Psychometrie voor psychologen

Sijtsma, K.

Published in:
Werken en laten werken. Bijdragen vanuit de arbeids- en organisatiepsychologie

Publication date:
2000

19 Psychometrie voor psychologen: over de betekenis van de item-respons-theorie voor de psychologische test

Klaas Sijtsma

In de jaren zestig, toen Drenth zijn boek over testtheorie (Drenth, 1965a) en zijn eerste tests (Drenth, 1965b; Drenth & Van Wieringen, 1969; Drenth & Hoolwerf, 1970) publiceerde, was de psychologische test een pen-en-papier standaardtest. Bij een dergelijke test krijgt elk individu dezelfde verzameling van items voorgelegd. Afhankelijk van het niveau van het geteste individu kunnen sommige items voor hem/haar te gemakkelijk of te moeilijk zijn en dit kan de nauwkeurigheid van de meting nadelig beïnvloeden. De gangbare testtheorie was de klassieke testtheorie (KTT; Lord & Novick, 1968), die gericht is op de evaluatie van items door middel van p-waarden en item-restcorrelaties, en de evaluatie van testscores door middel van bijvoorbeeld Cronbachs alfa. Sindsdien is er veel gebeurd in de testtheorie en de testconstructie. De item-respons-theorie (IRT; Van der Linden & Hambleton, 1997) heeft de klassieke testtheorie voorbijgestreefd, althans binnen de psychometrie. Tests kunnen nu – in principe – als adaptieve test worden aangeboden, hetgeen inhoudt dat verschillende individuen niet langer dezelfde test krijgen aangeboden, maar op hun individuele niveau toegespitste testversies. Dit leidt tot een optimale meetnauwkeurigheid per individu. Deze ontwikkelingen zijn mede sterk gestimuleerd door de snelle opkomst van de computer, die in de jaren zestig nog maar een zeer bescheiden rol speelde.

Ook tegenwoordig zijn vele psychologische tests echter nog steeds pen-en-papier standaardtests en worden vele tests nog steeds geconstrueerd volgens de principes van de KTT. Daarom lijkt dit een uitgelezen gelegenheid om na te gaan waarom nog niet iedere testconstructeur de IRT heeft omhelst. Door middel van een vergelijking van KTT en IRT wordt vastgesteld wat hiervan de oorzaak zou kunnen zijn. Ook zal ik uitleggen wat de IRT toevoegt aan de KTT en voor welke praktische toepassingen van tests de IRT vooral geschikt is. Daarbij maak ik een onderscheid tussen psychologisch testen en onderwijskundig toetsen. Uiteindelijk is de conclusie dat de IRT kan wat de KTT ook kan, maar vaak iets beter en soms veel beter, en dat in sommige gevallen de IRT zelfs iets heel nieuws heeft te bieden. De IRT is dus superieur, maar ook

zal blijken dat het gebruik van de KTT in aanvulling op de IRT nog niet zo onverstandig is.

19.1 Het klassieke complex en het moderne complex

Van der Linden (1983) maakt binnen de testtheorie onderscheid tussen het klassieke complex en het moderne complex. Met het klassieke complex bedoelt hij het geheel van standaardtest en KTT. Tegenover de standaardtest staat de itembank, een reservoir van items die alle dezelfde psychologische eigenschap meten en daarbij tot op zekere hoogte gelijkwaardig zijn volgens een definitie die is gegeven door het gebruikte IRT-model. Het moderne complex bestaat uit het geheel van itembank en IRT. We behandelen eerst het klassieke complex en daarna het moderne complex.

