sopivuuden arvioinnille.[11]

Alan Birnbaumilla oli merkittävä rooli Lordin mallien kehityksessä. Hän toi 1950-luvulla käyttöön logistisen linkkifunktion, mikä teki mallien käsittelystä matemaattisesti helpompaa. Lisäksi hän ehdotti arvausparametrin lisäämistä, mikä johti kolmiparametrin logistisen (3PL) mallin syntyyn. Lordin ja Novickin teos *Statistical theories of mental test scores* teki tunnetuksi Birnbaumin työn sekä esitteli Fisherin informaation keskeisenä tunnuslukuna osiovasteteoriassa.[2, 11]

Myöhemmät tutkijat, kuten Bock ja Samejima, laajensivat malleja dikotomisista vasteista polytomisiin[17, 18, 19]. Samejiman porrastetuille (englanniksi *graded*) vasteille kehittämä malli on edelleen laajassa käytössä.[11]

Tietokoneiden laskentatehon kasvu 1990-luvulta lähtien mahdollisti IRT:n laajemman soveltamisen ja uusien monimutkaisten mallien kuten moniulotteisten ja hierarkkisten mallien kehittämisen. Myös bayesilaiset menetelmät ovat nousseet tärkeään asemaan parametrien estimoinnissa ja mallien vertailussa.[11]

2.2.2 Osiovasteteoriamallit

Osiovasteteorian lähtökohtana on matemaattinen kuvaus siitä, miten vastaajan piilomuuttujan taso θ vaikuttaa hänen todennäköisyyteensä antaa tietty vastaus osioon. Tätä kuvausta kutsutaan osiovastefunktioksi (englanniksi *item response function*) tai osion ominaiskäyräksi[20].

Tässä työssä esiteltävät osiovastemallit perustuvat kahteen oletukseen, jotka ovat yksiulotteisuus ja paikallinen riippumattomuus.

Yksiulotteisuudella tarkoitetaan sitä, että testin kaikki osiot mittaavat vain yhtä piilomuuttujaa. Tämä on välttämätöntä, että yksiulotteiset osiovastemallit olisivat identtifiituvia. Jos osiot mittaisivat useampaa piilomuuttujaa, niin ei olisi mielekästä estimoida yksittäistä piilomuuttujan tason θ arvoa. Tilanteisiin, joissa mitataan useampaa kuin yhtä piilomuuttujaa samanaikaisesti, on kehitetty moniulotteisia osiovastemalleja[11]. Lord huomauttaa, että yksiulotteisuus ei riipu vastaajajoukon piilomuuttujan jakaumasta vaan on osioiden ominaisuus.[20]

Yksiulotteisuudesta seuraa paikallinen riippumattomuus. Jos vastaajan piilomuuttujan taso θ tunnetaan, se määrittää osion ominaispiirteiden kanssa vastaajan todennäköisyyden mille tahansa yksittäiselle vasteelle kuhunkin osioon. Tieto muiden osioiden vastauksista ei muuta tätä todennäköisyyttä. Täsmällisesti tämä tarkoittaa sitä, että osion i minkä tahansa yksittäisen vasteen todennäköisyys on sama riippumatta siitä tunnetaanko muiden osioiden $j, k, \dots$ vastaukset [20]:

$$P(u_i = v \mid \theta) = P(u_i = v \mid \theta, u_j, u_k, \dots), \quad i \neq j, k, \dots, \quad (7)$$

missä $u_i \in \{0, 1, \dots, m_i\}$ on vastaus osioon i , m_i on osion i korkein vastauskategoria ja v on mikä tahansa kyseisen osion vastauskategoria. Matemaattisesti yhtäpitävä esitys on, että kaikkien osioiden yhteistodennäköisyys hajoaa yksittäisten osioiden todennäköisyyksien tuloksi. Esimerkiksi kolmelle osiolle i, j ja k yhteistodennäköisyys on:

$$P(u_i = v_i, u_j = v_j, u_k = v_k \mid \theta) = P(u_i = v_i \mid \theta) P(u_j = v_j \mid \theta) P(u_k = v_k \mid \theta). \quad (8)$$

Paikallinen riippumattomuus ei kuitenkaan tarkoita sitä, että osiot olisivat korreloimattomia koko populaatiossa. Koska osiot mittaavat samaa piilomuuttujaa, vastaajan piilomuuttujan taso θ määrittää systemaattisesti hänen pistetasonsa kaikissa osioissa: korkea θ -arvo johtaa tyypillisesti korkeisiin pisteisiin ja matala θ -arvo mataliin pisteisiin kaikissa osioissa. Tästä syntyy osioiden välinen korrelaatio koko populaatiossa. Paikallinen riippumattomuus tarkoittaa kuitenkin sitä, että populaatiotasolla esiintyvä korrelaatio selittyy kokonaan yhteisellä piilomuuttujalla. Kun vastaajien piilomuuttujan taso θ asetetaan vakioksi, osioiden pisteet vaihtelevat toisistaan riippumattomasti.[20]

2.2.3 Kaksi- ja kolmiparametrinen logistinen malli

Dikotomisille tehtäville yleinen malli on Birnbaumin esittelemä[2] kolmiparametrinen logistinen malli (3PL), jossa vastauksen todennäköisyys piilomuuttujan tasolla θ on

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}}, \quad (9)$$

jossa a_i , b_i ja c_i ovat osion i parametrit ja $D = 1,7$ on skaalausvakio. Sen käyttö perustuu siihen, että logistinen funktio ja Lordin[15] normaalijakaumaan pohjautuva malli tuottavat tällä valinnalla lähes identtiset arvot kaikilla käytännössä relevanteilla θ -arvoilla[11].

Parametria a_i kutsutaan erottelukykyparametriksi (engl. *discrimination parameter*). Parametri a_i kuvaa osion erottelukykä eli kuinka voimakkaasti osio erittelee vastaajia taitotason mukaan. Mitä suurempi parametrin a_i arvo, sitä jyrkemmin käyrä nousee käännepisteessä ja sitä paremmin osio erottaa eri piilomuuttujan tasoilla θ olevat vastaajat. Kulmakerroin käännepisteessä on verrannollinen parametriin a_i , lausekkeeseen $(1 - c_i)$ ja skaalausvakioon D . [20]

Osiovastefunktion käännepisteellä tarkoitetaan pistettä, jossa sen toisen derivaatan etumerkki vaihtuu[21]. Osiovastefunktion käännepiste sijaitsee pisteessä $\theta = b_i$ [20].

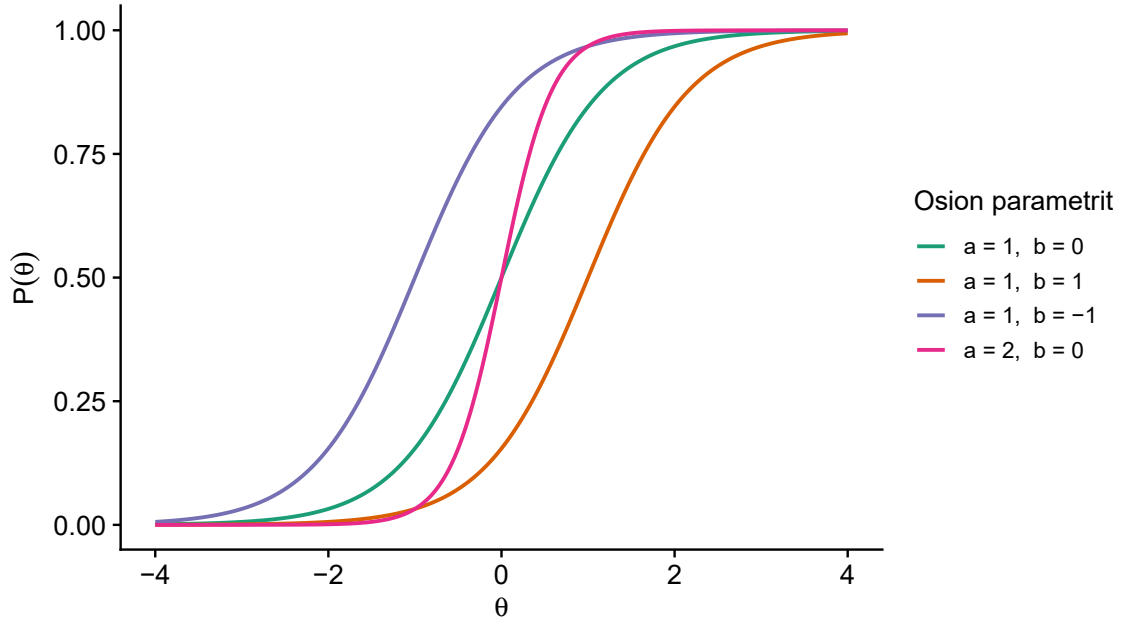
Parametria b_i kutsutaan vaikeusparametriksi (engl. *difficulty parameter*), joka määrittää osiovastefunktion sijainnin piilomuuttujan θ asteikolla. Mitä vaikeampi tehtävä, sitä enemmän oikealla käyrä sijaitsee. Kun osiossa $c_i = 0$, parametri b_i vastaa sitä piilomuuttujan tasoa θ , jolla oikean vastauksen todennäköisyys on $\frac{1}{2}$. Kun $c_i \neq 0$, parametri b_i vastaa piilomuuttujan tasoa θ , jolla todennäköisyys on $\frac{1+c_i}{2}$. [20]

Parametria c_i kutsutaan pseudoarvausparametriksi (englanniksi *pseudo-guessing parameter*). Parametri kuvaa todennäköisyyttä, jolla täysin taitamaton vastaaja eli $\theta \rightarrow -\infty$, vastaa oikein pelkästään arvaamalla. On huomattava, että $c_i > 0$ tekee yhtälön (9) vastefunktiosta epälogistisen, koska logistinen funktio edellyttää asymptootteja 0 ja 1. [11]

Jos $c_i = 0$, kolmiparametrinen malli (9) pelkistyy kaksiparametriseksi logistiseksi malliksi[20]:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}. \quad (10)$$

Kuvassa 1 esitetään 2PL-mallin osiovastefunktioita eri erottelukyky- ja vaikeusparametrien arvoilla. Vaikeusparametrin b vaikutus käyrään näkyy sen vaakasuuntaisena siirtymänä. Suuremmat vaikeusparametrin b arvot siirtävät käyrää oikealle. Erottelukykyparametrin a kasvu sen sijaa jyrkentää käyrää sen käännepisteessä, jolloin osio erottelee vastaajia tarkemmin piilomuuttujan suhteen.



Kuva 1: Kaksiparametrinen logistisen mallin osiovastefunktioita erottelukyky- ja vaikeusparametrien a ja b eri arvoilla.

2.2.4 Luokiteltu vastausmalli

Luokiteltu vastausmalli on Fumiko Samejiman[18] kehittämä osiovastemalli järjestetyille moniluokkaisille vastauksille, kuten Likert-asteikollisille osioille. Se on kaksiparametrinen logistisen mallin yleistys tilanteisiin, joissa vastausvaihtoehdot muodostavat järjestysasteikon[22]. Tällaisia vasteita esiintyy tyypillisesti persoonallisuus- ja asennemittareissa, joissa vastaa arvioi väittämää esimerkiksi asteikolla (*täysin eri mieltä*, $\dots$, *täysin samaa mieltä*).

Merkitään osion i vastauskategorioita $v = 0, 1, \dots, m_i$, jossa m_i on osion korkein vastauskategoria. Kullakin osiolla on yksi erottelukykyparametri $a_i > 0$ ja m_i kynnysparametria $b_{i1} < b_{i2} < \dots < b_{im_i}$, jotka on järjestetty kasvavaan järjestykseen. Malli perustuu kumulatiivisiin vastefunktioihin, joista luokkavastefunktiot johdetaan peräkkäisten funktioiden erotuksena.[22]

Kumulatiivinen vastefunktio $P_{iv}^*(\theta)$ kuvaa todennäköisyyttä, jolla vastaa piilomuuttujan tasolla θ valitsee vastauskategorian v tai sitä korkeamman[22]:

$$P_{iv}^*(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_{iv})}}, \quad v = 1, \dots, m_i, \quad (11)$$

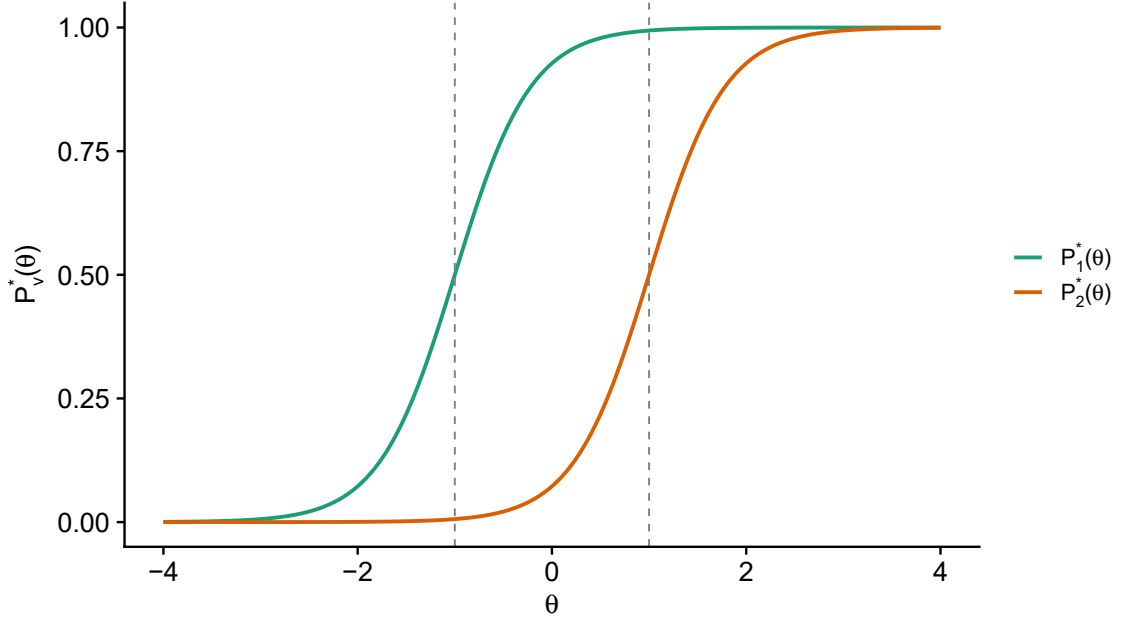
jossa $D = 1,7$ on sama skaalausvakio kuin kaksi- ja kolmiparametrisessä logistisessa mallissa.

Jokaista kynnysparametria b_{iv} kohden saadaan siis yksi 2PL-tyyppinen käyrä. Viidellä vastauskategorialla ($m_i = 5$) saadaan neljä kumulatiivista vastefunktiota $P_{i1}^*(\theta), \dots, P_{i4}^*(\theta)$. Lisäksi mallille on määritelty reunaehdot: todennäköisyys vastata vastauskategoriaan 0 tai sitä korkeampaan on 1 kaikilla θ , ja todennäköisyys vastata vastauskategoriaan $m_i + 1$ tai sitä korkeampaan on 0 kaikilla θ :

$$P_{i0}^*(\theta) = 1, \quad P_{i(m_i+1)}^*(\theta) = 0. \quad (12)$$

Kynnysparametri b_{iv} vastaa sitä piilomuuttujan arvoa θ , jolla todennäköisyys vastata vastauskategoriaan v tai sitä korkeampaan on täsmälleen $\frac{1}{2}$. [22]

Kuvassa 2 esitetään esimerkki GRM-mallin kumulatiivisista vastefunktioista kolmella vastauskategorialla $m_i = 2$. Kummankin kynnysparametrin kohdalla kumulatiivinen todennäköisyys on $\frac{1}{2}$, mikä on esitetty kuvassa katkoviivoin kynnysparametrien b_{iv} kohdalla.



