Admissible statistics from a latent variable perspective

Zand Scholten, A.

Publication date
2011

Dutch Summary/Samenvatting

In dit proefschrift wordt nagegaan hoe de invloed van het meetniveau van psychologische variabelen op de legitimiteit van inferenties over psychologische eigenschappen kan worden bezien vanuit een latente variabele modelbenadering. De algemene vraag is hoe latente variabele modellen zich verhouden tot, en in verband kunnen worden gebracht met representationele meetheorie (RMT), waarin formeel beschreven is aan welke voorwaarden moet worden voldaan om van kwantitatieve meting te kunnen spreken. Van veel psychologische variabelen wordt aangenomen dat de achterliggende, latente eigenschap een kwantitatieve structuur heeft. In latente variabele modellen wordt deze aanname soms impliciet, soms expliciet gemaakt, maar hij wordt bijna nooit getoetst. Aangezien veel statistische technieken alleen tot legitieme, consistente conclusies leiden als ze zijn gebaseerd op kwantitatieve metingen, lijkt het toetsen van deze aanname zeer belangrijk.

Meten en toegestane statistiek

In Hoofdstuk 2 van dit proefschrift wordt een belangrijk punt van kritiek op de theorie van meetniveaus en toegestane statistiek besproken. In deze theorie (Stevens, 1946) wordt gesteld dat er verschillende meetniveaus (nominaal, ordinaal, interval en ratio) zijn die corresponderen met een steeds rijkere structuur (onderscheidbaarheid, ordening, vergelijkbaarheid van ver-

schillen tussen objecten en directe vergelijkbaarheid van objecten). Meetniveaus kunnen worden gedefinieerd aan de hand van het type transformatie dat op de getallen kan worden uitgevoerd zonder dat de structuur verloren gaat die in eerste instantie door de getallen werd gevangen. Gekoppeld hieraan is een beperking in de toegestane wiskundige bewerking van de getallen. Een statistische toets die onder een structuurbehoudende, toegestane transformatie verandert is ambigu en ongewenst. De keuze voor een bepaalde statistische toets moet daarom afhangen van het meetniveau van de geobserveerde data.

Lord (1953) zette dit laatste deel van de theorie op losse schroeven door middel van een gedachte-experiment waarin een niet-toegestane t-toets werd uitgevoerd op nominale american footballnummers, om te bepalen of de nummers random uit een van tevoren bekende populatie getrokken waren. Deze toets leidde tot een zinvolle conclusie – namelijk dat er met de machine die de nummers uitgaf is geknoeid. De impliciete boodschap is dat een niet-toegestane toets wel degelijk informatief kan zijn en het gebruik ervan dus niet moet worden beperkt.

Dit argument heeft geleid tot een heftig debat over de relatie tussen meten en statistiek dat een aantal keer opnieuw is opgelaaid (Gaito, 1980; Townsend & Ashby, 1984; Velleman & Wilkinson, 1993). Lord's argument is echter niet waterdicht. In Hoofdstuk 2 wordt beargumenteerd dat de aanname onjuist is dat de nummers nominaal zijn en alleen de eigenschap van onderscheidbaarheid van de spelers representeren. Wanneer de nummers als representaties van een eigenschap van de machine en niet van de spelers wordt gezien, namelijk de interval eigenschap van bias naar lage nummers, dan is de toets een schoolvoorbeeld van een toegestane toets. De eigenschap kan zelfs gerepresenteerd worden volgens de strenge axioma's van RMT. Lord (1954) gaf zelf aan dat het beperken van het gebruik van statistische toetsen naar gelang het meetniveau van de geobserveerde scores wel belangrijk is als er een inferentie naar een achterliggende eigenschap wordt gemaakt. Nu de obscure, achterliggende eigenschap van bias is geïdentificeerd kunnen we het gedachte-experiment definitief als een invalide argument bestempelen.