19.1.1 Het klassieke complex

De KTT is de statistische theorie van de betrouwbaarheid en de validiteit van testscores. De betrouwbaarheid geeft een indruk van de mate waarin de test, indien onder identieke omstandigheden opnieuw voorgelegd aan dezelfde populatie van individuen, dezelfde testscores zou opleveren als de eerste keer. Technisch gezien is de betrouwbaarheid de correlatie tussen de scores op twee volstrekt gelijkwaardige tests, paralleltests genaamd, waarbij voor hetzelfde individu de scores alleen toevallig mogen verschillen. Hoe kleiner de toevallige verschillen, hoe hoger de betrouwbaarheid. Bij kleine verschillen kan op basis van een test goed worden voorspeld wat iemands score op een gelijkwaardige test zou zijn. Een hoge betrouwbaarheid is dus gewenst.

Betrouwbaarheid

De betrouwbaarheid is gebaseerd op een meetfoutentheorie, die uitgaat van de premisse dat een geobserveerde testscore X of, algemener, een geobserveerde meetwaarde, kan worden opgedeeld in een betrouwbaar deel T, ook wel bekend als "true score" of betrouwbare score, en een toevallige meetfout E, ook wel "error" genoemd. Ik spreek opzettelijk van "geobserveerde meetwaarde", om aan te geven dat de klassieke testtheorie, ofschoon al honderd jaar gemeengoed onder psychologen, niet typisch psychologisch is. Elke meetwaarde, of het nu een testscore, het afgelezen gewicht op een weegschaal, of de tijd op de 500 meter sprint voor schaatsers is, kan worden opgesplitst in een betrouwbaar en een toevallig deel.

Op zich is niet de betrouwbaarheid, maar de standaardmeetfout van belang voor het vaststellen van de nauwkeurigheid van de meting en de vergelijking van de meetwaarden van verschillende personen. Daarbij wordt in de context van de KTT verondersteld dat deze standaardmeetfout toepasbaar is op alle meetwaarden, dus over het gehele domein van de schaal. De praktische consequentie is dat alle respondenten door de test even betrouwbaar zouden worden gemeten. Dat dit onrealistisch is, blijkt uit het voorbeeld waarbij een lagere betrouwbare score op een studietoets bestaande uit meerkeuzevragen samengaat met een grotere meetfout tén gevolge van veelvuldiger gokken. Mensen met een lagere betrouwbare score worden dus onbetrouwbaarder gemeten. Hierop heeft de KTT geen antwoord.

Validiteit

De validiteit geeft aan in hoeverre de testscore datgene doet waarvoor hij bedoeld is. Twee vormen van validiteit zijn belangrijk. Ten eerste is de test bedoeld om een psychologische eigenschap te meten. Doordat de afstand tussen psychologisch construct en operationalisering in de vorm van een verzameling van items groot is, dient achteraf door middel van analyse van empirische testgegevens te worden vastgesteld in hoeverre de test het begrip-zoals-bedoeld meet. Zo zou het kunnen gebeuren dat een test voor rekenvaardigheid door een ingewikkelde formulering van de opgaven grotendeels taalvaardigheid meet. De vaststelling van dergelijke zaken valt onder het onderzoek naar de constructvaliditeit.

Ten tweede heeft een test een gebruiksdoel. Dit kan zijn een voorspelling van gedrag dat in de toekomst ligt, zoals de prestatie op de middelbare school, in een beroep of functie of in psychotherapie. Evenzo kan het doel de meting van intelligentie zijn of de diagnose van een neurotische stoornis. Ook hier kan men beargumenteren dat meting en diagnose geen op zichzelf staande doelen zijn, maar dat zij weer een voorspelling dienen. Bijvoorbeeld, het intelligentieniveau kan later worden gebruikt bij een beroepskeuzeadvies en de diagnose van neuroticisme kan worden gebruikt om het gesprek met een cliënt meer richting te geven, waarbij het uiteindelijke doel is hem of haar beter te laten functioneren. Het onderzoek naar de voorspellende kracht van de test resulteert in een uitspraak over de predictieve validiteit.