Kuva 2: Luokitellun vastausmallin kumulatiiviset vastefunktiot $P_{iv}^*(\theta)$ kolmella vastauskategorialla, kun $a_i = 1,5$ ja kynnysparametrit ovat $b_{i1} = -1$ ja $b_{i2} = 1$. Katkoviivat osoittavat kynnysparametrien sijainnin.

Luokkavastefunktiot $P_{iv}(\theta)$ kuvaavat todennäköisyyttä vastata osiossa i täsmälleen vastauskategoriaan v piilomuuttujan tasolla θ . Ne saadaan peräkkäisten kumulatiivisten vastefunktioiden $P_{iv}^*(\theta)$ erotuksena [22, 18]:

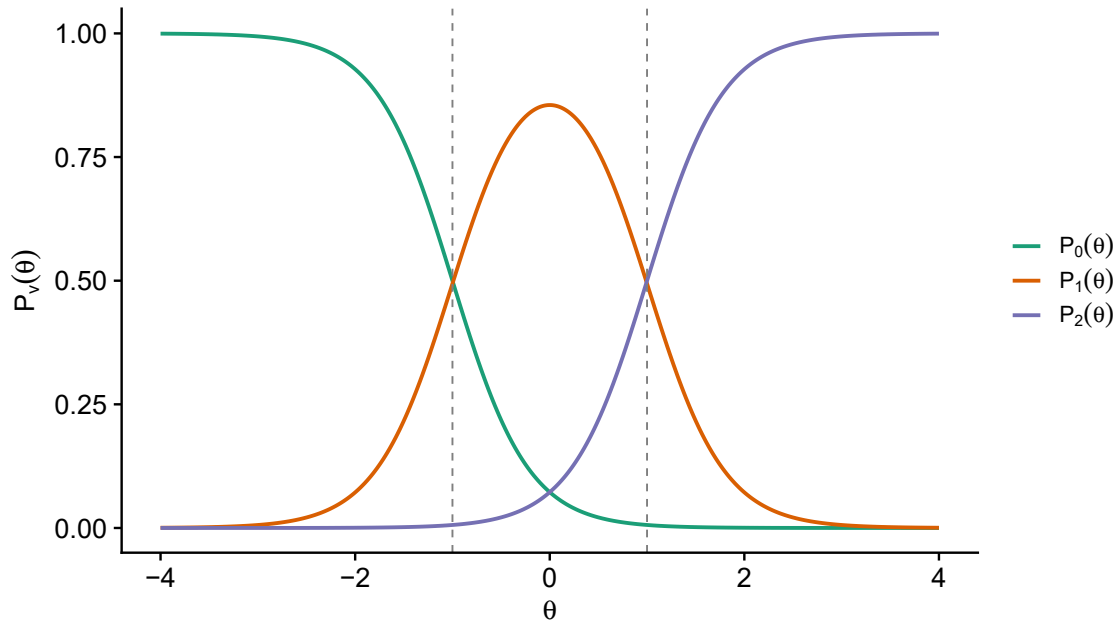
$$P_{iv}(\theta) = P_{iv}^*(\theta) - P_{i(v+1)}^*(\theta), \quad v = 0, 1, \dots, m_i. \quad (13)$$

Luokkavastefunktioilla on kolme tärkeää ominaisuutta: ne ovat ei-negatiivisia ja ne summautuvat kaikilla θ ykköseksi $\sum_{v=0}^{m_i} P_{iv}(\theta) = 1$, jokainen luokkavastefunktio on yksihuippuinen ja kukin vastauskategoria on todennäköisin tietyllä piilomuuttujan tason θ välillä, joka sijaitsee sitä vastaavien kynnysparametrien b_{iv} välissä. [22]

Erottelukykyparametri a_i vaikuttaa kaikkien kumulatiivisten vastefunktioiden jyrkkyyteen yhteisesti. Suuri a_i tuottaa jyrkästi nousevia käyriä, mikä tarkoittaa, että osio erottelee vastaajia piilomuuttujan suhteen tarkemmin. Kynnysparametrit $b_{i1} < \dots < b_{im_i}$ määrittävät kumulatiivisten vastefunktioiden sijainnit piilomuuttujan θ asteikolla ja siten sen, millä tasoilla siirtymät vastauskategoriasta toiseen tapahtuvat. [22]

Kuvassa 3 esitetään GRM-mallin luokkavastefunktiot samoilla osioparametreilla kuin kuvassa 2. Kuvasta huomataan, että kynnysparametrien b_{iv} kohdalla viereisten

vastauskategorioiden $P_{i(v-1)}(\theta)$ ja $P_{iv}(\theta)$ luokkavastefunktiot leikkaavat toisensa ja kun piilomuuttujan arvo θ kasvaa korkeammaksi kuin kynnysparametri b_{iv} todennäköisin vastauskategoria siirtyy kategoriasta $v - 1$ kategoriaan v . Lisäksi kuvasta 3 huomataan, että keskimäinen luokkavastefunktio $P_1(\theta)$ on yksihuippuinen ja se saavuttaa maksimiarvonsa kynnysparametrien välissä.



Kuva 3: GRM-mallin luokkavastefunktiot $P_{iv}(\theta)$ kolmella vastauskategorialla, kun $a_i = 1,5$ ja kynnysparametrit ovat $b_{i1} = -1$ ja $b_{i2} = 1$. Katkoviivat osoittavat kynnysparametrien sijainnin.

3 Reliabiliteetti

Reliabiliteetti on yksi psykologisen mittaamisen keskeisimmistä käsitteistä. Intuitiivisesti reliabiliteetilla tarkoitetaan mittauksen toistettavuutta: luotettava mittaus antaa johdonmukaisia tuloksia samasta ilmiöstä riippumatta siitä, milloin, missä tai kenen toimesta mittaus suoritetaan. Muodollisemmin ilmaistuna reliabiliteetti kuvaa sitä, missä määrin testipistemäärät ovat riippumattomia satunnaisista mittausvirheistä. *Standards for Educational and Psychological Testing* (jatkossa *Standardit*) käyttävät termiä reliabiliteetti/tarkkuus (englanniksi *reliability/precision*) viittamaan yleiseen käsitteeseen pistemäärien johdonmukaisuudesta eri mittauskertojen välillä. Tässä luvussa esitellään ensin reliabiliteettia yleisemmin ja sen jälkeen klassiseen testiteoriaan ja osiovasteteoriaan perustuvaa reliabiliteettia.[1]

3.1 Reliabiliteetin merkitys testaamisessa

Ennen matemaattista tarkastelua on perusteltua hahmottaa, miksi reliabiliteetti on niin keskeinen käsite. *Standardit*[1] toteavat, että reliabiliteetti/tarkkuus on aina tärkeää, mutta sen merkitys kasvaa sitä mukaa kuin päätösten seuraukset ovat vakavampia. Jos esimerkiksi testipistemäärä johtaa vaikeasti peruutettavaan päätökseen, kuten opiskelijan hylkäämiseen tai hyväksymiseen, edellytetään korkeampaa reliabiliteettia kuin tilanteessa, jossa virhe voidaan helposti korjata. Vastaavasti, jos virheellinen päätös voidaan korjata muiden tietolähteiden avulla, vaatimattomampi reliabiliteetti voi riittää.

Testipistemäärien tulkinnat nojaavat oletukseen, jonka mukaan yksilöt ja ryhmät osoittavat tiettyä johdonmukaisuutta pistemäärissään mittauksen toistuessa. Käytännössä tämä oletus ei kuitenkaan täyty täydellisesti. Saman henkilön suoritukset, tuotokset ja vastaukset tehtäviin vaihtelevat sekä testiversiosta että mittauskerrasta toiseen, vaikka olosuhteet pidettäisiin mahdollisimman vakioina. Myös eri arvioijat voivat antaa saman suorituksen perusteella erilaisia pistemääriä. Kaikki nämä vaihtelun lähteet heijastuvat testipistemääriin, jotka vaihtelevat eri mittauskertojen välillä.[1]

Pistemäärien reliabiliteetti/tarkkuus riippuu siitä, kuinka paljon pistemäärät vaihtelevat mittauksen toistuessa, mutta myös siitä, minkälaisen vaihtelun katsotaan kuuluvan mittausprosessiin ja minkälaisen vaihtelun katsotaan olevan virhettä. Tämä ratkaisee, mitä virhelähteitä otetaan reliabiliteettianalyysissä huomioon. *Standardit*[1] erittelevät kolme keskeistä vaihtelun lähdettä:

1. **Vaihtelu mittauskertojen välillä.** Jos testipistemäärän tulkinta edellyttää, että mitattava konstrukti pysyy vakaana eri mittauskertojen välillä, pistemäärävaihtelu mittauskerrasta toiseen luokitellaan mittausvirheeksi. Esimerkiksi jos testi mittaa pysyvää kognitiivista kykyä, pistemäärän muutos lyhyen ajan sisällä kahdella eri mittauskerralla heijastaa mittausvirhettä eikä todellista muutosta mitattavassa kyvyssä.
2. **Vaihtelu rinnakkaisversioiden välillä.** Jos testin eri versioiden tehtävät vaihtelevat ja suorituksia käsitellään otoksena laajemmasta tehtäväjoukosta, versioiden välinen pistemäärävaihtelu luokitellaan mittausvirheeksi. Tällöin

reliabiliteetti arvioi, kuinka hyvin yksittäinen testiversio edustaa koko tehtäväjoukkoa. Rinnakkaisversioista käytetään myös nimitystä rinnakkaistesti (englanniksi *alternate form*).

3. **Vaihtelu arvioijien välillä.** Jos suoritusten pisteyttämiseen käytetään arvioijia, heidän välisensä arvioinnin epäyhdenmukaisuus on lähde mittausvirheelle. Pistemäärät voivat vaihdella sen mukaan, kuka arvioi, vaikka arvioitava suoritus olisi sama.

Yleisesti ilmaistuna kaikki sellaiset pistemääriin heijastuvat muutokset, jotka eivät johdu mitattavan konstruktin todellisesta vaihtelusta, luetaan mittausvirheeksi. Tästä seuraa, että reliabiliteetti ei ole testin ominaisuus, vaan se riippuu siitä, minkälaisia vaihtelun lähteitä otetaan analyysissä huomioon ja minkälaista pistemäärien tulkintaa tavoitellaan. Sama testi voi näyttäytyä eri reliabiliteettiarvoin riippuen siitä, arvioidaanko johdonmukaisuutta yli tehtävien, mittauskertojen vai arvioijien.[1]

Yksinkertaisin tapa arvioida pistemäärien johdonmukaisuutta on tarkastella, kuinka paljon vastaajien pistemäärät vaihtelevat mittauksen toistuessa. Käytännössä testi teetetään kerran ja lyhyen ajan kuluttua, jolloin mitattavan ominaisuuden ei odoteta muuttuvan, kun sama testi tai vastaava rinnakkaisversio teetetään uudelleen. Oletetaan, että ensimmäinen mittauskerta ei vaikuta toiseen. Mittauskertojen ollessa toisistaan riippumattomia, suurempi vaihtelu kertojen välillä viittaa suurempaan mittausvirheeseen ja siten heikompaa reliabiliteettia/tarkkuutta.[1]

Useimmissa testaustilanteissa toistomittaukset eivät kuitenkaan ole mahdollisia, joten yksilötason keskivirhettä ei voida estimoida suoraan. Klassisessa testiteoriassa mittausvirheen suuruus estimoidaan populaatiotasolla ja tätä populaation yli laskettua virheen keskiarvoa kutsutaan mittauksen keskivirheeksi (englanniksi *standard error of measurement*, SEM). SEM kuvaa mittausprosessin tuottamien pistemäärien epäjohdonmukaisuutta populaatiossa. Suuri SEM ilmaisee heikkoa reliabiliteettia/tarkkuutta. Osiovasteteoriassa mittausvirhe sen sijaan ilmaistaan käytännössä aina ehdollisena piilomuuttujan tasoon θ . Tällaista yksittäiselle piilomuuttujan tasolle θ laskettua keskivirhettä kutsutaan ehdolliseksi mittauksen keskivirheeksi (englanniksi *conditional standard error of measurement*, CSEM).[1]

Se, että pistemäärä sisältää mittausvirhettä, edellyttää teoreettisen mittausvirheettömän arvon olemassaoloa. Tilastollisesti todellinen pistemäärä on tuntematon parametri ja havaittu pistemäärä on satunnaismuuttuja, joka vaihtelee todellisen pistemäärän ympärillä[1].

Kolme edellä mainittua 2 testiteoreettista viitekehystä käsittelevät reliabiliteetiongelmia eri tavoin[1]:

1. **Klassinen testiteoria** olettaa yksittäisen virhejakauman vastaajan pistemäärille. Henkilön todellinen pistemäärä on parametri, jonka ympärillä havaittu pistemäärä vaihtelee mittauskertojen välillä. Reliabiliteettia klassisessa testiteoriassa arvioidaan reliabiliteettikertoimilla, joilla tarkastellaan henkilön todellisen pistemäärän varianssin suhdetta havaittujen pistemäärien varianssiin.
2. **Yleistettävyysteoria** tarjoaa laajemman viitekehysten, jossa arvioidaan useiden eri virhelähteiden, esimerkiksi tehtävien, mittauskertojen ja arvioijien, yh-

täaikaista osuutta kokonaisvirheeseen. Henkilön universumipistemäärä määritellään odotusarvona kaikkien mahdollisten mittaustilanteiden yli ja vastaa käsitteellisesti klassisen testiteorian todellista pistemäärää.

3. **Osiovasteteoria** lähestyy reliabiliteetti/tarkkuus-ongelmaa informaatiofunktioiden avulla. Nämä funktiot ilmaisevat, kuinka tarkasti havaitut osiovas-teet antavat tietoa vastaajan piilomuuttujan tasosta θ sen kullakin alueella. IRT-pohjaisten kertoimien laskemisessa hyödynnetään osioinformaatiofunktioita sekä populaation piilomuuttuja jakaumaa.

Käytännössä reliabiliteetti/tarkkuus estimoidaan näistä viitekehyksistä riippuen eri menetelmillä: klassisen testiteorian tai yleistettävyysteorian reliabiliteettikertomilla, tai IRT:n informaatiofunktio pohjaisilla tunnusluvuilla. Tämä riippuu analyysin tavoitteesta ja käytetystä mallista. Kaikille näille menetelmille on yhteistä, että ne saavat suuria arvoja silloin, kun virheeseen liittyvä vaihtelu on pientä suhteessa estimoitavien pistemäärien kokonaisvaihteluun.[1]

Reliabiliteetilla on myös suora yhteys validiteettiin. Koska epäjohdonmukaiset pistemäärät heijastavat satunnaisia mittausvirheitä, ne rajoittavat pistemäärien kykyä ennustaa kriteereitä, tukea diagnostisia johtopäätöksiä ja ohjata päätöksentekoa[1]. Toisin sanoen, jos mittaus on epäluotettava, sen tuloksia ei voida yleistää edes silloin, kun muut validiteetin ehdot täyttyvät. Reliabiliteetti on validiteetin välttämätön, muttei riittävä ehto.