Hoewel het 'meten versus statistiek' debat nooit definitief is beëindigd, is het wel langzaam verstomd nadat RMT ten tonele verscheen. RMT is een volledig geformaliseerde versie van Stevens' theorie, waarin meten wordt gedefinieerd als de mogelijkheid om een isomorfisme (structuurbehoudende

relatie) aan te tonen tussen een empirische relationeel systeem (objecten of personen die een bepaalde eigenschap vertonen) en een numeriek relationeel systeem. Een set axioma's specificeert of het mogelijk is om een numerieke representatie voor een bepaald systeem te vinden. Voor verschillende empirische systemen is bewezen wat de relatie is tussen alle mogelijke schalingen van het numerieke relationele systeem, oftewel, wat het bijbehorende meetniveau van zo'n systeem is. Het concept toegestane statistiek komt in RMT terug in de vorm van 'betekenisvolheid'. Hierbij ligt de nadruk op de invariantie van de waarheidswaarde van de conclusie die voortkomt uit een statistische toets, niet op de invariantie van de waarde van de toetsingsgrootheid zelf. Hiermee lijkt voor statistici althans de angel uit het debat te zijn verwijderd.

RMT lijkt een welkome formalisering van ons begrip van meten en kwantiteit, die ons beter in staat stelt om kwantitatieve eigenschappen als zodanig te herkennen. Men zou verwachten dat deze theorie en de bijbehorende axioma's die ons in principe in staat stellen het meetniveau empirisch vast te stellen, omarmd zouden worden door psychologisch onderzoekers. Vooralsnog is de theorie echter alleen bekend onder meettheoretici en een kleine groep psychometrici, (Hand, 2004; Embretson & Reise, 2000; Bond & Fox, 2007). RMT is nooit gemeengoed geworden in het bredere onderzoeksveld. Leerboeken behandelen meetniveaus en de bijbehorende beperkingen aan statistische toetsen nog steeds in termen van Stevens, zonder zijn informele concept van representatie en betekenisvolheid van conclusies uit te leggen, laat staan dat de verbeterde versies van deze concepten, geformaliseerd in RMT worden besproken (Tabachnick & Fidell, 2007; Agresti & Franklin, 2008). De redenering achter legitieme inferentie en de invloed van meetniveau hierop wordt niet besproken, de lezer wordt aangeraden om voor Likert-items en soortgelijke meetinstrumenten aan te nemen dat het interval niveau is bereikt. De noodzaak om empirisch aan te tonen welk meetniveau met een meetmethode kan worden geassocieerd wordt daarmee genegeerd.

Aan de andere kant kan men zich afvragen wat de praktische consequenties van deze aanpak zijn. In Lord's gedachte-experiment resulteerde een niet-toegestane toets in een nuttig resultaat; Het resultaat was echter nuttig omdat het in feite een interval eigenschap, namelijk bias in de machine, betrof. Een alternatief moraal in het verhaal van Lord zou dus kunnen zijn dat we ons niet zozeer moeten bezighouden met de legitimiteit van de toets, maar met het zoeken naar resultaten op basis van niet-toegestane toetsen,

die bij herhaalde replicatie met verschillende instrumenten steeds consistent optreden. Hoewel dit geen direct bewijs is dat er sprake is van kwantitatieve structuur, vormen dergelijke resultaten toch een aanknopingspunt om naar kwantitatieve structuur op zoek te gaan. Of we nu besluiten om ons wel of niet te beperken tot het uitvoeren van toegestane toetsen, de algemene conclusie is in beide gevallen dat er meer aandacht uit zou moeten gaan naar het onderzoeken van de structuur van psychologische eigenschappen en het toetsen van onze aannames over deze structuur.

Het Rasch Model als Additief Conjunct Meten

In Hoofdstuk 3 wordt de bruikbaarheid van een specifiek latent variabele model, namelijk het Rasch model, besproken om het meetniveau van psychologische variabelen te bepalen. Het Rasch model is volgens sommigen een vorm van Additief Conjunct Meten (ACM). ACM is een meetstructuur gedefinieerd door RMT, welke op indirecte wijze een additieve, interval representatie van een empirische structuur mogelijk maakt voor twee eigenschappen die samen een derde eigenschap bepalen. In het Rasch model bepalen item moeilijkheid en persoonsvaardigheid samen de kans dat iemand een item goed beantwoordt. Om te kunnen beoordelen of het Rasch model ons kan helpen om kwantitatieve eigenschappen te identificeren, worden verschillende punten van kritiek besproken op de stelling dat het Rasch model interval niveau metingen garandeert.