In de literatuur worden overigens tientallen soorten van validiteit onderscheiden. Zij kunnen alle worden herkend als specifieke vormen van de hier genoemde construct- en predictieve validiteit. Zo wordt bij

convergente validering onderzocht in hoeverre een test hetzelfde meet als andere tests en bij divergente validering in hoeverre een test juist iets anders meet. Zonder veel fantasie kunnen hier deelaspecten van een onderzoek naar constructvaliditeit in worden herkend.

Hoewel de klassieke testtheorie bijvoorbeeld via theorie over de invloed van meetfouten op correlaties en de formule voor attenuatiecorrectie wel degelijk specifieke bijdragen levert aan de validiteitstheorie, is het onderzoek naar validiteit in de eerste plaats onderzoek naar relaties tussen testscores enerzijds en andere variabelen (waarmee andere eigenschappen worden gemeten, maar ook succes in beroep of opleiding) anderzijds. Dit onderzoek gebeurt met allerlei statistische methoden, die niet typisch voor de klassieke testtheorie zijn. Gedacht kan worden aan factoranalyse of lineair structurele vergelijkingsmodellen voor het onderzoek naar de structuur in een test of een testbatterij of de plaats van een test in een verzameling van verwante tests. Ook kan men denken aan regressieanalyse voor de voorspellende kracht van een test of een testbatterij.

De standaardtest
Kenmerkend voor de standaardtest is dat elk individu precies dezelfde test krijgt aangeboden. De condities waaronder dit gebeurt, zijn voor iedereen zoveel mogelijk gelijk. Iedereen krijgt dus dezelfde instructie en proefvragen en wordt op hetzelfde tijdstip van de dag getest. Ook krijgt iedereen de test via hetzelfde medium voorgelegd, dus schriftelijk, mondeling of via het beeldscherm van een computer. Deze standaardisering maakt de vergelijkbaarheid van de prestaties van de respondenten mogelijk.

19.1.2 Het moderne complex

De IRT gaat uit van dezelfde basisgegevens – itemscores en totaalscores – als de KTT en streeft dezelfde doelen na: een betrouwbare en valide meting. Daartussenin zitten dus de verschillen, maar omdat uitgangspositie en einddoel dezelfde zijn, is te verwachten dat beide theorieën ook een aantal overeenkomsten hebben.

Betrouwbaarheid
IRT-modellen drukken de kans op een goed antwoord op een item of de kans op en bepaalde score op een rating scale uit als een functie van een of meer persoonskenmerken en een of meer itemkenmerken. Nemen we aan dat de itemkenmerken, zoals de moeilijkheid en het onderscheidend

vermogen tussen lage en hoge waarden van de gemeten eigenschap, per item vastliggen, dan is de kans op een bepaalde itemscore een functie van de persoonskenmerken. Deze persoonskenmerken zijn de meetwaarden op de schaal van de psychologische eigenschap en geven iemands positie op deze schaal aan. Doorgaans wordt aangenomen dat een hogere meetwaarde inhoudt dat de kans op bijvoorbeeld een goed antwoord ook hoger is.

De aannemelijkheidsfunctie kan worden gedefinieerd door van alle respondenten uit de steekproef en voor alle items uit de test de kansen op goede antwoorden te combineren in de gezamelijke verdeling van de gegevens onder het IRT-model. De grootste aannemelijkheidschatting geeft vervolgens die waarden van de persoons- en itemkenmerken – dit zijn de parameters van het IRT-model – waarvoor de aannemelijkheidsfunctie de maximale waarde heeft. Bij deze schattingen van de modelparameters zijn de gegevens het meest waarschijnlijk. De omgekeerde redenering is dan dat dit het meest plausibele model is dat de gegevens heeft gegenereerd en de schattingen van de parameters nemen we vervolgens voor waar aan.