3.2 Reliabiliteetti klassisessa testiteoriassa

Reliabiliteetti klassisessa testiteoriassa rakentuu luvussa 2.1 esitellyn klassisen testimallin $X = T + E$ ja sen neljän perusoletuksen (KTT.1)–(KTT.4) varaan. Reliabiliteetin ρ matemaattinen perusmääritelmä on todellisen vaihtelun osuus kokonaisvaihtelusta:

$$\rho = \frac{\sigma_{\text{tosi}}^2}{\sigma_{\text{tosi}}^2 + \sigma_{\text{virhe}}^2}. \quad (14)$$

Lord ja Novick määrittelevät testin reliabiliteetin havaitun pistemäärän X ja todellisen pistemäärän T välisen korrelaation neliönä $\rho_{X,T}^2$ [2]. Klassisen testiteorian oletuksista (KTT.1) ja (KTT.2) Lord ja Novick osoittavat, että havaitun ja todellisen pistemäärän välinen kovarianssi on $\sigma_{X,T} = \sigma_T^2$, josta seuraa[2]

$$\rho_{X,T}^2 = \frac{\sigma_T^2}{\sigma_X^2}. \quad (15)$$

Koska $\rho_{X,T}^2$ sisältää todellisen pistemäärän T , jota ei voida havaita, sitä ei voida estimoida suoraan aineistosta. Lord ja Novick osoittavat, että kahden rinnakkaisen testin $X = T + E$ ja $X' = T + E'$ havaittujen pistemäärien välinen korrelaatio on

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}, \quad (16)$$

josta huomataan kahden havaittavan muuttujan välisen korrelaatio olevan yhtä suuri kuin testin reliabiliteetti (15)[2]. Tällöin testin reliabiliteetti voidaan estimoida käyttäen rinnakkaisia testejä ilman tietoa havaitsemattomasta todellisesta pistemäärästä T . Testin reliabiliteetti on luku välillä $[0, 1]$, sillä $\sigma_X^2 \geq \sigma_E^2$, kun $\sigma_T^2 \geq 0$.

Virhepistemäärän varianssin σ_E^2 ja mittauksen keskivirheen σ_E Lord ja Novick määrittävät käyttäen rinnakkaisten testien korrelaatiota seuraavasti[2]:

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX'}), \quad (17)$$

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}. \quad (18)$$

Näin ollen klassisessa testiteoriassa reliabiliteettia voidaan estimoida teettämällä sama tai vastaava rinnakkainen testiversio samalle vastaajajoukolle kahdesti ja laskemalla havaittujen pistemäärien välinen Pearsonin korrelaatiokerroin. Toistomittaukset ovat kuitenkin ongelmallisia psykologisessa testauksessa, koska jo muuttaman testikerran jälkeen oppimis- ja harjoitusvaikutus tai väsymys voivat vaikuttaa testattavien vastauksiin[2]. Tämä rajoittaa klassisen testiteorian reliabiliteetin arviointia käytännössä ja on yksi keskeisimmistä klassisen testiteorian rajoituksista, jotka ovat motivoineet osiovasteteorian kehittämistä.

3.3 Reliabiliteetti osiovasteteoriassa

Osiovasteteoria lähestyy reliabiliteettiongelmaa eri tavalla kuin klassinen testiteoria. Sen sijaan, että tarkasteltaisiin pistemäärien korrelaatioita testikertausten välillä, osiovasteteoria tarkastelee informaatiota eli kuinka paljon kukin osio ja koko testi kertovat vastaajan piilomuuttujasta sen eri tasoilla θ [1].

3.3.1 Fisher-informaatio reliabiliteetin perustana

Osiovasteteorian informaatiokäsite pohjautuu Fisherin informaatioon, jonka matemaattiset perusteet hän esitteli vuonna 1922 tilastollisen estimointiteorian yhteydessä[23]. Olkoon Y satunnaismuuttuja, jolla on tiheysfunktio $f(y; \theta)$, jossa θ on parametri ja θ_0 sen todellinen, mutta tuntematon arvo. Fisherin informaatio parametrin todellisella arvolla θ_0 määritellään log-uskottavuusfunktion derivaatan neliön odotusarvona, jossa odotusarvo lasketaan jakauman $f(y; \theta_0)$ mukaan[24]:

$$I(\theta_0) = E_{\theta_0} \left[\left(\frac{\partial \log f(Y; \theta_0)}{\partial \theta_0} \right)^2 \right]. \quad (19)$$

Vaihtoehtoinen, usein laskennallisesti kätevämpi muoto saadaan log-uskottavuuden toisen derivaatan negatiivisena odotusarvona[24]:

$$I(\theta_0) = -E_{\theta_0} \left[\frac{\partial^2 \log f(Y; \theta_0)}{\partial \theta_0^2} \right]. \quad (20)$$

Fisherin informaation keskeinen tulos on, että suurimman uskottavuuden estimaattorin (SU-estimaattori) asymptoottinen varianssi on $1/I(\theta_0)$ tiettyjen säännöllisyysehtojen täyttyessä. Fisherin informaatio on ei-negatiivinen luku, kun θ_0 on skalaari, ja positiivisesti semidefiniitti matriisi, kun θ_0 on vektori. Se asettaa ylärajan minkä tahansa harhattoman estimaattorin varianssille tai kovarianssimatriisille moniulotteisessa tapauksessa. Toisin sanoen SU-estimaattori on asymptoottisesti optimaalinen varianssin minimoinnin suhteen.[23, 24]

Birnbaum[2] sovelsi Fisherin informaatiota osiovasteteorian kontekstiin hyödyn-
täen Pitmanin[25] esittämää asymptoottisen tehokkuuden teoriaa. Lähtökohtana on
oletus, että tarkasteltavaa piilomuuttujaa voidaan mitata testillä, jonka osiot nou-
dattavat yksiulotteista osiovasteteoriamallia, ja että henkilön piilomuuttujan taso
on θ . Olkoon y testin havaittu pistemäärä. Jos sama testi teetetettäisiin henkilöllä
toistuvasti ja riippumattomasti, havaittu pistemäärä y muodostaisi jakauman, jolla
on odotusarvo $\mu_{y|\theta}$ ja keskihajonta $\sigma_{y|\theta}$. [24]

Lordin[20] mukaan intuitiivinen tapa mitata, kuinka hyvin testi erottaa kaksi
piilomuuttujan tasoa θ_1 ja θ_2 toisistaan, on seuraava signaali-kohinasuhde:

$$\frac{(\mu_{y;\theta_2} - \mu_{y;\theta_1})/\sigma_{y;\theta_1,\theta_2}}{\theta_2 - \theta_1}, \quad (21)$$

missä $\sigma_{y;\theta_1,\theta_2}$ on mittauksen keskivirheen yhdistetty estimaatti kahdella taitotasolla.
Nimittäjässä käytetty $\theta_2 - \theta_1$ perustuu oletukseen, että odotetun pistemäärän etäis-
syys kasvaa lineaarisesti piilomuuttujien tasojen etäisyyden kasvaessa. Tämä ole-
tus toteutuu vain, kun θ_1 ja θ_2 ovat hyvin lähellä toisiaan. Ottamalla raja-arvo
 $\theta_2 \rightarrow \theta \leftarrow \theta_1$ saadaan:

$$\lim_{\theta_2 \rightarrow \theta \leftarrow \theta_1} \frac{(\mu_{y;\theta_2} - \mu_{y;\theta_1})/\sigma_{y;\theta_1,\theta_2}}{\theta_2 - \theta_1} = \frac{\partial \mu_{y;\theta} / \partial \theta}{\sigma_{y;\theta}}. \quad (22)$$

Tämä liittyy läheisesti Pitmanin esittämään asymptoottiseen tehokkuuteen[25] ja
kuva, kuinka hyvin piilomuuttujien väliset erot ovat havaittavissa havaitun piste-
määrän y avulla. [24]

Birnbaum[2] nimesi kaavan (22) neliön havaitun pistemäärän y informaatiofunk-
tioksi:

$$I(\theta, y) = \frac{\left[\frac{\partial E(y; \theta)}{\partial \theta} \right]^2}{\sigma_{y|\theta}^2}. \quad (23)$$

Birnbaum osoitti, että kaava (23) on aina ylhäältä rajoitettu Fisherin informaatiolla
 $I(\theta)$. Kun $y = u_i$ on yksittäinen dikotominen vastaus osiolle i , kaava (23) yksinker-
taistuu muotoon:

$$I(\theta, u_i) = \frac{\left[\frac{\partial E(u_i|\theta)}{\partial \theta} \right]^2}{P_i(\theta) Q_i(\theta)} = \frac{[P'_i(\theta)]^2}{P_i(\theta) Q_i(\theta)}, \quad (24)$$

missä $P_i(\theta) = 1 - Q_i(\theta)$ on todennäköisyys oikealle vastaukselle osiolla i ja $P'_i(\theta) =$
 $\frac{d}{d\theta} P_i(\theta)$ on sen ensimmäinen derivaatta. Birnbaum osoitti, että kaava (24) saavuttaa
Fisherin informaation havainnolle $U_i = u_i$, ja nimesi sen *osioinformaatiofunktio*ksi
(englanniksi *item information function*) [2]. Kaavasta nähdään, että osio tuottaa
eniten informaatiota siellä, missä $P'_i(\theta)$ on suurimmillaan. [24]

Polytomisille osioille, kuten luokitellussa vastausmallissa, osioinformaatiofunktio
yleistyy summaksi yli kaikkien vastausluokkien. Merkitään $P_{u_i}(\theta)$ luokkavastefunk-
tiota ja $P'_{u_i}(\theta)$ sen derivaattaa. Tällöin osion i informaatiofunktio on [18]:

$$I_i(\theta) = \sum_{u_i=0}^{m_i} \frac{[P'_{u_i}(\theta)]^2}{P_{u_i}(\theta)}. \quad (25)$$

Kaava (24) on kaavan (25) erikoistapaus, kun $m_i = 1$.

Koska osiot ovat paikallisesti riippumattomia θ :n suhteen ja Fisherin informaatiolla on additiivisuusominaisuus, *testi-informaatiofunktio* (englanniksi *test information function*) saadaan osioinformaatioiden summana[2]:

$$I(\theta) = \sum_{i=1}^n I_i(\theta). \quad (26)$$

Koska SU-estimaattorin asymptoottinen varianssi on $1/I(\theta)$, saadaan taitotason SU-estimaattorin $\hat{\theta}$ ehdolliselle mittauksen keskivirheelle approksimaatio:

$$\text{CSEM}(\theta) \approx \frac{1}{\sqrt{I(\theta)}}. \quad (27)$$

Tätä kutsutaan myös henkilöparametrin θ keskivirheeksi, josta käytetään merkintää $\text{SE}(\theta)$. Toisin kuin klassisen testiteorian SEM, joka on yksi kiinteä luku koko populaatiolle, osiovasteteorian ehdollinen mittauksen keskivirhe $\text{CSEM}(\theta)$ vaihtelee θ mukaan. Joten suuri informaation määrä tietyllä θ välillä merkitsee pientä mittauksen keskivirhettä kyseisellä alueella.[1, 24]

Testi-informaatiofunktion additiivisuus merkitsee osiovasteteorian ja klassisen testiteorian välistä ratkaisevaa eroa. Yksittäisen osion panos mittaustehokkuuteen ei riipu siitä, mitä muita osioita testissä on[20]. Klassisessa testiteoriassa osion poistaminen tai lisääminen muuttaa koko testin reliabiliteettia tavoilla, jotka riippuvat muista osioista. Osiovasteteoriassa kukin osio lisää oman informaationsa piilomuuttujan suhteen riippumatta testin muista osioista.

3.3.2 Marginaalinen reliabiliteetti

Testi-informaatiofunktio kuvaa mittaustarkkuutta yksittäisillä piilomuuttujan arvoilla θ , mutta se ei sellaisenaan tuota yhtä lukua koko populaation mittaustarkkuudelle. Tähän tarkoitukseen Green et al.[26] esittelivät marginaalisen reliabiliteetin tietokonepohjaisten mukautuvien testien (englanniksi *Computerized Adaptive Tests*, *CAT*) yhteydessä.

Marginaalinen reliabiliteetti on verrattavissa populaatiotason reliabiliteettiin klassisessa testiteoriassa. Klassisen testiteorian reliabiliteetti kuvaa, kuinka suuri osa havaittujen pisteiden varianssista on todellisen taitotason vaihtelua koko populaatiossa. Marginaalinen reliabiliteetti ilmaisee saman asian osiovasteteorian kehityksessä kuvaamalla mittaustarkkuutta koko populaation yli eikä yksittäisellä piilomuuttujan tasolla θ [27]. Lisäksi marginaalinen reliabiliteetti kuvaa saadun informaation määrää niiltä piilomuuttujan alueilta, joissa populaation yksilöt ovat.

Cheng, Yuan ja Liu[27] osoittavat, että kun piilomuuttujan varianssiksi oletetaan $\sigma_\theta^2 = 1$ ja piilomuuttujan suurimman uskottavuuden estimaattorin $\hat{\theta}_{\text{ML}}$ virhetermin varianssi on Fisherin informaation käänteisluku $\sigma_e^2 = 1/I$, niin marginaalisen reliabiliteetin perusmääritelmä (14) saa tällöin muodon

$$\rho = \frac{1}{1 + 1/I} = \frac{I}{I + 1}. \quad (28)$$

Osiovasteteoriassa informaatio riippuu piilomuuttujan tasosta θ eli $I = I(\theta)$. Tämä tekee reliabiliteetista paikallisen käsitteen. Green et al.[26] ehdollistivat reliabiliteetin piilomuuttujan tason θ suhteen, jolloin yhtälöstä (28) saadaan ehdollinen reliabiliteetti kiinteällä piilomuuttujan tasolla θ :

$$\rho(\theta) = \frac{I(\theta)}{I(\theta) + 1}. \quad (29)$$

Marginaalinen reliabiliteetti saadaan integroimalla ehdollinen reliabiliteetti (29) populaatiojakauman $g(\theta)$ yli, jonka varianssi on $\sigma_\theta^2 = 1$:

$$\rho_{\text{marg}} = \mathbb{E}_\theta[\rho(\theta)] = \int_{-\infty}^{\infty} \frac{I(\theta)}{I(\theta) + 1} g(\theta) d\theta, \quad (30)$$

missä $I(\theta)$ on testi-informaatiofunktio ja $g(\theta)$ on populaation taitotason tiheysfunktio. Tätä marginaalisen reliabiliteetin muotoa kutsutaan myös marginaaliseksi reliabiliteetikertoimeksi ja käytetään tämän työn simulaatiotutkimuksessa tunnuslukuksi.

Yhtälöstä (30) havaitaan, että marginaalinen reliabiliteetti riippuu sekä testin informaatiosta $I(\theta)$ että populaatiojakaumasta $g(\theta)$. Tämä tarkoittaa, että saman testin marginaalinen reliabiliteetti voi saada eri arvoja eri populaatioissa. Populaatiojakauman $g(\theta)$ vinouden vaikutus marginaaliseen reliabiliteettiin on tämän työn tutkimuksen kohde.