Om deze kritiek te kunnen bespreken is het eerst van belang om duidelijk te maken wat er wordt bedoeld met kwantitatieve of interval metingen. Volgens RMT vereist het meten van een eigenschap dat bepaalde kwalitatieve relaties opgaan voor een verzameling objecten die de eigenschap vertonen en dat het empirisch relationele systeem van objecten en relaties isomorf is met een numeriek relationeel systeem. Een eigenschap is kwantitatief als de numerieke afbeelding uniek is tot affiene transformatie of vermenigvuldiging met een constante. RMT is echter niet de enige theorie die zich aan een definitie van meten waagt (Michell, 1997). Een alternatief is de traditionele meettheorie waarin meten beschouwd wordt als het ontdekken van ratio's tussen grootheden. Getallen zijn geen mathematische constructen maar instantiëringen van ratio's en daarmee empirische entiteiten. Welke van deze theorieën men aanhangt heeft echter geen directe praktische consequenties

voor zover de discussie hier reikt, aangezien beide theorieën de empirische demonstratie vereisen van de axioma's zoals geformuleerd in RMT.

De vraag blijft daarom of aan deze axioma's van ACM wordt voldaan door data die goed kan worden beschreven met het Rasch model en of dit daadwerkelijk metingen op het interval niveau oplevert. Het eerste deel van de vraag is weinig controversieel. Geen enkele criticaster heeft geageerd tegen de stelling dat het Rasch model structureel equivalent is aan ACM. Het model is speciaal ontwikkeld om de eigenschap van specifieke-objectiviteit te hebben, welke correspondeert met het 'single cancellation' axioma in ACM. Specifieke-objectiviteit houdt in dat de persoonsvaardigheid onafhankelijk van de itemmoeilijkheden is te bepalen en andersom. Tegemoetkoming aan een ander belangrijk axioma, dat van 'double cancellation', is gegarandeerd vanwege de keuze voor de logistische functie om het verschil in persoonsvaardigheid en itemmoeilijkheid om te zetten in de kans om een item goed te beantwoorden. Kritiek op de claim van interval niveau metingen moet daarom niet gericht worden op de structurele equivalentie met ACM, maar op de verschillende interpretaties van de concepten 'objecten' en 'empirische demonstratie' in RMT.

In RMT moet de verzameling objecten die de te meten eigenschap vertonen duidelijk gedefinieerd zijn. Helaas is het onmogelijk voor de meeste psychologische eigenschappen die zich laten meten met behulp van test- of vragenlijst items, om van tevoren onomstotelijk vast te stellen wat de eigenschappen van een valide item zijn. Een Rasch model past vaak pas goed op data nadat slecht passende items een voor een verwijderd zijn, waarbij het onduidelijk is waarom deze items gebrekkige passing vertonen. Dit resulteert in ad hoc, arbitraire tests of vragenlijsten die een kunstmatig gecreeërde kwantitatieve structuur vertonen. Het ligt voor de hand dat wanneer er genoegen wordt genomen met dergelijke schijn-kwantitativiteit zonder dat de reden van gebrekkige passing wordt onderzocht, er weinig vooruitgang zal worden geboekt op het terrein van kwantitatieve meting van psychologische variabelen.

Een ander probleem betreft de onmogelijkheid voor veel psychologische variabelen om steeds preciezer onderscheid te maken tussen itemmoeilijkheden. Als een eigenschap werkelijk een kwantitatieve structuur heeft dan moet het in principe mogelijk zijn om altijd een object te vinden tussen twee andere objecten, hoe dicht deze twee ook bij elkaar liggen. Hoe dit mogelijk

zou moeten zijn voor eigenschappen als extraversie of rekenvaardigheid is onduidelijk.

De empirische status van kansen is een ander problematisch onderwerp wanneer we het Rasch model als vorm van ACM zien. RMT vereist dat empirisch wordt gedemonstreerd dat de objecten die de te meten eigenschap vertonen aan bepaalde axioma's voldoen. Dit betekent dat de objecten en hun onderlinge relatie op een of andere manier moet kunnen worden geobserveerd. Kyngdon (2008a) stelt dat de kans om een item goed te beantwoorden niet aan deze eis van observeerbaarheid voldoet. Het is de vraag of een strikte representationalist kansen zoals beschreven in het Rasch model niet zou accepteren, aangezien RMT niet helder beschrijft wat onder 'observeerbaar' moet worden verstaan. De kansen in het Rasch model kunnen door een meer tastbaar kenmerk worden vervangen, bijvoorbeeld door de proportie goed beantwoorde items na herhaalde afname of de proportie goed beantwoorde items van dezelfde moeilijkheid. Wellicht is een dergelijke interpretatie in termen van proporties of kansen als benaderingen van proporties wel acceptabel voor representationalisten.