Een handig bijproduct van de grootste aannemelijkheidsschatting is de informatiefunctie. Deze functie neemt in de IRT de plaats in van de betrouwbaarheidscoëfficiënt in de KTT. De informatiefunctie geeft de nauwkeurigheid van de schatting van een persoonsscore en varieert over de schaal waarop de persoonsscore wordt gemeten. Daarmee is gezegd dat de nauwkeurigheid van de meting verschilt voor personen met verschillende scores. Intuïtief kan men zich dit voorstellen door te bedenken dat een ruimtelijk inzichttest, die bijvoorbeeld bestaat uit overwegend gemakkelijke items, voor iemand met een goed ruimtelijk inzicht maar weinig informatief is. De test geeft voor deze persoon hooguit een ondergrens aan en over het preciese ruimtelijk inzichtniveau is dus maar weinig bekend. Dit komt tot uiting in een geringe nauwkeurigheid van de meting. Het is alsof we met een gewone thermometer voor huishoudelijk gebruik temperaturen meten boven de honderd graden Celcius. Daarvoor is die thermometer niet geschikt en de meetwaarden die we aflezen zeggen bijzonder weinig. De thermometer is echter wel geschikt voor temperaturen tussen min twintig en plus vijftig graden.

Vertaald naar de ruimtelijk inzichttest zouden we deze test dus moeten gebruiken om personen te meten met een matig ruimtelijk inzicht. Voor deze personen levert de test de meest betrouwbare metingen en dit is wat

de informatiefunctie laat zien. Naarmate het niveau van de test en het niveau van de geteste persoon beter overeenkomen, is de functiewaarde hoger, wat betekent dat de meting nauwkeuriger is. Een directe consequentie is dat we personen het beste items kunnen aanbieden die wat betreft hun moeilijkheid bij het niveau van deze personen passen. Dit zou echter betekenen dat in principe iedereen een andere test krijgt voorgelegd, een situatie die in de KTT volstrekt ondenkbaar is, vanwege de verwarring die dan ontstaat tussen het niveau van de persoon en het niveau van de test: Als Jan op een gemakkelijker test een hogere score heeft dan Piet op een moeilijker test, hoe moeten we dan ooit nog bepalen of dit nu komt door Jans hogere niveau of door de gemakkelijker test?

Validiteit

Elk IRT-model geeft een definitie van de equivalentie van de items uit een test. Deze definitie beperkt de verzameling van items die bijeen in een test kunnen komen aanzienlijk en dit houdt in dat niet alle denkbare items voor bijvoorbeeld de meting van ruimtelijk inzicht of introversie in dezelfde test worden toegelaten. IRT-modellen bepalen dus met welke items een eigenschap wordt gemeten en als zodanig bepalen zij mede de constructvaliditeit van de test. Zo definieert het een-parameter logistisch model of Rasch-model equivalentie van items door te veronderstellen dat de items uit een test inwisselbaar zijn op hun moeilijkheid na. In het Rasch-model is bijvoorbeeld de sterkte van de relatie met de gemeten eigenschap voor alle items gelijk en daarmee huldigt het Rasch-model een restrictieve opvatting over validiteit: items meten alleen hetzelfde, eventueel met verschillende moeilijkheidsgraad, indien zij even sterk samenhangen met de onderliggende eigenschap. Het twee-parameter logistisch model of Birnbaum-model laat variatie in moeilijkheid toe en laat ook de samenhang met de onderliggende eigenschap binnen zekere beperkingen vrij. Daarmee is een ruimere definitie van constructvaliditeit gegeven. Weer andere IRT-modellen laten bijvoorbeeld het aantal gemeten aspecten van een eigenschap vrij en zijn daarmee meerdimensioneel. Zo is uiteindelijk elk IRT-model een variant op het thema van de equivalentie van items. Daarbij hebben sommige modellen onderling een hiërarchische relatie, zoals de zojuist genoemde Rasch- en Birnbaum-modellen, terwijl andere modellen variatie vertonen op verschillende kenmerken, waardoor ze niet hiërarchisch zijn te ordenen.