4 Menetelmät

Työn empiirinen osuus toteutettiin Monte Carlo -simulaatiotutkimuksena. Simulaatiossa manipuloitiin neljää tekijää: osiovastemallia, piilomuuttujan θ jakauman vinousparametria α , otoskokoa ja testin pituutta. Jokainen näiden tekijöiden yhdistelmä tutkittiin erikseen, mikä tuotti yhteensä 168 parametriyhdistelmää. Simulaatio toteutettiin R-ohjelmiston versiolla 4.4.3[28] käyttäen seuraavia paketteja: *mirt* (versio 1.44.0)[29] osiovastemallien sovittamiseen ja testi-informaatiofunktioiden laskemiseen, *mirtCAT* (versio 1.14)[30] mirt-objektien luomiseen populaation marginaalisten reliabiliteettikertoimien laskemista varten, *sn* (versio 2.1.3)[31] vinonormaalijakaumasta poimintaan, *furrr* (versio 0.3.1)[32] simulaation rinnakkaisajoon, sekä *ggplot2* (versio 4.0.2)[33], *RColorBrewer* (versio 1.1-3)[34] ja *dplyr* (versio 1.1.4)[35] tulosten visualisointiin ja aineiston käsittelyyn.

Työn tutkimuskysymykset ovat:

1. Miten piilomuuttujan jakauman vinous vaikuttaa marginaalisen reliabiliteettikertoimen estimaattorin harhaan kaksiparametrisessa logistisessa mallissa ja luokitellussa vastausmallissa?
2. Miten vastauskategorioiden lukumäärä vaikuttaa estimaattorin harhaan?
3. Miten otoskoko vaikuttaa estimaattorin harhaan?
4. Miten testin pituus vaikuttaa estimaattorin harhaan?

Otoskoon kasvaessa parametriestimoinnista johtuvan satunnaisen vaihtelun odotetaan pienenevän, mutta estimaattorin harhan ei odoteta häviävän kokonaan, kun vinonormaalijakauman parametri $\alpha \neq 0$. Tämä johtuu siitä, että marginaalisen reliabiliteetin estimaattorin standardinormaalijakaumaoletus poikkeaa systemaattisesti todellisesta piilomuuttujan jakaumasta otoskoosta riippumatta. Testin pituuden osalta hypoteesina on, että pidemmissä testeissä jakaumaoletusten kohtaamattomuudesta aiheutuva harha pienenee, koska testi-informaatiofunktion arvot ovat tällöin korkeammat laajemmalla piilomuuttujan tason θ välillä[18].

4.1 Osiovastemallit

Simulaatiossa tarkasteltiin kahta osiovastemallia: kaksiparametrista logistista mallia (2PL) sekä luokiteltua vastausmallia (GRM). Luokiteltua vastausmallia tutkittiin kolmella ja viidellä vastauskategorioiden. Tällä tavoin vastausasteikon pituuden vaikutusta voitiin tarkastella. Mallien valinta perustui niiden yleisyyteen psykologisessa mittauksessa sekä siihen, että 2PL-malli voidaan esittää luokitellun vastausmallin erikoistapauksena, kun vastauskategorioita on kaksi[18]. Koska molemmat mallit perustuvat samaan matemaattiseen viitekehykseen, vinon piilomuuttujan jakauman vaikutusta voidaan verrata dikotomisen (2PL) ja polytomisen (GRM) mallin tuottamia reliabiliteettikertoimia toisiinsa mielekkäästi.

2PL-mallissa osioiden vastaukset ovat dikotomisissa (esimerkiksi 0 tai 1), ja malli sisältää kaksi osioparametria: erotteluparametrin a ja vaikeusparametrin b . Luokitellussa vastausmallissa vastaukset ovat tässä työssä järjestysasteikollisia kolmella

(GRM-3) tai viidellä vastauskategoriolla (GRM-5). Luokiteltu vastausmalli sisältää jokaiselle osiolle erotteluparametrin sekä $m - 1$ kynnysparametria, joissa m on vastauskategorioiden lukumäärä. Luokiteltu vastausmalli viidellä vastauskategoriolla vastaa tyypillistä viisiportaista Likert-asteikkoa.

Vastauskategorioiden lukumäärien 3 ja 5 sisällyttäminen simulaatioon mahdollisti sen arvioimisen, missä määrin ne vaikuttavat marginaalisen reliabiliteettiker-toimeen vinon θ -jakauman kanssa. Samejima[18] osoitti, että vastausluokkien lisääminen kasvattaa osion tuottamaa informaatiota, edellyttäen että kynnysparametrit ovat riittävän etäällä toisistaan. Vinon piilomuuttujan jakauman vaikutus tähän on toistaiseksi tutkimatta.

4.2 Piilomuuttujan jakauma

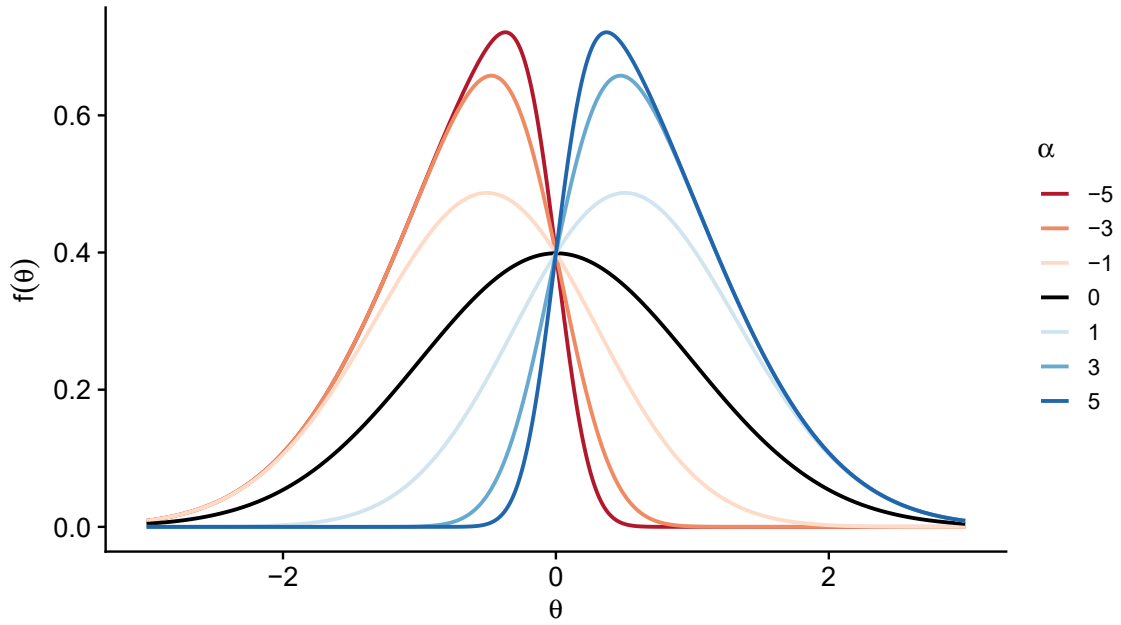
Piilomuuttujan jakaumana käytettiin vinonormaalijakaumaa, jonka Azzalini[6] esit-teli normaalista poikkeavien jakaumien mallintamiseen parametrisena jakaumaper-heenä. Vinonormaalijakauman $SN(\alpha)$ tiheysfunktio on määritelty muodossa

$$\phi(z; \alpha) = 2 \phi(z) \Phi(\alpha z), \quad -\infty < z < \infty, \quad (31)$$

jossa ϕ ja Φ ovat standardinormaalijakauman tiheys- ja kertymäfunktio ja $\alpha \in \mathbb{R}$ on jakauman muotoa ohjaava vinousparametri. Kun $\alpha = 0$, yhtälö supistuu muotoon $\phi(z; 0) = \phi(z)$, eli jakauma palautuu standardinormaalijakaumaksi $N(0, 1)$. Negatiiviset arvot $\alpha < 0$ tuottavat vasemmalle vinon jakauman ja positiiviset arvot $\alpha > 0$ oikealle vinon jakauman. Lisäksi jos $Z \sim SN(\alpha)$, niin $-Z \sim SN(-\alpha)$. Jakauman sijaintia ja skaalaa voidaan säädetään lineaarisella muunnoksella $Y = \xi + \omega Z$, jossa ξ on sijaintiparametri ja $\omega > 0$ skaalaparametri. Simulaatiossa asetettiin $\xi = 0$ ja $\omega = 1$ kaikissa ehdoissa, jolloin ainoastaan vinousparametri α vaihteli.[6]

Piilomuuttujan arvot θ generoitiin vinonormaalijakaumasta käyttäen vinouspara-metrin arvoja $\alpha \in \{-5, -3, -1, 0, 1, 3, 5\}$, joissa negatiiviset arvot tuottavat vasem-malle vinon jakauman ja positiiviset arvot oikealle vinon jakauman. Arvo $\alpha = 0$ toi-mi kontrolliehtoina, jolloin jakauma palautuu standardinormaalijakaumaksi $N(0, 1)$. Arvot valittiin siten, että ne mahdollistavat sekä vinouden suunnan että voimakkuu-den vaikutuksen tarkastelun symmetrisesti molempiin suuntiin. Theta-arvot gene-roitiin R-paketin *sn* funktiolla *rsn*[31].

Kuvassa 4 esitetään simulaatiotutkimuksessa käytetyt piilomuuttujan vinonor-maalijakaumat kaikilla vinousparametrin α arvoilla. Kuvasta nähdään, että vinous-parametrin α etumerkki ohjaa jakauman vinouden suuntaa ja sen itseisarvon kasvu lisää vinouden voimakkuutta.



Kuva 4: Simulaatiossa käytetyt vinonormaalijakaumat $\text{SN}(\xi = 0, \omega = 1, \alpha)$ vinousparametrin arvoilla $\alpha \in \{-5, -3, -1, 0, 1, 3, 5\}$. Kun $\alpha = 0$, jakauma on standardinormaali $N(0, 1)$.

4.3 Otokoot, replikaatioiden määrä ja testin pituus

Simulaatiossa käytettiin neljää otoskokoa: $n \in \{250, 500, 1000, 5000\}$. Otokoot $n = 250$, $n = 500$ ja $n = 1000$ valittiin, koska ne vastaavat aiemmissa ei-normaalisuutta tarkastelevissa IRT-simulaatiotutkimuksissa käytettyjä otoskokoja[36]. Suurin otoskoko $n = 5000$ lisättiin mahdollistamaan marginaalisen reliabiliteettikertoimen asympotoottisen käyttäytymisen tarkastelun. Riittävän suurilla otoksilla estimaatin odotetaan stabiloituvan populaatioarvoaan vastaavalle tasolle. Tähän tarkoitukseen Andersson ja Xin[7] sisällyttivät tutkimukseensa vastaavasti suuret otokoot $n = 4000$ ja $n = 8000$.

Simulaatiossa tarkasteltiin kahta testin pituutta: 10 osiota ($k = 10$) ja 40 osiota ($k = 40$). Molemmat pituudet ovat peräisin Manapatin ja Edwardsin[36] tutkimuksesta, jossa tarkasteltiin lisäksi 20 osion testiä. Keskimäinen pituus jätettiin pois simulaatioiden kokonaismäärän pitämiseksi hallittavana, sillä $k \in \{10, 40\}$ mahdollistaa jo lyhyen ja pitkän testin vaikutuksen merkityksellisen vertailun.

Kutakin parametriyhdistelmää toistettiin 1000 kertaa, mikä on linjassa aiempien simulaatiotutkimusten kanssa[36]. Taulukossa 1 esitetään simulaatiotutkimuksen rakenne tiivistetysti.

Taulukko 1: Simulaatioasetelman rakenne.

Tekijä	Tasoja	Arvot
Osiovastemalli	3	2PL, GRM-3, GRM-5
Vinousparametri (α)	7	-5, -3, -1, 0, 1, 3, 5
Otoskoko (n)	4	250, 500, 1000, 5000
Testin pituus (k)	2	10, 40
Replikaatiot / ehto	–	1000
Parametriyhdistelmiä	–	168
Replikaatioita yhteensä	–	168000

4.4 Marginaalisen reliabiliteettikertoimen laskenta

Estimoitu marginaalinen reliabiliteettikerroin $\hat{\rho}$ laskettiin kahdessa vaiheessa jokaisessa replikaatiossa. Ensin osiovastemalli sovitettiin simuloituun vastausaineistoon käyttäen R-paketin `mirt` funktiota `mirt`[29]. Mallien sovittamiseen käytettiin funktion oletusalgoritmia, joka on Bockin ja Aitkinin[37] esittämä EM-algoritmi. EM-algoritmi olettaa piilomuuttujan θ jakaumaksi standardinormaalijakauman $N(0, 1)$ [37]. Toisessa vaiheessa sovitetun mallin testi-informaatiofunktio $I(\theta)$ laskettiin `mirt`-paketin `testinfo`-funktioilla[29], minkä jälkeen suoritettiin yhtälön (32) numeerinen integrointi R:n `integrate`-funktioilla yli koko reaaliakselin.

$$\hat{\rho} = \int_{-\infty}^{\infty} \frac{I(\theta)}{1 + I(\theta)} \phi(\theta) d\theta, \quad (32)$$

jossa $\phi(\theta)$ on standardinormaalijakauman tiheysfunktio. Viitearvona toimii populaation marginaalinen reliabiliteettikerroin, joka laskettiin todellisten osioparametrien avulla, mutta integroimalla vinonormaalijakauman $SN(0, 1, \alpha)$ suhteen:

$$\rho_{\text{pop}} = \int_{-\infty}^{\infty} \frac{I(\theta)}{1 + I(\theta)} f_{\text{SN}}(\theta; 0, 1, \alpha) d\theta. \quad (33)$$

Kun piilomuuttujan θ jakauma on vino eli $\alpha \neq 0$, estimaattorin (32) käyttämä standardinormaalijakaumaoletus ja populaatioarvon (33) määrittävä todellinen jakauma eivät kohtaa. Tästä voi seurata systemaattista harhaa, joka on otoskoosta riippumaton. Koska todelliset osioparametrit pidettiin vakioina kaikissa replikaatioissa, populaatioarvo ρ_{pop} laskettiin vain kerran kutakin parametriyhdistelmää kohden.

4.5 Uudelleenpöimintaehto

IRT-mallien sovittamiseen käytettiin R-paketin `mirt` funktiota `mirt`[29]. Luokiteltu vastausmalli edellyttää, että jokaisessa vastausluokassa on riittävästi havaintoja parametriestimoinnin onnistumiseksi. Voimakas vinous voi johtaa tilanteeseen, jossa ääriluokkiin (esim. matalin tai korkein vastauskategoria) kertyy hyvin vähän tai ei lainkaan vastauksia. Tällöin `mirt`-funktio yhdistää automaattisesti tyhjiä tai lähes tyhjiä vastauskategorioita viereisiin kategorioihin, jolloin sovitetun mallin vastauskategorioiden rakenne poikkeaa suunnitellusta. Tämä vääristäisi mallien välistä vertailua sekä tulosten vertailukelpoisuutta eri vinousehtojen välillä.

Kuten Manapat ja Edwards[36] menettelivät tutkimuksessaan, simuloitu vastausaineisto tarkistettiin ennen estimointia jokaisella replikaatiokerralla. Jos missään luokitellun vastausmallin vastausluokassa oli alle 5 havaintoa, hylättiin kyseinen aineisto ja generoitiin uusi vastaajajoukko samoilla parametreilla. Uudelleenpoimintojen enimmäismääräksi asetettiin 500 replikaatiota kohden. Jos raja ylittyi, replikaatio kirjattiin epäonnistuneeksi ja suljettiin pois analyysistä. Uudelleenpoimintojen lukumäärä kirjattiin jokaisessa ehdossa, jotta voidaan arvioida, missä vinousarvoissa uudelleenpoimintaa tarvittiin merkittävässä määrin.