Ofschoon de stelling dat het Rasch model interval metingen oplevert dus in twijfel kan worden getrokken op basis van wetenschapsfilosofische gronden, lijken de meer praktische bezwaren dat weinig psychologische variabelen een kwantitatieve structuur hebben en het gevaar dat itemselectie tijdens modelpassing de illusie van kwantitatieve structuur wekt veel grotere problemen op te leveren.

De Guttman-Rasch paradox

In Hoofdstuk 3 werden verschillende punten van kritiek besproken op de stelling dat het Rasch model een vorm van ACM is en dus metingen op interval niveau oplevert. In Hoofdstuk 4 wordt een andere vorm van kritiek op deze stelling besproken. Dit punt van kritiek betreft het ogenschijnlijk paradoxale verschil in meetniveau dat kan worden geassocieerd met het Rasch model en het sterk gerelateerde Guttman model. Het Guttman model kan worden gezien als het Rasch model, ontdaan van meetfout. Wanneer de precisie (discriminatie) in het Rasch model wordt opgevoerd verandert dit probabilistische model in de limiet in het deterministische Guttman model, dat

is geassocieerd met een ordinaal meetniveau. De claim dat het Rasch model interval metingen oplevert moet volgens sommigen betwijfeld worden omdat het verlies van meetniveau, veroorzaakt door het verhogen van de precisie en verwijderen van meetfout, paradoxaal is (Michell, 2008a, 2009).

Een nadere analyse van dit argument laat echter zien dat de toevoeging van meetfout niet de essentiële factor is die voor een hoger, interval niveau zorgt. Er zijn verschillende manieren om meetfout aan het Guttman model toe te voegen die geen hoger meetniveau ten gevolge hebben. Voor beide modellen is nagegaan waarom ze precies wel en niet voldoen aan de axioma's van ACM. Hieruit komt naar voren dat het continue versus het discrete karakter van de 'meetfout' en de directe relatie tussen de grootte van de meetfout en de onderliggende eigenschap zijn die leiden tot een hoger meetniveau in het Rasch model.

Dit neemt niet weg dat meetfout een belangrijke rol speelt bij het bereiken van het interval niveau in het Rasch model. Het argument dat meetfout niet zou horen te leiden tot een toename in precisie lijkt overtuigend. Dit argument is echter misleidend geformuleerd in termen van precisie, niet meetniveau. Afgezien van het feit dat niet de meetprecisie maar het niveau in twijfel wordt getrokken, is het argument, zoals geformuleerd, niet juist. Het toevoegen van meetfout aan een meetprocedure kan wel degelijk leiden tot een nauwkeuriger waarneming. Dit fenomeen staat bekend als stochastische resonantie in onder andere de biologie en de natuurkunde (Gammaitoni et al., 1989; Chatterjee & Robert, 2001; Ries, 2007).

Met een simpele simulatie van een gedachte-experiment waarbij lengte wordt gemeten met behulp van gepaarde vergelijkingen wordt getoond hoe meetfout die samenhangt met de latente eigenschap tot hogere nauwkeurigheid kan leiden. Met een tweede simulatie wordt geïllustreerd hoe dit effect op kan treden onder het Rasch model, namelijk wanneer items qua moeilijkheid ongelijk verdeeld zijn. Aangezien de grootte van de meetfout direct gerelateerd is aan het verschil tussen persoonsvaardigheid en itemmoeilijkheid staat het verwijderen van meetfout gelijk aan het verwijderen van waardevolle informatie.

Het verschil in meetniveau tussen het Guttman en Rasch model moet niet als paradoxaal worden gezien. Het Rasch model maakt gebruik van continue kansen die gerelateerd zijn aan de onderliggende eigenschap; het

Guttman model maakt gebruik van discrete kansen van nul en een, die niet onderscheiden kunnen worden van de ruwe scores en daarom geen extra informatie bevatten. Een verschil in precisie – laat staan meetniveau – tussen deze modellen is niet paradoxaal aangezien het in feite twee zeer verschillende meetmethoden betreft, die meer van elkaar verschillen dan alleen in de aanwezigheid van meetfout, zoals de criticasters het doen voorkomen.