Overigens is met de inherente definitie van validiteit het validiteitsonderzoek van de test niet afgerond. Ten eerste is het maar de vraag of men bijvoorbeeld het Rasch-model als definitie van de

constructvaliditeit wil accepteren of dat men niet liever een ruimer gedefinieerd IRT-model zou willen kiezen. Ook kan een eigenschap diverse facetten omvatten, waardoor een meerdimensioneel IRT-model adequater is dan een eendimensioneel model. Ten tweede is het maar de vraag of een test die voldoet aan de eisen van een IRT-model een goede voorspeller is van geschiktheid voor beroep, opleiding of therapie. Evenals bij tests die zijn geconstrueerd volgens de principes van de KTT, dient een test die voldoet aan de eisen van een IRT-model apart te worden onderzocht op geschiktheid voor het beoogde gebruiksdoel. Hierin verschillen KTT en IRT dus niet.

De itembank

Indien een IRT-model de testgegevens adequaat beschrijft, nemen we vervolgens aan dat de eigenschappen van het model gelden voor de meting in kwestie. Dit wordt wel "measurement by implication" genoemd. In veel IRT-modellen wordt de persoonsscore gemeten op een intervalschaal en bij een passend model nemen we dus aan dat de meting van bijvoorbeeld ruimtelijk inzicht intervalniveau heeft. Doordat het meetniveau bekend is, kunnen we de schalen van verschillende tests voor ruimtelijk inzicht, bijvoorbeeld een gemakkelijke, een middelmatig moeilijke en een moeilijke test, in principe "over elkaar heen leggen" en corrigeren voor verschillende nulpunten en verschillende meeteenheden. Hiervoor is nodig dat verschillende tests enkele items gemeen hebben, die als ankerpunten dienen.

Dit equivaleren van schalen opent de mogelijkheid om schalen voor een eigenschap te maken die gebaseerd zijn op honderden items, zonder dat alle proefpersonen alle items hebben gemaakt. Zo ontstaat een itembank, waaruit diverse groepstests kunnen worden samengesteld, maar ook tests die op maat zijn gemaakt voor de te testen respondent. Respondenten met verschillende niveaus krijgen in dit geval verschillende tests, maar hun testprestaties zijn toch vergelijkbaar doordat van alle items bekend is hoe ze ten opzichte van elkaar op de schaal liggen. Bij het toekennen van scores aan personen wordt hiermee rekening gehouden, zodat die scores vergelijkbaar zijn. Vanuit het oogpunt van acceptabiliteit en transparantie zijn er verder geen problemen te verwachten, want personen zijn zich er niet van bewust dat zij een unieke test maken, en bovendien zal de afstemming van de opgaven op het persoonlijke niveau bewerkstelligen dat niemand zich verveelt (veel gemakkelijke items) of gefrustreerd raakt (veel moeilijke items).

19.2 De praktijk van testconstructie en testgebruik

Voor testconstructeurs en testgebruikers is het meest opvallende kenmerk van de IRT waarschijnlijk de mogelijkheid om itembanken te maken en hieruit grote aantallen groepstests te selecteren of individuele tests op maat samen te stellen. Grote aantallen groepstests zijn nodig als men parallelle of anderszins equivalente tests wil maken, bijvoorbeeld als op verschillende tijdstippen grote aantallen personen moeten worden getest en geheimhouding van de items lastig is. Binnen de psychologie zijn voorbeelden van tests op maat of adaptieve tests de test voor woordenschat van Schoonman (1989; test niet gepubliceerd) en de Snijders-Oomentest voor niet-verbale intelligentie van Laros en Tellegen (1991; zie Evers, Van Vliet-Mulder & Ter Laak, 1992, voor een beoordeling). Ook het CITO te Arnhem maakt op kleine schaal adaptieve tests, maar verder zijn mij geen voorbeelden bekend van Nederlandstalige adaptieve psychologische tests.