Uudelleenpoimintaehto pyrittiin soveltamaan ainoastaan luokiteltuihin vastausmalleihin (GRM-3 ja GRM-5). Perusteena tälle oli, että kaksiparametrisen logistisen mallin dikotomisissa osioissa vastaukset jakautuvat vain kahteen luokkaan, jolloin ongelmaa vastausten kasautumisesta vain yhteen vastauskategoriaan ei odotettu esiintyvän. Simulaation testausvaiheessa kuitenkin huomattiin, että pienellä otoskoollla $n = 250$ ja voimakkaasti vinolla piilomuuttujan θ jakaumalla myös 2PL-mallissa esiintyi tilanteita, jossa vastaajien voimakas kasaantuminen jakaumansa toiseen ääripäähän johti siihen, että toiseen vastauskategoriaan kertyi alle viisi havaintoa. Näin ollen uudelleenpoimintaehto päätettiin ulottaa koskemaan kaikkia malleja estimointivirheiden välttämiseksi ja tulosten vertailukelpoisuuden varmistamiseksi.

4.6 Osioparametrien generointi

Osioparametrit generoitiin erikseen jokaiselle mallin ja testin pituuden yhdistelmälle, ja näin saatuja osioparametreja käytettiin kiinteinä kyseisen yhdistelmän kaikissa replikaatioissa sekä kaikilla vinous- ja otoskokoehdoilla. Yhteensä generoitiin kuusi toisistaan riippumatonta osioparametrijoukkoa. Tämä kiinteiden parametrien asetelma mahdollisti sen, että replikaatioiden välinen vaihtelu heijastaa puhtaasti vastaajajoukon otannasta johtuvaa satunnaisvaihtelua eikä sekoitu osioparametrien satunnaisesta vaihtelusta johtuvaan varianssiin. Tulosten toistettavuuden varmistamiseksi parametrien generoinnissa käytettiin satunnaislukugeneraattorin siementä. Työssä käytetyt generoidut osioparametrit on saatavilla työn liitteessä B.

Luokitellun vastausmallin osioparametrit generoitiin Houtsin ja Edwardsin[38] kuvaamalla menettelyllä, jota myös Manapat ja Edwards[36] sovelsivat. Erottelukykyparametrit a generoitiin normaalijakaumasta, jonka odotusarvo ja keskihajonta perustuivat Hillin[39] empiiriseen dokumentaatioon todellisista luokitellun vastausmallin sovituksista:

$$a \sim N(1.7, 0.3). \quad (34)$$

GRM-5 -mallin kynnysparametrit b_i generoitiin kumulatiivisella prosessilla, joka varmistaa parametrien järjestyshdon $b_1 < b_2 < b_3 < b_4$. Ensimmäinen kynnysparametri generoitiin normaalijakaumasta:

$$b_1 \sim N(-1.5, 0.5). \quad (35)$$

Seuraavat kynnysparametrit muodostettiin lisäämällä edelliseen positiivinen erotusparametri d_i , joka generoitiin normaalijakaumasta:

$$d_i \sim N(1.0, 0.2), \quad d_i > 0, \quad (36)$$

$$b_2 = b_1 + d_1, \quad b_3 = b_2 + d_2, \quad b_4 = b_3 + d_3. \quad (37)$$

Erotusparametrien positiivisuus varmistettiin hylkäämällä negatiiviset arvot ja generoimalla uusi arvo, kunnes saatiin positiivinen erotus. GRM-3 -mallille käytettiin vastaavaa menettelyä kahdella kynnysparametrilla, jolloin $b_2 = b_1 + d_1$.

Kaksiparametrinen logistisen mallin erotteluparametrit a generoitiin normaalijakaumasta Manapatin ja Edwardsin[36] mukaan:

$$a \sim N(1.19, 0.3). \quad (38)$$

Vaikeusparametrit b generoitiin standardinormaalijakaumasta Anderssonin ja Xinin[7] mukaisesti:

$$b \sim N(0, 1). \quad (39)$$

4.7 Simulaatiotulosten raportointi

Raportoinnissa noudatettiin Anderssonin ja Xinin [7] käyttämää menettelyä. Jokaisessa parametriyhdistelmässä laskettiin Monte Carlo -keskivirhe (englanniksi *Monte Carlo standard error*, MCSE) sekä keskimääräinen poikkeama populaatioarvosta vähentämällä populaatioarvo ρ_{pop} replikaatioiden yli lasketusta marginaalisen reliabiliteettikertoimen keskiarvosta:

$$\hat{b}(\hat{\rho}) = \sum_{r=1}^R \frac{\hat{\rho}_r}{R} - \rho_{\text{pop}}, \quad (40)$$

jossa R on onnistuneiden replikaatioiden määrä ja $\hat{\rho}_r$ on r :nnen replikaation marginaalisen reliabiliteettikertoimen estimaatti. Harhalla tarkoitetaan estimaattorin odotusarvon poikkeamaa todellisesta arvosta ja yhtälössä (40) laskettu keskimääräinen poikkeama on tämän harhan estimaattori. MCSE puolestaan kuvaa harhan estimaattorin otantavaihtelua simulaatiossa. Jatkossa keskimääräisestä poikkeamasta, joka on siis harhan estimaattori, käytetään lyhyesti termiä harha.

5 Tulokset

5.1 Uudelleenpoiminta, konvergenssi ja Monte Carlo -keskivirhe

Tulosten analysoinnista jätettiin pois parametriyhdistelmät, joissa onnistuneita replikaatioita kertyi alle 100 kappaletta. Epäonnistuneeksi replikaatio luokiteltiin, jos ei löydetty 500 poiminnalla piilomuuttujan jakaumasta vastaajajoukkoa, jolla simuloidussa aineistossa jokaisessa vastauskategoriassa oli 5 tai enemmän vastaajaa. Lisäksi epäonnistuneeksi replikaatioksi luokiteltiin ne, joiden mallin sovitus ei konvergoinut.

Kummallakin vastauskategorian lukumäärän ja testipituuden arvolla, otoskoilla $n = 250$ ja $n = 500$ sekä vinousparametrin $|\alpha| \in \{3, 5\}$ arvoilla luokitellulla vastausmallilla onnistuneiden replikaatioiden määrä jäi alle 100 kappaleen. GRM-5 -mallilla testipituudella $k = 40$ ja otoskoolla $n = 250$ sekä vinousparametrin $|\alpha| \in \{3, 5\}$ ei saatu yhtään onnistunutta replikaatiota. GRM-3 -mallissa sama ilmiö huomattiin samalla otoskoolla ja testipituudella, kun vinousparametrin arvo $\alpha \in \{3, 5\}$.

Epäonnistuneiden replikaatioiden lukumäärä ei ylittänyt 100 kappaleen rajaa 2PL-mallin tuloksissa. Toisin kuin GRM-malleissa 2PL-mallissa mallin sovituksessa esiintyi kuitenkin ilmiötä, jossa mallin sovitus ei konvergoitunut. Mallin konvergoitongelmia esiintyi testipituudella $k = 10$, otoskoolla $n = 250$ ja vinousparametrin arvoilla $\alpha \in \{-5, 3, 5\}$. Suurin replikaatioiden lukumäärä, joissa ei tapahtunut konvergoitua, oli 89 kappaletta.

Tulosten analysoinnista poistetut parametriyhdistelmät, niiden onnistuneiden replikaatioiden lukumäärä ja harha on esitetty taulukossa 2.

Taulukko 2: Parametriyhdistelmät, jotka jätettiin pois analyysistä. Viiva merkitsee, että arvoa ei pystytty laskemaan, koska onnistuneita replikaatioita oli nolla kappaletta.

Malli	k	n	α	Onnistuneet	Harha
GRM-3	10	250	3	29	-0,203
GRM-3	10	250	5	6	-0,248
GRM-3	40	250	3	0	–
GRM-3	40	250	5	0	–
GRM-3	40	500	5	69	-0,128
GRM-5	10	250	-5	1	-0,139
GRM-5	10	250	-3	5	-0,112
GRM-5	40	250	-5	0	–
GRM-5	40	250	-3	0	–
GRM-5	40	250	3	0	–
GRM-5	40	250	5	0	–
GRM-5	40	500	-5	1	-0,042
GRM-5	40	500	-3	15	-0,036
GRM-5	40	500	3	11	-0,035
GRM-5	40	500	5	2	-0,039

MCSE-arvot olivat kaikissa analyysiin mukaan otetuissa parametriyhdistelmissä pieniä ja vaihtelivat välillä $[0,00002; 0,00159]$. Koska nämä simulaation keskivirheen

arvot olivat kauttaaltaan pieniä, estimaattorin havaitun harhan ei oleteta selittyvän simulaation tuottamasta kohinasta.

5.2 Harha vinousparametrin α suhteen

Kuvassa 5 esitetään marginaalisen reliabiliteettikertoimen estimaattorin harha piilomuuttujan vinonormaalijakauman vinousparametrin α suhteen kaikissa simulaation parametrijhdistelmissä. Kun vinousparametri $\alpha = 0$ estimaattorin jakaumaoletus vastasi piilomuuttujan todellista jakaumaa, harha oli käytännöllisesti katsottuna nolla kaikissa parametrijhdistelmissä. Tällöin harha vaihteli välillä $[-0,0001; 0,0070]$. Tulos vahvistaa sen, että estimaattori toimii harhattomasti, kun piilomuuttujan jakauma vastaa estimaattorin standardinormaalijakaumaa.

Kun vinousparametri α poikkesi nolasta huomattiin, että reliabiliteettikertoimen harha on negatiivista kaikissa parametrijhdistelmissä. Lisäksi harha kasvoi monotonisesti vinousparametrin α itseisarvon kasvaessa. Harha oli suurimmillaan 0,280, kun mallina oli kolmikategorinen luokiteltu vastausmalli (GRM-3), vinousparametri $\alpha = 5$, otoskoko $n = 5000$ ja testipituus $k = 10$.

Harhan huomattiin käyttäytyvän lähes symmetrisesti vinousparametrin α itseisarvon suhteen. Poikkeuksen tähän muodosti kolmikategorinen luokiteltu vastausmalli, jossa keskimääräinen harha oli systemaattisesti suurempaa positiivisilla vinousparametrin α arvoilla kuin vastaavilla negatiivisilla arvoilla. Esimerkiksi parametrijhdistelmällä testin pituus $k = 10$, otoskoko $n = 5000$ ja vinousparametri $\alpha = -5$, kolmikategorisen luokitellun vastausmallin harha oli noin 0,144. Samalla parametrijhdistelmällä paitsi vinousparametrin α arvolla 5, estimaattorin harha oli noin 0,28.

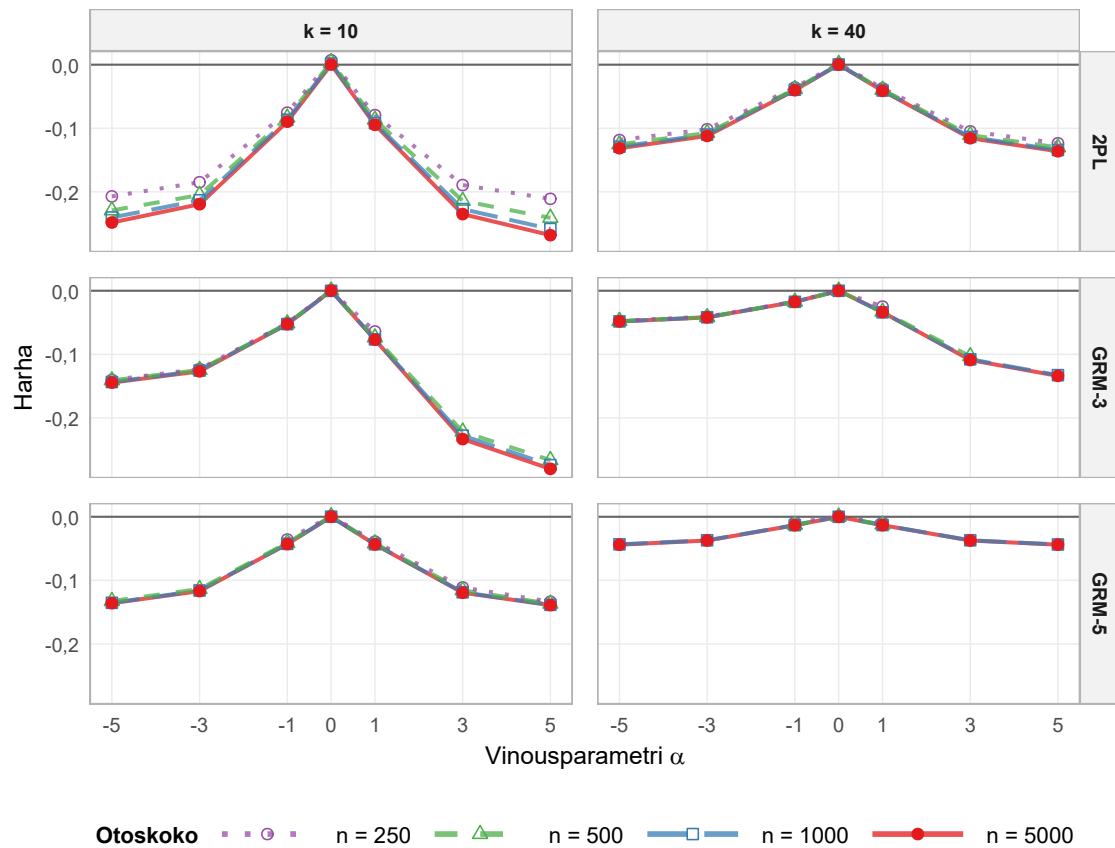
5.3 Muiden parametrijhdistelmien vaikutus estimaattorin harhaan

5.3.1 Vastauskategorioiden lukumäärän vaikutus

Liitteen A taulukossa 3 esitetään harha otoskoolla $n = 5000$ ja kaikilla vinousparametrin α arvoilla kummallakin testipituudella k . Vastauskategorioiden lukumäärän huomattiin vaikuttavan marginaalisen reliabiliteettikertoimen harhaan vähentävästi ja sen riippuvan vinousparametrin α etumerkistä.

Negatiivisilla vinousparametrin α arvoilla harha pieneni johdonmukaisesti vastauskategorioiden lukumäärän kasvaessa. Siirryttäessä kahdesta vastauskategorias- ta kolmeen harha väheni huomattavasti ja viisikategorinen luokiteltu vastausmalli tuotti pienemmän harhan kuin kolmikategorinen.

Positiivisilla vinousparametrin α arvoilla siirryttäessä kahdesta vastauskategorias- ta kolmeen ei vähentänyt harhaa ja kasvatti sitä joissain ehdoissa. Viisikategorinen luokiteltu vastausmalli tuotti pienimmän harhan myös positiivisilla vinousparametrin α arvoilla.



Kuva 5: Marginaalisen reliabiliteettikertoimen estimaattorin harha vinousparametrin α suhteen osiovastemalleittain (rivit), testipituuksittain (sarakkeet) ja otoskoottain (viivat).