Inschatten van risico op inferentie-fouten met IRT

Zelfs als we accepteren we dat het Rasch model een vorm van ACM is en zo interval metingen oplevert, dat geschatte kansen metingen kunnen vormen en dat er een eindeloze hoeveelheid mensen en items beschikbaar is om een bepaalde eigenschap te meten, dan nog is het onwaarschijnlijk dat het Rasch model gebruikt kan worden om voor meer dan enkele eigenschappen een interval niveau vast te stellen. Weinig psychologische eigenschappen hebben een structuur die oneindige verfijning tussen nabijgelegen objecten toelaat. Andere, meer complexe IRT modellen vertonen geen structurele gelijkenis met ACM. Deze modellen zouden in eerste instantie dus van nog minder nut lijken om aannames over het meetniveau te onderzoeken. Hoewel ze niet direct bruikbaar zijn om dergelijke aannames te toetsen, kunnen deze modellen wel ingezet worden om op indirecte wijzen vragen omtrent het meetniveau en de legitimiteit van inferenties te beantwoorden.

In veel gevallen nemen we aan dat een onderliggende eigenschap kwantitatief is terwijl de geobserveerde scores ten hoogste ordinaal zijn. De relatie tussen deze ordinale data en de latente trek wordt beschreven door een onbekende monotoon stijgende functie. Latente variabele modellen, in dit geval IRT modellen, kunnen behulpzaam zijn bij het inschatten van het risico op het maken van een verkeerde inferentie op basis van niet-toegestane statistiek. In IRT modellen worden parameters gespecificeerd die inhoudelijke eigenschappen van items en personen vertegenwoordigen en het daarom mogelijk maken om een bepaalde, theoretisch plausibele relatie te veronderstellen tussen geobserveerde en latente variabelen.

Hierdoor is het mogelijk om de risicofactoren om een foute inferentie te maken meer nauwkeurig in kaart te brengen. In Hoofdstuk 5 en 6 worden het Twee-Parameter Logistisch (2PL) model en het Graded-Response Model (GRM) gebruikt om effecten op het latente niveau om te zetten in geob-

serveerde scores om het risico op inferentie-fouten in factoriëel onderzoek met interactie-effecten te bepalen.

Deze effecten zijn gevoelig voor inferentie-fouten veroorzaakt door de arbitraire keuze van meetschaal wanneer data ordinaal zijn. Bij standaard vergelijkingen van twee groepen is het onmogelijk om een inferentie-fout te maken als aan de aannamen van de statistische toets – onafhankelijke, normaal-verdeelde steekproeven met gelijke varianties – is voldaan (Davison & Sharma, 1988). bij simultane vergelijkingen van groepen op twee of meer factoren geldt dit echter niet (Davison & Sharma, 1990, 1994).

In eerder simulatie onderzoek met behulp van het Rasch en 2PL model is aangetoond dat extreme test moeilijkheid, hoge item discriminatie, testlengte en effectgrootte het aantal significante interacties deed toenemen dat werd gevonden in data gegenereerd met een model dat geen interacties bevatte (Embretson, 1996; Kang & Waller, 2005). Deze simulaties werden gedaan aan de hand van Moderated Multiple Regression (MMR) opzet met random factoren en een grote steekproef. Onder zulke omstandigheden kan het risico van inferentie-fouten tegengegaan worden door latente trekwaarden te schatten met behulp van een IRT model en deze te analyseren in plaats van de geobserveerde scores. Voor typisch experimentele onderzoeksopzetten met kleine steekproeven is het niet mogelijk om een IRT model te passen. Om het risico op inferentïele fouten in te schatten voor typisch experimentele omstandigheden werd in Hoofdstuk 5 en 6 een twee-bij-twee fixed factors design met kleine steekproefgrootte, kleine tot middelgrote effecten en verschillende typen hoofd- en interactie-effecten op het latente niveau gesimuleerd.