Toepassingen die itembanken vereisen, worden juist meer aangetroffen in het onderwijskundig toetsen. De belangrijkste reden is dat in het onderwijs vaak enorme aantallen leerlingen worden getoetst, waarbij de toetsing doorgaans in groepen plaatsvindt en groepen vaak op verschillende tijdstippen worden getoetst. Daarentegen zijn er eindexamens van opleidingen, die wel gelijktijdig voor iedereen plaatsvinden. In dergelijke grootschalige toepassingen, waar bovendien soms grote belangen op het spel staan, raken items snel bekend, zodat zij regelmatig vervangen moeten worden. Items die kennis en vaardigheden meten, eigenschappen die bij uitstek in het onderwijs worden getoetst, zijn relatief gemakkelijk in groten getale te maken. Juist de combinatie van "overexposure" en het relatief grote gemak waarmee items kunnen worden bijgemaakt, maakt het aantrekkelijk om grote aantallen te produceren en toetsen vaak te vervangen. Dit zijn ideale condities voor de productie van itembanken en het vervaardigen van grote aantallen verschillende toetsen.

Hoe anders verloopt het psychologisch testen. Ofschoon ook in de psychologie de groepstest niet onbekend is – denk aan de vroegere psychologische keuring voor militaire dienst – en ook hier kennis en vaardigheden worden gemeten – bijvoorbeeld ten behoeve van selectie voor diverse functies – is het testen hier vaak een individuele aangelegenheid. Zo worden de intelligentie en de persoonlijkheid van kinderen en volwassenen vaak in individuele sessies vastgesteld. Doel is dan bijvoorbeeld een diagnose van iemands verstandelijke vermogens in

verband met tegenvallende prestaties op school, een diagnose van leesproblemen of meer specifiek dyslexie, een vaststelling van het persoonlijkheidsprofiel in verband met de geschiktheid voor een bepaalde functie of een diagnose van neuroticisme met het oog op verdere behandeling. In de sfeer van de intelligentiemeting en de meting van cognitieve vermogens hebben de items bovendien vaak het karakter van opdrachten en taken in de vorm van doolhoven, blokken, vergelijking van lengtes, gewichten en oppervlakken van objecten, mentale rotatie van geometrische figuren, en het vinden van figurele, verbale en geometrische analogieën. Duidelijk is dat dergelijke taken vaak moeilijk zijn te onthouden en na te vertellen. Het individuele testproces en vaak ook de abstractie of de complexiteit van de taken maakt dat "overexposure" hier niet een erg groot probleem is. Daar komt dan nog bij dat mensen niet kunnen "zakken", zodat zij de test maar één keer maken.

Wat dus vooral opvalt aan de IRT – de itembank en zijn spectaculaire gebruiksmogelijkheden – lijkt in het psychologisch meten voorlopig minder dominant. Wat is dan de betekenis van de IRT voor de psychologische test? We noemen vier voorbeelden.

19.2.1 Vraagpartijdigheid (differential item functioning)

In de Verenigde Staten begon reeds in de jaren zestig en zeventig het grootschalige onderzoek naar "differential item functioning", aanvankelijk "item bias" geheten en met een goede Nederlandse term vraagpartijdigheid genoemd (Kok, 1988). Hierbij is de vraag of de leden van een maatschappelijke minderheidsgroep – in de VS veelal zwarten of "hispanics" en in Nederland bijvoorbeeld allochtonen van Turkse of Marokkaanse afkomst – door een test of vragenlijst benadeeld worden. Zo zou het kunnen gebeuren dat een taalachterstand of minder goed onderwijs ervoor zorgen dat de leden van een minderheid in het nadeel zijn, dus meer fouten maken, op tests die taalvaardigheid of andere vaardigheden vereisen die vooral via het onderwijs worden geleerd. Een aardig voorbeeld is een rekenvaardigheidtest, die ook taalvaardigheid vereist. Nemen we iemand uit de minderheidsgroep en iemand uit de meerderheidsgroep, beiden met een even goede rekenvaardigheid, dan zou degene uit de minderheidsgroep toch een slechtere testprestatie kunnen leveren door een minder goede taalvaardigheid. Bij selectie van personen voor functies in het bedrijfsleven of bij de overheid, waarvoor mede wordt geselecteerd op rekenvaardigheidstestscores, zouden leden uit de minderheid dus systematisch in het nadeel zijn omdat de test ook nog "stiekem" taalvaardigheid meet. Het onderzoek naar vraagonzuiverheid