5.3.2 Otoskoon vaikutus

Liitteen A taulukossa 4 esitetään estimaattorin harha kaikilla tutkimuksen otoskoilla ja testipituudella $k = 10$. Otoskoon vaikutus harhaan oli vastakkainen verrattuna piilomuuttujan θ jakauman vinousparametrin α vaikutukseen. Harha kasvoi otoskoon kasvaessa, kun $|\alpha| \neq 0$. Kontrolliehdon $\alpha = 0$ harha pieneni nollaan otoskoon kasvaessa kaikissa malleissa.

Kun $|\alpha| \neq 0$, estimaattorin harha kasvoi otoskoon kasvaessa kaikissa malleissa. Kasvu oli suurinta kaksiparametrisessa logistisessa mallissa (2PL) ja luokitellussa vastausmallissa vähäisempää. Harhan kasvu oli 2PL-mallissa suurimmillaan noin 0,05 ja luokitellussa vastausmallissa noin 0,02, joten otoskoon vaikutus harhan suuruuteen oli kaikissa ehdoissa vähäistä. Myös testipituudella $k = 40$ otoskoko kasvatti harhaa, mutta kasvu oli vähäisempää kuin testipituudella $k = 10$.

5.3.3 Testin pituuden vaikutus

Testin pituuden vaikutus estimaattorin harhaan oli johdonmukainen kaikissa malleissa ja parametrijohdistelmissä. Testipituus $k = 40$ tuotti pienemmän harhan kuin testipituus $k = 10$. Liitteen A taulukossa 5 esitetään harha kummallakin testipituudella.

della ja otoskoolla $n = 5000$. Harha pieneni testipituuden kasvaessa kaikissa malleissa, kaikissa vinousparametrin α arvoilla ja kaikilla otoskoilla. Harha pieneni noin puoleen siirryttäessä testipituudesta $k = 10$ testipituuteen $k = 40$.

6 Pohdinta

Tässä työssä käytetyssä simulaatioasetelmassa marginaalisen reliabiliteettikertoimen estimaattoriin tuottaa harhaa kaksi lähdettä. Ensimmäinen harhan lähde on mallin sovituksessa käytetty Bockin ja Aitkinin [37] esittämä EM-algoritmi suurimman uskottavuuden menetelmälle, jossa käytetään oletuksena standardinormaalijakaumaa $N(0, 1)$. Toinen harhan lähde on marginaalisen reliabiliteettikertoimen estimaattorin laskennassa käytetty $N(0, 1)$ oletus. Tämän työn tutkimusasetelmassa näiden kahden harhan lähteen suuruutta ei voida erottaa toisistaan, mikä on työn keskeinen rajoitus. Jatkotutkimuksessa näitä harhan lähteitä voitaisiin tarkastella erikseen asetelmalla, jossa joko mallin sovitus tai reliabiliteettikertoimen estimaattorin laskenta suoritettaisiin todellisella vinonormaalijakaumalla $SN(0, 1, \alpha)$ ja toinen standardinormaalijakaumaoletuksella. Tällainen tutkimusasetelma paljastaisi kummankin harhan lähteen osuuden marginaalisen reliabiliteettikertoimen estimaattorin harhaan.

Simulaatiotutkimuksessa havaittiin, että marginaalisen reliabiliteettikertoimen estimaattorin harha on käytännössä nolla, kun piilomuuttujan jakauma vastaa estimaattorin oletusta. Harhan havaittiin olevan johdonmukaisesti negatiivista kaikissa parametriyhdistelmissä, kun piilomuuttujan vinonormaalijakauman vinousparametri $|\alpha| \neq 0$. Kahden tekijän huomattiin vähentävän estimaattorin itseisarvoista harhaa: vastauskategorioiden lukumäärän ja testin pituuden kasvattaminen. Näiden lisäksi tuloksista tehtiin kolme havaintoa:

1. kolmen vastauskategorian luokitellussa vastausmallissa (GRM-3) huomattiin harhan epäsymmetria vinousparametrin α suhteen,
2. estimaattorin harhan itseisarvo kasvoi otoskoon kasvaessa kaksiparametrisessä logistisessa mallissa testipituudella $k = 10$, kun vinousparametri $|\alpha| \neq 0$,
3. testipituudella $k = 40$ luokitellussa vastausmallissa estimaattorin harha oli lähes riippumaton vinousparametrin α arvosta lukuun ottamatta edellä mainittua epäsymmetriaa.

6.1 Estimaattorin harhan epäsymmetria kolmen vastauskategorian luokitellussa vastausmallissa

Tuloksia tarkasteltaessa havaittiin, että kolmen vastauskategorian luokitellussa vastausmallissa (GRM-3) marginaalisen reliabiliteettikertoimen estimaattorin harhan itseisarvo ei ollut symmetrinen vinousparametrin α suhteen kaikissa parametriyhdistelmissä. GRM-3-mallilla estimaattorin harhan itseisarvo oli johdonmukaisesti suurempi positiivisilla vinousparametrin α arvoilla kuin negatiivisilla arvoilla. Tämä epäsymmetria poikkeaa kahdesta muusta tarkastellusta mallista, joiden estimaattorin harhan itseisarvo käyttäytyi likimain symmetrisesti vinousparametrin α suhteen.

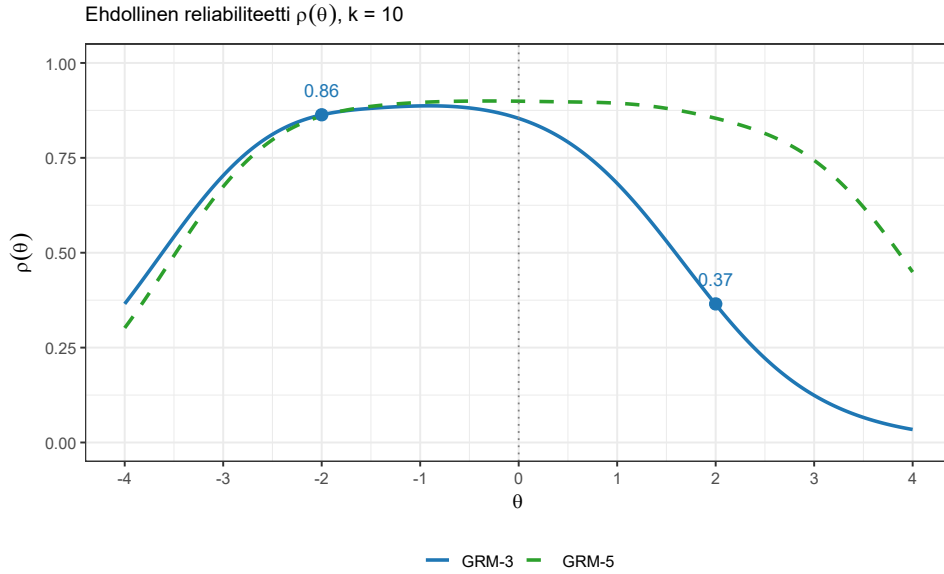
Epäsymmetrian syitä voidaan tarkastella ensin marginaalisen reliabiliteettikertoimen estimaattorin populaatiotason analyysillä, jossa sen harha lasketaan todellisilla osioparametreilla ilman mallien estimointia. Populaatiotason analyysi estimaattorin harhasta paljastaa, että estimaattorin $N(0, 1)$ jakaumaoletus on symmetrinen

vinousparametrin α suhteen. Populaatiotason estimaattorin harha kaikilla malleilla on esitetty liitteen C taulukosta 12. Lisäksi harhan etumerkki vastaa vinousparametrin α etumerkkiä. Liitteen C taulukosta 12 huomataan, että esimerkiksi GRM-3-mallilla testipituudella $k = 10$ todellisilla osioparametreilla laskettu estimaattorin harha on itseisarvoltaan noin 0,02 vinousparametrin äärimmäisillä arvoilla $|\alpha| = 5$. Simulaatiotutkimuksen tuloksissa estimaattorin harha oli sen sijaan johdonmukaisesti negatiivinen ja likimain symmetrinen vinousparametrin α suhteen kaksiparametrisessa ja viiden vastauskategorian luokitellussa vastausmallissa. GRM-3-mallilla harha oli myös negatiivinen vinousparametrin α suhteen, mutta itseisarvoltaan epäsymmetrinen. Koska estimaattorin jakaumaoletuksesta aiheutuva harha on populaatiotasolla positiivinen, kun vinousparametri $\alpha > 0$, sen vaikutuksen pitäisi pienentää harhan itseisarvoa vinousparametrin α positiivisilla arvoilla. Simulaatiotutkimuksen tuloksissa harhan itseisarvo on kuitenkin GRM-3-mallissa suurempi positiivisilla vinousparametrin α arvoilla, joten estimaattorin jakaumaoletuksen tuottama harha ei selitä kaikkea havaittua epäsymmetriaa.

Yksi mahdollinen syy estimaattorin harhan epäsymmetrialle on GRM-3-mallin osioparametrien generointiprosessin tuottama ehdollisen testi-informaatiofunktion $I(\theta)$ epäsymmetria. Mallin kynnysparametrit b_1 ja b_2 generoitiin prosessilla, jossa kaikki b_1 parametrit ja 80% b_2 parametreista olivat negatiivisia kummallakin testipituudella k . Kaksiparametrisessa logistisessa mallissa osion informaatio saa suurimman arvonsa vaikeusparametrin b ympäristössä[22], ja tämä ominaisuus periytyy luokiteltuun vastausmalliin, jolloin GRM-3-mallin testi-informaatiofunktio $I(\theta)$ painottuu piilomuuttujan θ negatiiviselle puolelle. Testi-informaatiofunktion $I(\theta)$ epäsymmetria vaikuttaa osioparametrien estimoinnin keskivirheeseen $\frac{1}{\sqrt{I(\theta)}}$ [22] kasvattavasti piilomuuttujan tason θ positiivisilla arvoilla. Tästä seuraa, että piilomuuttujan todellisen jakauman ollessa vino positiiviseen suuntaan, kun $\alpha > 0$, suurempi osa populaatiosta sijoittuu tälle matalan informaation alueelle, jossa osioparametrien estimointi on epätarkempaa ja estimaattorin harha kasvaa. Vastaavasti viiden vastauskategorian luokitellussa vastausmallissa kynnysparametrit b_i jakautuivat tasaisemmin nollan ympärille, jolloin testi-informaatio on likimain symmetrinen eikä vastaavaa epäsymmetriaa esiintynyt estimaattorin harhassa.

Tämä epäsymmetrisyys on huomattavissa myös marginaalisen reliabiliteettiker-toimen estimaattorissa esiintyvässä ehdollisessa reliabiliteetissa $\rho(\theta) = \frac{I(\theta)}{I(\theta)+1}$. Kuvassa 6 esitetään $\rho(\theta)$ GRM-3- ja GRM-5-malleilla todellisilla osioparametreilla ja testipituudella $k = 10$. Kuvaaajasta huomataan, että GRM-3-mallin $\rho(\theta)$ on epäsymmetrinen piilomuuttujan tason θ suhteen, kun taas GRM-5-mallin $\rho(\theta)$ on likimain symmetrinen.

Työn tuloksissa havaittu epäsymmetria GRM-3-mallin harhassa on mahdollisesti osioparametrien generointiprosessin tuottama ilmiö, ei mallin yleinen piirre. Osioparametrien generointiprosessilla, jossa kynnysparametrit b_i jakautuisivat tasaisemmin nollan ympärille, voidaan olettaa epäsymmetrian katoavan kuten GRM-5-mallissa. Tämä voitaisiin vahvistaa jatkotutkimuksella, jossa kynnysparametrien generointiprosessia muutettaisiin GRM-3-mallissa.



Kuva 6: Ehdollinen reliabiliteetti $\rho(\theta) = I(\theta)/(I(\theta)+1)$ GRM-3- ja GRM-5-malleille testipituudella $k = 10$ todellisilla osioparametreilla. GRM-3-mallin $\rho(\theta)$ arvot on esitetty pisteissä $\theta \in [-2, 2]$ epäsymmetrian havainnollistamiseksi.

6.2 Otoksoon vaikutus estimaattorin harhaan

Havainto estimaattorin harhan itseisarvon kasvusta otoksoon kasvaessa on yllättävä, koska estimaattorin harhan odottaisi intuitiivisesti pienenevän otoksoon kasvaessa ja se käyttäytyikin näin vinousparametrin ollessa $\alpha = 0$. Tämä estimaattorin harhan itseisarvoinen kasvu otoksoon kasvaessa saattaa selittyä kahden ilmiön yhteisvaikutuksesta. Estimaattorin huomattiin harhan olevan johdonmukaisesti negatiivista. Tähän johdonmukaiseen harhaan lisätään mallin sovituksessa käytetyn suurimman uskottavuuden menetelmän tuottama harha osioparametreille, kun populaation jakauma ei vastaa sen jakaumaoletusta $N(0, 1)$. Tämä harha on suuruusluokkaa $\frac{1}{n}$ [40]. Tämän harhan oletetaan olevan positiivista ja se vaikuttaa estimaattoriin (32) testi-informaatiofunktion kautta. Otoksoon n kasvaessa rajatta tämä harha lähestyy nollaa ja estimaattorin harhaan jää vain asymptoottinen harha, jonka lähteenä on väärä jakaumaoletus reliabiliteettikertoimen estimaattorin laskennassa.

Estimaattorin harhan itseisarvon kasvun voimakkuus vaihteli mallien välillä. Vaikutus oli suurin kaksiparametrisessa logistisessa mallissa ja vähäisempi luokitellussa vastausmallissa. Tämän vaihtelun syiden selvittäminen vaatisi jatkotutkimusta, jossa mallien estimointivirheiden vaikutusta marginaalisen reliabiliteetin estimaattoriin tarkasteltaisiin erikseen.

On kuitenkin tärkeää huomata, että marginaalisen reliabiliteettikertoimen estimaattorin harhan itseisarvon kasvu otoksoon kasvaessa on sen alkuperäisellä asteikolla $[0, 1]$ kokonaisuudessaan pientä, suurimmillaan noin 0,05. Tulokset kuitenkin osoittavat, että otoksoon kasvattaminen ei korjaa väärin määritellyn jakaumaoletuksen aiheuttamaa harhaa estimaattoriin vaan paljastaa sen selkeämmin.

6.3 Testipituuden vaikutus estimaattorin harhaan

Simulaatiotutkimuksen tuloksissa havaittiin, että testipituudella $k = 40$ luokitellussa vastausmallissa marginaalisen reliabiliteetin estimaattorin harha oli lähes riippumaton vinousparametrin α arvosta. Erityisesti viiden vastauskategorian luokitellussa vastausmallissa (GRM-5) harha oli käytännössä nolla kaikilla vinousparametrin α arvoilla ja sama ilmiö oli havaittavissa GRM-3-mallissa lukuunottamatta luvussa 6.2 esitettyä harhan epäsymmetriaa.

Havainto selittyy mahdollisesti ehdollisen reliabiliteetin $\rho(\theta)$ käyttäytymisellä testipituuden k kasvaessa. Testi-informaatiofunktio $I(\theta)$ on yksittäisten osioiden informaatiofunktioiden summa, joten testipituuden k kasvaessa $I(\theta)$ kasvaa. Kun testi-informaatiofunktion $I(\theta)$ arvo on suuri samalla piilomuuttujan θ välillä, jolle populaation jakauma $g(\theta)$ keskittyy, ehdollisen reliabiliteetin $\rho(\theta)$ arvo lähestyy arvoa yksi. Tällöin marginaalisen reliabiliteettikertoimen integraalissa $\int \rho(\theta)g(\theta)d\theta$ arvo lähestyy arvoa $\int 1 \cdot g(\theta)d\theta = 1$ riippumatta jakauman $g(\theta)$ muodosta. Kun sekä populaation todellisella jakaumalla $SN(0, 1, \alpha)$ että estimaattorin oletetulla jakaumalla $N(0, 1)$ lasketut integraalit lähestyvät samaa arvoa, niiden erotuksena muodostuva virhe pienenee.