Het effect van extreme test moeilijkheid, item discriminatie en steekproefgrootte is allereerst onderzocht aan de hand van het 2PL model (Hoofdstuk 5). Dit model is een uitbreiding van het Rasch model waarbij items verschillend discrimineren tussen personen op de latente trek. Uit de resultaten blijkt dat een te moeilijke (of makkelijke) test een bodem-effect tot gevolg heeft op de geobserveerde schaal, wat de oorspronkelijke interactie-effecten op latent niveau verstoort en tot inferentie-fouten leidt. Hoge item discriminatie verergert dit risico. Dit geldt echter alleen voor een beperkt aantal effecten. Alleen als er twee hoofdeffecten zonder interactie zijn, of een ordeonafhankelijke interactie op het latente niveau aanwezig zijn is het risico op verkeerde conclusies groot. Een aantal andere interacties, namelijk partiële

en minimale, kunnen dit risico ook lopen vanwege sampling error. voor deze effecten is het risico wel vele malen kleiner. Steekproefgrootte tenslotte, vervormde de resultaten meer wanneer er op latent niveau geen interactie aanwezig dan wanneer deze wel aanwezig was.

Toetsing van de aannamen van normaliteit en homogeniteit van de varianties werden beide vaker verworpen door het bodem-effect veroorzaakt door de niet-lineaire transformatie volgens het 2PL model. De frequentie waarmee dit gebeurde was echter gelijk voor verschillende typen effecten. Hoewel relatief zeldzaam, kwam het een aantal keer voor dat verkeerde conclusies over de aanwezigheid van een interactie getrokken werden terwijl zowel normaliteit als homogeniteit niet verworpen konden worden. In de experimentele situatie met fixed effects en een kleine steekproef loont het dus om met het toetsen van de statistische aannames verkeerde conclusies te elimineren, maar hierbij is het risico een valide conclusie te elimineren groot en de methode is niet waterdicht.

Een betere manier om het maken van inferentiële fouten te voorkomen is om een test te gebruiken die een gepaste moeilijkheidsgraad heeft voor de steekproef. Als dit niet mogelijk is kan een normaliserende transformatie het risico op een inferentie-fout tenminste verkleinen. Een dergelijke transformatie kan gezien worden als een poging om de latente waarden te benaderen zonder de trek daadwerkelijk met een formeel model te beschrijven en te schatten.

Deze en eerdere resultaten hebben betrekking op binair gescoorde items. Een logische uitbreiding is om te zien of hetzelfde risico op inferentie-fouten optreedt wanneer er meer antwoordcategorieën beschikbaar zijn. Aangezien van dergelijke Likert-schaal items wordt gezegd dat ze het interval niveau voldoende benaderen om dit meetniveau aan te nemen, verwachten we een kleiner risico voor deze items. In Hoofdstuk 6 werd deze verwachting bevestigd in een simulatie-onderzoek waarbij latente effecten omgezet werden naar geobserveerde scores met behulp van het GRM. De opzet was hetzelfde, met als enig verschil dat alleen een kleine steekproef werd onderzocht en verschillende aantallen antwoordcategorieën werden gesimuleerd.

De resultaten waren vergelijkbaar maar minder extreem dan de resultaten gebaseerd op de data die werden gesimuleerd aan de hand van het binaire 2PL model. Naarmate het aantal antwoordcategorieën steeg nam

het aantal inferentie-fouten onder risicovolle omstandigheden (extreme test moeilijkheid, hoge itemdiscriminatie, orde-onafhankelijke effecten) af. Opnieuw bleek een normaliserende transformatie een geschikte manier om het aantal verkeerde inferenties nog verder te doen afnemen. Voor hogere aantallen antwoordcategorieën (6,7) leidde dit tot zeer acceptabele risico-niveaus.

Natuurlijk hangen al deze conclusies af van de aanname dat een kwantitatieve eigenschap aan de scores ten grondslag ligt, dat de geobserveerde scores daadwerkelijk ordinaal zijn en dat de data gegenereerd zijn aan de hand van een proces dat dezelfde structuur oplevert als het gebruikte IRT model. Of deze aannamen plausibel zijn vormt het onderwerp voor een heel nieuw proefschrift. Als we de modellen in ieder geval voorlopig accepteren als werkbaar, dan staan ze ons toe om een beter inschatting te maken van het risico op inferentie-fouten en het identificeren van de factoren die dit risico verhogen wanneer we toetsen uitvoeren die strikt genomen niet zijn toegestaan.