houdt zich bezig met het opsporen en verwijderen van de items die deze ongewenste dubbelrol vervullen. Als taalvaardigheid belangrijk is voor de betreffende functie, dan dient men hier door middel van een aparte test expliciet op te selecteren. Het is verder een beleidskwestie wat men wil doen met een eventuele geringere taalvaardigheid van sommige groepen: een vast aantal personen met voorrang toelaten, eerst bijscholingsonderwijs geven, enzovoorts.

De IRT heeft een eenvoudige definitie van vraagonzuiverheid. Een item is *zuiver* als voor een gegeven meetwaarde (bijvoorbeeld, een bepaald rekenvaardigheidsniveau) de kans om het item goed te beantwoorden gelijk is voor personen uit de meerderheids- en de minderheidsgroep en *onzuiver* als deze kans afhankelijk is van groepslidmaatschap. In het laatste geval verschillen de groepen systematisch op een eigenschap, bijvoorbeeld taalvaardigheid, waar het item gevoelig voor is. Vraagonzuiverheid is dus een vorm van meerdimensionaliteit. In de IRT zijn diverse methoden bedacht waarmee op een eenvoudige manier kan worden onderzocht of items zuiver zijn (Holland & Wainer, 1993). Nu Nederland in hoog tempo "multicultureler" wordt, valt te verwachten dat vooral de relatieve nieuwkomers door verschillen in taalvaardigheid en onderwijs te maken krijgen met tests die hierop nog niet zijn ingesteld. Dit speelt niet alleen bij selectie voor functies, maar ook in alle vormen van selectie in het onderwijs. Het onderzoek naar vraagonzuiverheid is een belangrijke psychologische toepassing van de IRT.

19.2.2 Afwijkende scorepatronen (person fit)

Het beoordelen van testprestaties gebeurt meestal met behulp van totaalscores of testscores, die een indicatie geven van het niveau op de gemeten eigenschap. Wat wel eens over het hoofd wordt gezien is dat ook het patroon van de scores op de afzonderlijke items informatie geeft over de testprestatie. Zo is bekend dat sommige respondenten langzaam op gang komen en de eerste items slechter maken dan op basis van hun prestatie op de overige items zou kunnen worden verwacht. Dit zijn de "faalangstigen". Verder ziet men wel eens dat iemand over de hele test genomen een veel mindere prestatie levert dan op basis van enkele goede antwoorden op moeilijke items kan worden verwacht. Zo iemand was wellicht ongeconcentreerd, ongeïnteresseerd, of zag het belang van de testsituatie onvoldoende in. Dit is een "slaper". Ook gebeurt het dat iemand van de moeilijker items opvallend veel goed beantwoordt in vergelijking met zijn/haar prestatie op de gemakkelijker items. Als dit het gevolg is van afkijken of van een andere vorm van bedrog, hebben we te

maken met een "bedrieger". Zo zijn er meer voorbeelden bekend. Waar het om gaat, is dat in al deze gevallen de totaalscore niet representatief is voor wat iemand werkelijk kan, terwijl de itemscores hier aanvullende informatie over bevatten.