Tulos osoittaa, että pitkällä testillä ja suurella osiokohtaisella informaatiolla marginaalisen reliabiliteettikertoimen estimaattori on käytännössä robusti piilomuuttujan vinonormaalijakauman vinoudelle. Tämä on käytännön kannalta merkittävä havainto, koska estimaattorin väärin määritellyn jakaumaoletuksen vaikutus pienenee testipituuden kasvaessa ilman, että jakaumaoletusta tarvitsee erikseen korjata.

7 Yhteenveto

Tässä työssä tutkittiin Monte Carlo -simulaatiolla, miten piilomuuttujan jakauman vinous vaikuttaa marginaalisen reliabiliteettikertoimen estimaattorin harhaan. Simulaatioasetelma kattoi kaksi osiovastemallia, seitsemän vinousparametrin α arvoa, neljä otoskokoa ja kaksi testipituutta. Simulaatiossa oli 168 parametriyhdistelmää ja kussakin ehdossa suoritettiin 1000 replikaatiota. Osioparametrit generoitiin erikseen jokaiselle mallin ja testin pituuden yhdistelmälle.

Simulaation tulokset osoittivat, että estimaattorin harha oli lähes nolla tilanteissa, joissa reliabiliteettikertoimen laskemisessa käytetty piilomuuttujan jakauma vastasi todellista jakaumaa. Kaikissa parametriyhdistelmissä, piilomuuttujan jakauma oli vino, estimaattorin harha oli sen sijaan negatiivinen. Estimaattorin harhan itseisarvo kasvoi vinouden kasvaessa ja oli suurimmillaan noin 0,28. Vastauskategorioiden lukumäärän kasvattaminen ja pidempi testi käytettäessä luokiteltua vastausmallia vähensivät estimaattorin harhaa. Testipituudella $k = 40$ luokitellussa vastausmallissa estimaattorin harha oli lähes riippumaton vinousparametrin α arvosta.

Yksi työn keskeisimmistä ja samalla yllättävimmistä tuloksista oli se, että otoskoon kasvattaminen ei vähentänyt estimaattorin harhaa vaan lisäsi sitä, kun $|\alpha| \neq 0$. Tämä poikkeaa yleisestä käsityksestä, että otoskoon kasvattaminen korjaa estimaattorin ongelmia. Mahdollinen syy tälle ilmiölle on, että otoskoon kasvaessa osioparametrien estimaatit suppenevat kohti arvoja, jotka maksimoivat estimoinnissa käytetyn uskottavuusfunktion, kun standardinormaalijakaumaoletus pätee. Koska piilomuuttujan θ noudattaa todellisuudessa vinonormaalijakaumaa $SN(0, 1, \alpha)$, nämä arvot poikkeavat osioparametrien todellisista arvoista, jolloin otoskoon kasvattaminen ei korjaa estimaattorin harhaa.

Työn tutkimusasetelma tuotti rajoituksia tulosten tulkintaan ja yleistettävyyteen. Marginaalisen reliabiliteettikertoimen estimaattorin harhalle on kaksi lähdettä: mallin sovituksessa käytetyn EM-algoritmin jakaumaoletus $N(0, 1)$ sekä estimaattorin laskennassa tehty vastaava oletus. Käytetyssä tutkimusasetelmassa näitä kahta harhan lähdettä ei voitu erottaa toisistaan. Toinen tutkimusasetelman rajoitus liittyy vinonormaalijakauman ominaisuuksiin. Vinousparametrin muutos muuttaa jakauman odotusarvoa ja varianssia, jolloin havaittu harha voi heijastaa myös näiden muutosten vaikutusta eikä pelkästään jakauman vinouden vaikutusta.

Lisäksi kiinteiden osioparametrien käyttö toi mukanaan mahdollisuuden, että työn tulokset heijastavat osittain käyttöön valikoituneiden osioiden ominaisuuksia eivätkä ne välttämättä ole yleistettävissä laajempaan osiojoukkoon. Esimerkki tästä on kolmen vastauskategorian luokitellussa vastausmallissa havaittu estimaattorin harhan epäsymmetria vinousparametrin α suhteen. Todennäköinen selitys tälle harhan epäsymmetrialle on työssä käytetty osioparametrien generointiprosessi. Tämän rajoitteen ja muiden mallien tulosten perusteella on todennäköistä, että ilmiö itsessään on symmetrinen eikä mallin yleinen ominaisuus. Epäsymmetrian alkuperää voitaisiin selvittää muuttamalla osioparametrien generointiprosessia ja tarkastelemalla useampia osiokombinaatioita. Vaihtoehtoinen tutkimusasetelma olisi generoida tuhat keskenään riippumatonta osioparametrijoukkoa ja käyttää näitä kaikilla parametriyhdistelmillä siten, että kutakin replikaatiota vastaa sama osioparametrijoukko kaikissa vinous- ja otoskokoehdossa. Tällöin parametriyhdistelmien vertailukelpoi-

suus säilyisi ja tulokset kuvaisivat keskimääräistä harhan vaikutusta laajemmin.

Jatkotutkimuksessa tulisi tarkastella estimaattorin kahden harhan lähteen erotamista, mikä voitaisiin toteuttaa asetelmalla, jossa joko mallin sovitus tai estimaattorin laskenta suoritettaisiin todellisella vinonormaalijakaumalla $SN(0, 1, \alpha)$ ja toinen standardinormaalijakaumaoletuksella. Tällöin olisi mahdollista havaita kumpunkin harhan lähteen osuus kokonaisharhasta. Tämä asetelma voisi mahdollisesti tarjota tietoa myös otoskoon kasvun aiheuttamasta harhan kasvusta, jos otoskoon vaikutus liittyy ensisijaisesti jompaan kumpaan harhan lähteistä. Lisäksi vinonormaalijakauman sijainti- ja skaalaparametrit voitaisiin jatkotutkimuksessa valita siten, että jakauman odotusarvo ja varianssi vastaisivat standardinormaalijakauman vastaavia kaikilla tutkimuksessa käytettävillä vinousparametrin arvoilla. Tällöin vinouden vaikutus voitaisiin erottaa odotusarvon ja varianssin muutoksen vaikutuksesta.

Käytännön sovellustilanteissa piilomuuttujan todellista jakaumaa ei tunneta. Jatkotutkimuksessa olisi siksi aiheellista tarkastella menetelmiä, jotka estimoivat piilomuuttujan jakauman aineistosta ennalta määritellyn jakauman sijaan. Tällaisia menetelmiä ovat esimerkiksi empiirinen histogrammimenetelmä[37] ja Davidian-käyriin perustuva estimointi[41], jotka molemmat ovat toteutettuna `mirt`-paketissa [29]. Näissä menetelmissä piilomuuttujan jakauma estimoidaan samanaikaisesti osio-parametrien kanssa EM-algoritmin aikana, jolloin estimoitu jakauma voi saada lähes minkä tahansa muodon[41].

Näiden menetelmien käyttö mahdollistaisi estimaattorin harhan tarkastelun toteutuksen kahdessa vaiheessa. Ensimmäisessä vaiheessa osiovastemalli sovitettaisiin käyttäen estimoitua piilomuuttujan jakaumaa, mutta marginaalinen reliabiliteettikerroin laskettaisiin yhä käyttäen standardinormaalioletusta. Näin voitaisiin tarkastella piilomuuttujan jakauman estimoinnin vaikutusta mallin sovituksessa estimaattorin harhaan. Toisessa vaiheessa piilomuuttujan jakauman estimointi käytetään myös marginaalisen reliabiliteettikertoimen laskennassa. Näiden kahden vaiheen vertailu tämän työn tutkimusasetelman tuloksiin paljastaisi harhan lähteiden suhteet ja kuinka aineistosta estimoitu piilomuuttujan jakauma vaikuttaa vinouden aiheuttamaan estimaattorin harhaan.

Koska tämän työn tulokset osoittavat, että piilomuuttujan jakauman vinous vaikuttaa marginaalisen reliabiliteettikertoimen estimaattorin harhaan, normaalijakaumaoletusta estimaattorissa olisi syytä arvioida. Soveltaja voisi toteuttaa tämän soveltamalla halutun osiovastemallin sekä standardinormaalioletuksella että käyttäen aineistosta estimoitua piilomuuttujan jakaumaa. Tämän jälkeen mallien sopivuutta voitaisiin verrata käyttämällä esimerkiksi Hannan-Quinn -indeksiä[42]. Woods ja Lin[41] havaitsivat tämän indeksin perusteella tehdyjen päätösten tunnistavan piilomuuttujan ei-normaaliuden luotettavasti ja valitsevan standardinormaalioletuksella sovitettun mallin, kun piilomuuttujan jakauma on todellisuudessa normaali. He huomauttivat kuitenkin, että ei-normaaliuden havaitsemiseen vaikuttavat todennäköisesti testin pituus ja jakauman vinouden aste, joiden vaikutusta tulisi tarkastella jatkotutkimuksissa.

Osiovastemallinnusta sovelletaan laajasti psykologisessa ja osaamisen arvioinnissa, joissa mitattavien piilomuuttujien jakaumat eivät aina ole normaaleja. Normaalijakaumaoletuksen vaikutuksen tutkiminen auttaa soveltajaa punnitsemaan, onko

oletuksen käyttö perustelua. Tämä työ on kuitenkin vain aiheen alustava tarkastelu, ja oletuksen vaikutusta marginaaliseen reliabiliteettikertoimen estimaattoriin olisi syytä tutkia jatkossa laajemmin. Estimaattorin käyttäytymisen tarkempi tunteminen parantaa arvioita testien käyttökelpoisuudesta ja niiden pohjalta tehdyistä johtopäätöksistä.

Viitteet

- [1] *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014. ISBN: 978-0-935302-35-6.
- [2] Frederic M. Lord, Melvin R. Novick ja Allan Birnbaum. *Statistical theories of mental test scores*. Oxford: Addison-Wesley, 1968.
- [3] Ronald K. Hambleton ja Hariharan Swaminathan. *Item Response Theory*. Dordrecht: Springer Netherlands, 1985. ISBN: 978-94-017-1988-9. DOI: 10.1007/978-94-017-1988-9.
- [4] Theodore Micceri. “The unicorn, the normal curve, and other improbable creatures”. *Psychological Bulletin* 105.1 (1989), s. 156–166. ISSN: 1939-1455. DOI: 10.1037/0033-2909.105.1.156.
- [5] Andrew D. Ho ja Carol C. Yu. “Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects”. *Educational and Psychological Measurement* 75.3 (2015), s. 365–388. ISSN: 1552-3888. DOI: 10.1177/0013164414548576.
- [6] A. Azzalini. “A Class of Distributions Which Includes the Normal Ones”. *Scandinavian Journal of Statistics* 12.2 (1985), s. 171–178. ISSN: 0303-6898.
- [7] Björn Andersson ja Tao Xin. “Large Sample Confidence Intervals for Item Response Theory Reliability Coefficients”. *Educational and Psychological Measurement* 78.1 (2018), s. 32–45. ISSN: 0013-1644. DOI: 10.1177/0013164417713570.
- [8] Ronald K. Hambleton ja Russell W. Jones. “Comparison of classical test theory and item response theory and their applications to test development”. *Educational Measurement: Issues and Practice* 12.3 (1993), s. 38–47. ISSN: 1745-3992. DOI: 10.1111/j.1745-3992.1993.tb00543.x.
- [9] Howard Wainer. “Choosing Research Topics Wisely: The Irony of Item Response Theory as an Example”. *Journal of Educational and Behavioral Statistics* (2026). ISSN: 1076-9986. DOI: 10.3102/10769986261432203.
- [10] Bruno D. Zumbo. “Reliability as Projection in Operator-Theoretic Test Theory: Conditional Expectation, Hilbert Space Geometry, and Implications for Psychometric Practice”. *Educational and Psychological Measurement* 86.3 (2026), s. 433–469. ISSN: 0013-1644. DOI: 10.1177/00131644251389891.
- [11] Wim J. van der Linden. *Handbook of Item Response Theory, Volume One: Models*. Boca Raton: CRC Press, 2016. ISBN: 978-1-4665-1431-7.
- [12] Charles Spearman. “The Proof and Measurement of Association between Two Things”. *The American Journal of Psychology* 15.1 (1904), s. 72. ISSN: 00029556. DOI: 10.2307/1412159.
- [13] Alfred Binet ja Théodore Simon. “Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux”. *L’Année psychologique* 11.1 (1904), s. 191–244. DOI: 10.3406/psy.1904.3675.

- [14] Louis L. Thurstone. “A method of scaling psychological and educational tests”. *Journal of Educational Psychology* 16.7 (1925), s. 433–451. ISSN: 1939-2176. DOI: 10.1037/h0073357.
- [15] Frederic M. Lord. *A Theory of Test Scores*. Psychometric Corporation, 1952.
- [16] Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut, 1960. ISBN: 0-598-55451-3.
- [17] R. Darrell Bock. “Estimating item parameters and latent ability when responses are scored in two or more nominal categories”. *Psychometrika* 37.1 (1972), s. 29–51. ISSN: 1860-0980. DOI: 10.1007/BF02291411.
- [18] Fumiko Samejima. “Estimation of latent ability using a response pattern of graded scores”. *Psychometrika* 34.1 (1969), s. 1–97. ISSN: 1860-0980. DOI: 10.1007/BF03372160.
- [19] Fumiko Samejima. “A general model for free-response data”. *Psychometrika Monograph Supplement* 37.1 (1972), s. 68–68.
- [20] Frederic M. Lord. *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, 1980. ISBN: 0-89859-006-X.
- [21] Petteri Harjulehto, Riku Klén ja Mika Koskenoja. *Analyysiä reaaliluvuilla*. 4. painos. Helsinki: Unigrafia Oy, 2017. ISBN: 978-952-93-4162-7.
- [22] Susan E. Embretson ja Steven P. Reise. *Item Response Theory for Psychologists*. Multivariate Applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000. ISBN: 978-0-8058-2818-7.
- [23] R. A. Fisher. “On the mathematical foundations of theoretical statistics”. *Philosophical Transactions of the Royal Society of London. Series A*. 222.594-604 (1922), s. 309–368. DOI: 10.1098/rsta.1922.0009.
- [24] Wim J. van der Linden. *Handbook of Item Response Theory, Volume Two: Statistical Tools*. Boca Raton: CRC Press, 2016. ISBN: 978-1-4665-1443-0.
- [25] Edwin J.G. Pitman. *Lecture Notes on Nonparametric Statistical Inference: Lectures Given for the University of North Carolina*. 1948.
- [26] Bert F. Green et al. “Technical Guidelines for Assessing Computerized Adaptive Tests”. *Journal of Educational Measurement* 21.4 (1984), s. 347–360. ISSN: 1745-3984. DOI: 10.1111/j.1745-3984.1984.tb01039.x.
- [27] Ying Cheng, Ke-Hai Yuan ja Cheng Liu. “Comparison of Reliability Measures Under Factor Analysis and Item Response Theory”. *Educational and Psychological Measurement* 72.1 (2012), s. 52–67. ISSN: 0013-1644. DOI: 10.1177/0013164411407315.
- [28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2025.
- [29] R. Philip Chalmers. “mirt: A Multidimensional Item Response Theory Package for the R Environment”. *Journal of Statistical Software* 48.6 (2012), s. 1–29. DOI: 10.18637/jss.v048.i06.

- [30] R. Philip Chalmers. “Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications”. *Journal of Statistical Software* 71.5 (2016), s. 1–38. DOI: 10.18637/jss.v071.i05.
- [31] Adelchi Azzalini. *The R package sn: The skew-normal and related distributions such as the skew-t and the SUN*. Università degli Studi di Padova, Italia, 2026.
- [32] Davis Vaughan, Henrik Bengtsson ja Matt Dancho. *furrr: Apply Mapping Functions in Parallel using Futures*. 2026.
- [33] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4.
- [34] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. 2022.
- [35] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. 2026.
- [36] Patrick D. Manapat ja Michael C. Edwards. “Examining the Robustness of the Graded Response and 2-Parameter Logistic Models to Violations of Construct Normality”. *Educational and Psychological Measurement* 82.5 (2022), s. 967–988. ISSN: 0013-1644. DOI: 10.1177/00131644211063453.
- [37] R. Darrell Bock ja Murray Aitkin. “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm”. *Psychometrika* 46.4 (1981), s. 443–459. ISSN: 1860-0980. DOI: 10.1007/BF02293801.
- [38] Carrie R. Houts ja Michael C. Edwards. “The Performance of Local Dependence Measures With Psychological Data”. *Applied Psychological Measurement* 37.7 (2013), s. 541–562. ISSN: 0146-6216. DOI: 10.1177/0146621613491456.
- [39] Cheryl D. Hill. “Precision of parameter estimates for the graded item response model”. *University of North Carolina at Chapel Hill* (2004).
- [40] D. R. Cox ja E. J. Snell. “A General Definition of Residuals”. *Journal of the Royal Statistical Society: Series B (Methodological)* 30.2 (1968), s. 248–265. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1968.tb00724.x.
- [41] Carol M. Woods ja Nan Lin. “Item Response Theory With Estimation of the Latent Density Using Davidian Curves”. *Applied Psychological Measurement* 33.2 (2009), s. 102–117. ISSN: 0146-6216. DOI: 10.1177/0146621608319512.
- [42] E. J. Hannan. “Rational Transfer Function Approximation”. *Statistical Science* 2.2 (1987), s. 135–151. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1177013343.

Liitteet

A Tulossiossa tarkasteltujen ehtojen harhataulukot

Taulukko 3: Harha malleittain vinousparametrin α arvoilla, kun $n = 5000$.

α	$k = 10$			$k = 40$		
	2PL	GRM-3	GRM-5	2PL	GRM-3	GRM-5
-5	-0,248	-0,144	-0,136	-0,131	-0,048	-0,044
-3	-0,219	-0,127	-0,117	-0,112	-0,042	-0,037
-1	-0,090	-0,053	-0,043	-0,040	-0,017	-0,013
0	0,000	0,000	0,000	0,000	0,000	0,000
1	-0,095	-0,077	-0,044	-0,041	-0,034	-0,013
3	-0,235	-0,233	-0,119	-0,116	-0,109	-0,037
5	-0,268	-0,280	-0,139	-0,136	-0,134	-0,044

Taulukko 4: Harha otoskoottain vinousparametrin α arvoilla, kun $k = 10$.

Malli	α	$n = 250$	$n = 500$	$n = 1000$	$n = 5000$
2PL	-5	-0,207	-0,229	-0,240	-0,248
	-3	-0,185	-0,205	-0,213	-0,219
	-1	-0,075	-0,083	-0,087	-0,090
	0	0,007	0,004	0,002	0,000
	1	-0,079	-0,086	-0,092	-0,095
	3	-0,189	-0,214	-0,227	-0,235
	5	-0,211	-0,241	-0,258	-0,268
GRM-3	-5	-0,140	-0,141	-0,144	-0,144
	-3	-0,123	-0,125	-0,126	-0,127
	-1	-0,051	-0,051	-0,052	-0,053
	0	0,000	0,000	0,000	0,000
	1	-0,064	-0,074	-0,077	-0,077
	3	-	-0,222	-0,228	-0,233
	5	-	-0,266	-0,274	-0,280
GRM-5	-5	-	-0,133	-0,135	-0,136
	-3	-	-0,114	-0,116	-0,117
	-1	-0,036	-0,041	-0,043	-0,043
	0	0,001	0,001	0,000	0,000
	1	-0,039	-0,042	-0,043	-0,044
	3	-0,111	-0,116	-0,119	-0,119
	5	-0,134	-0,137	-0,139	-0,139

Taulukko 5: Harha testipituuksittain vinousparametrin α arvoilla, kun $n = 5000$.

Malli	α	$k = 10$	$k = 40$
2PL	-5	-0,248	-0,131
	-3	-0,219	-0,112
	-1	-0,090	-0,040
	0	0,000	0,000
	1	-0,095	-0,041
	3	-0,235	-0,116
	5	-0,268	-0,136
GRM-3	-5	-0,144	-0,048
	-3	-0,127	-0,042
	-1	-0,053	-0,017
	0	0,000	0,000
	1	-0,077	-0,034
	3	-0,233	-0,109
	5	-0,280	-0,134
GRM-5	-5	-0,136	-0,044
	-3	-0,117	-0,037
	-1	-0,043	-0,013
	0	0,000	0,000
	1	-0,044	-0,013
	3	-0,119	-0,037
	5	-0,139	-0,044

B Simulaatiossa käytetyt kiinteät osioparametrit

Taulukko 6: 2PL-mallin kiinteät osioparametrit, $k = 10$.

Osio	a	b
1	1,346	-0,408
2	0,866	-0,730
3	1,232	-0,221
4	1,165	-0,226
5	0,990	-2,547
6	0,435	1,347
7	0,969	0,616
8	0,884	0,218
9	1,224	-0,805
10	1,048	0,690

Taulukko 7: GRM-3-mallin kiinteät osioparametrit, $k = 10$.

Osio	a	b_1	b_2
1	2,065	-1,747	-0,366
2	1,632	-2,056	-0,629
3	1,999	-1,219	-0,173
4	1,231	-0,759	0,420
5	1,349	-2,026	-1,374
6	1,258	-2,097	-1,003
7	1,596	-1,234	-0,342
8	1,487	-1,599	-0,633
9	1,616	-0,915	0,196
10	2,387	-2,034	-0,828

Taulukko 8: GRM-5-mallin kiinteät osioparametrit, $k = 10$.

Osio	a	b_1	b_2	b_3	b_4
1	1,611	-1,156	-0,086	1,038	2,275
2	1,580	-1,552	-0,601	0,671	1,541
3	2,159	-1,817	-0,850	-0,174	1,083
4	1,632	-0,749	0,425	1,441	2,549
5	1,597	-1,777	-0,656	0,447	1,452
6	1,777	-1,193	-0,008	0,948	2,181
7	1,504	-1,569	-0,478	0,357	1,284
8	1,663	-1,982	-1,143	-0,345	0,582
9	1,706	-0,654	0,246	1,503	2,501
10	1,966	-1,865	-1,108	-0,150	0,830

Taulukko 9: 2PL-mallin kiinteät osioparametrit, $k = 40$.

Osio	a	b
1	1,091	-0,141
2	1,141	0,167
3	0,772	-0,199
4	1,630	-0,291
5	1,204	-1,734
6	1,762	-0,273
7	1,709	-1,800
8	1,207	1,153
9	1,384	-1,003
10	1,708	0,148
11	1,031	0,519
12	1,240	0,005
13	1,114	1,347
14	1,290	-0,847
15	1,245	0,443

Osio	a	b
16	1,539	-0,977
17	1,368	2,124
18	0,922	0,688
19	1,363	-0,343
20	0,943	0,785
21	0,843	-1,156
22	1,423	1,509
23	0,829	-1,155
24	1,282	-1,557
25	0,940	-0,055
26	1,615	0,849
27	1,404	-0,011
28	1,069	-0,760
29	1,430	1,176
30	1,318	2,414
31	0,839	-1,959
32	1,128	1,474
33	0,911	-0,475
34	1,325	0,981
35	0,997	-1,824
36	1,121	-0,261
37	0,819	-0,959
38	0,902	-0,490
39	1,230	-1,077
40	0,890	0,369

Taulukko 10: GRM-3-mallin kiinteät osioparametrit, $k = 40$.

Osio	a	b_1	b_2
1	1,686	-0,181	0,645
2	1,877	-1,601	-0,445
3	1,832	-1,550	-0,673
4	1,372	-1,248	-0,314
5	1,985	-1,309	-0,380
6	1,811	-1,421	-0,265
7	1,954	-2,142	-1,020
8	2,047	-1,955	-0,739
9	1,800	-1,112	0,101
10	1,771	-2,280	-0,848
11	1,707	-1,333	-0,319
12	1,997	-1,250	-0,757
13	1,873	-0,929	-0,037
14	1,488	-1,809	-0,964
15	1,698	-1,331	-0,390

Osio	a	b_1	b_2
16	2,122	-1,933	-0,702
17	1,438	-1,113	-0,227
18	1,579	-1,608	-0,769
19	1,882	-1,807	-0,709
20	1,483	-1,628	-0,488
21	1,582	-1,843	-0,892
22	1,963	-1,741	-0,658
23	1,721	-1,431	-0,332
24	1,142	-1,327	-0,253
25	2,150	-1,507	-0,621
26	1,467	-0,844	0,303
27	1,778	-0,993	-0,447
28	1,765	-1,193	-0,070
29	2,134	-1,171	-0,090
30	1,813	-1,837	-0,427
31	1,874	-1,707	-0,654
32	2,354	-1,329	-0,306
33	1,461	-0,679	0,560
34	1,562	-2,293	-1,147
35	1,387	-1,479	-0,480
36	1,570	-1,271	-0,275
37	1,631	-2,428	-1,401
38	1,613	-0,616	0,170
39	1,567	-1,794	-0,547
40	1,664	-0,673	0,219

Taulukko 11: GRM-5-mallin kiinteät osioparametrit, $k = 40$.

Osio	a	b_1	b_2	b_3	b_4
1	1,655	-1,591	-0,562	0,625	1,695
2	1,443	-1,021	0,031	1,072	2,286
3	1,448	-1,880	-1,141	0,149	1,332
4	1,440	-0,670	0,428	1,172	2,315
5	1,867	-1,199	-0,302	0,682	1,893
6	1,109	-1,690	-1,022	0,198	1,268
7	2,047	-0,805	0,319	1,181	1,794
8	1,097	-2,668	-1,665	-1,010	-0,170
9	1,289	-2,187	-1,333	-0,220	1,248
10	1,921	-1,273	-0,218	0,663	2,047
11	2,233	-1,403	-0,774	0,128	1,106
12	1,597	-0,694	0,722	1,876	2,373
13	1,933	-1,148	-0,620	0,803	1,651
14	2,022	-1,607	-0,542	0,467	1,166
15	1,809	-0,958	-0,004	1,186	1,850

Osio	a	b_1	b_2	b_3	b_4
16	1,114	-1,492	-0,728	-0,153	1,037
17	1,439	-1,950	-0,884	0,134	1,159
18	1,467	-1,319	-0,401	0,251	1,539
19	1,695	-2,059	-0,754	-0,095	0,640
20	1,876	-1,257	-0,478	0,804	1,603
21	1,842	-1,319	-0,052	1,088	2,216
22	2,065	-1,837	-0,726	0,389	1,621
23	1,520	-1,132	-0,343	0,550	1,515
24	1,881	-0,946	0,016	0,760	1,567
25	1,697	-1,197	-0,379	0,707	2,077
26	2,037	-1,853	-1,253	-0,479	0,858
27	2,229	-1,865	-0,483	1,001	2,212
28	1,210	-1,562	-0,248	0,767	1,930
29	1,761	-1,842	-0,814	0,250	1,011
30	1,395	-1,836	-1,120	-0,436	0,159
31	2,037	-2,550	-1,448	-0,344	1,081
32	1,073	-2,146	-1,343	-0,116	0,944
33	2,022	-1,053	0,067	1,129	1,964
34	1,447	-1,477	-0,303	0,708	2,013
35	1,887	-1,257	-0,220	1,093	1,971
36	1,860	-1,401	-0,415	0,729	1,637
37	1,798	-2,220	-1,005	0,122	1,211
38	2,334	-1,592	-0,378	0,835	1,556
39	1,848	-2,181	-1,096	-0,028	0,872
40	1,590	-1,122	-0,147	1,482	2,270

C Marginaalisen reliabiliteettikertoimen estimaattorin harha populaatiotasolla

Taulukko 12: Marginaalisen reliabiliteettikertoimen (MRC) harha todellisilla osioparametreilla kaikissa simulaatiotutkimuksen parametrijhdistelmissä.

Malli	k	α	Pop. MRC	Est. MRC	Harha
2PL	10	-5	0,6749	0,6525	-0,0224
2PL	10	-3	0,6741	0,6525	-0,0216
2PL	10	-1	0,6683	0,6525	-0,0158
2PL	10	0	0,6525	0,6525	0,0000
2PL	10	1	0,6368	0,6525	0,0158
2PL	10	3	0,6309	0,6525	0,0216
2PL	10	5	0,6301	0,6525	0,0224
2PL	40	-5	0,9003	0,8985	-0,0018
2PL	40	-3	0,9003	0,8985	-0,0018
2PL	40	-1	0,8997	0,8985	-0,0012

Malli	k	α	Pop. MRC	Est. MRC	Harha
2PL	40	0	0,8985	0,8985	0,0000
2PL	40	1	0,8973	0,8985	0,0012
2PL	40	3	0,8967	0,8985	0,0018
2PL	40	5	0,8967	0,8985	0,0018
GRM3	10	-5	0,8738	0,7904	-0,0834
GRM3	10	-3	0,8716	0,7904	-0,0812
GRM3	10	-1	0,8524	0,7904	-0,0620
GRM3	10	0	0,7904	0,7904	-0,0000
GRM3	10	1	0,7284	0,7904	0,0620
GRM3	10	3	0,7092	0,7904	0,0812
GRM3	10	5	0,7070	0,7904	0,0834
GRM3	40	-5	0,9673	0,9327	-0,0345
GRM3	40	-3	0,9666	0,9327	-0,0339
GRM3	40	-1	0,9602	0,9327	-0,0274
GRM3	40	0	0,9327	0,9327	0,0000
GRM3	40	1	0,9053	0,9327	0,0274
GRM3	40	3	0,8988	0,9327	0,0339
GRM3	40	5	0,8982	0,9327	0,0345
GRM5	10	-5	0,8915	0,8907	-0,0008
GRM5	10	-3	0,8914	0,8907	-0,0007
GRM5	10	-1	0,8911	0,8907	-0,0004
GRM5	10	0	0,8907	0,8907	0,0000
GRM5	10	1	0,8903	0,8907	0,0004
GRM5	10	3	0,8900	0,8907	0,0007
GRM5	10	5	0,8899	0,8907	0,0008
GRM5	40	-5	0,9699	0,9699	0,0000
GRM5	40	-3	0,9699	0,9699	0,0000
GRM5	40	-1	0,9698	0,9699	0,0001
GRM5	40	0	0,9699	0,9699	-0,0000
GRM5	40	1	0,9700	0,9699	-0,0001
GRM5	40	3	0,9699	0,9699	-0,0000
GRM5	40	5	0,9699	0,9699	-0,0000