In vrijwel alle IRT-modellen geldt dat een toenemende meetwaarde van personen inhoudt dat de kans op het goede antwoord op een item toeneemt. Volgens dit principe verwachten we dus dat als iemand fouten maakt, het de relatief moeilijke items zal betreffen. Toevalsmechanismen maken dat iemand soms wel eens tegen de verwachting in een moeilijk item goed maakt of faalt op een gemakkelijk item, maar als we hier even vanaf zien, dan verwachten we dus dat bij bijvoorbeeld tien van de 15 items goed het de tien gemakkelijkste items betreft terwijl de vijf moeilijkste items fout zijn gemaakt. Bij deze verwachting is het opvallend dat de eerste (relatief gemakkelijke) items fout zijn (faalangstigen); afwisselend moeilijke en gemakkelijke items fout worden gemaakt (slapers), en vooral moeilijke items goed worden gemaakt terwijl gemakkelijker items veelvuldig fout worden gemaakt (bedriegers).

In de IRT zijn diverse methoden bedacht die bij een gegeven IRT-model volgens bovengenoemd principe per itemscorepatroon, dus voor elke respondent apart, aangeven in hoeverre dit patroon afwijkt van wat men zou verwachten (Meijer & Sijtsma, in press). Op zich zegt gevonden afwijking alleen nog maar dat het patroon van itemscores niet bij het gebruikte IRT-model past. Daar staat tegenover dat het wel *opmerkelijk* is dat iemand zich niet houdt aan het verwachte patroon van moeilijker items fout en gemakkelijker items goed. Verder dient de onderzoeker bij gevonden afwijkendheid aanvullende evidentie te verzamelen die kan verklaren waarom een patroon van itemscores anders is dan verwacht. Deze evidentie kan worden verzameld uit interviews met de geteste personen of uit achtergrondvariabelen. Zo kan bijvoorbeeld blijken dat vooral de leden van een minderheidsgroepering bepaalde relatief gemakkelijke items fout hebben beantwoord, terwijl zij moeilijker items wel goed beantwoorden. Als vervolgens blijkt dat de gemakkelijker items een veel groter beroep doen op taalvaardigheid dan de moeilijker items, dan is hier tevens een link gelegd naar het onderzoek van vraagonzuiverheid.

19.2.3 Verklaring van onderliggende processen (cognitive modeling)

Test- en vragenlijstconstructie is een nogal technologische aangelegenheid. Testconstructeurs zijn de instrumentenbouwers onder de

psychologen. De wetenschappelijke bijdrage van tests ligt dan vervolgens in hun toepassing in psychologisch onderzoek, waarin de testscores fungeren als onafhankelijke of afhankelijke variabele. In deze zin is validiteitsonderzoek meer wetenschappelijk en minder technologisch dan testconstructie.

De IRT kent een klasse van modellen, de componentiële IRT-modellen, die niet alleen tot doel hebben om kenmerken van items vast te stellen en meetwaarden voor respondenten te schatten, maar die tevens een verklaring geven voor de totstandkoming van de itemscores. Dit gebeurt bijvoorbeeld door per item uit de test te hypothetiseren welke deelvaardigheden nodig zijn om het item op te lossen. Ook moet vooraf worden vastgesteld in welke mate elk van deze deelvaardigheden nodig zijn (Kelderman & Rijkes, 1994). Rijkes (1996, h.2) bespreekt vier verschillende hypothesen over de oplossing van de taken uit Ravens Progressive Matrices Test en toetst deze hypothesen door middel van vier componentiële IRT-modellen. Een andere invalshoek is om per item te hypothetiseren welke aspecten van het item verantwoordelijk zijn voor het moeilijkheidsniveau van het item (Fischer, 1995). Zo kan het ene item een ingewikkelder samenstelling hebben dan het andere item en daardoor naar verwachting moeilijker zijn. Deze aanpak werd gevolgd door Van Maanen, Been en Sijtsma (1989) bij hun onderzoek naar balanstaken (taken over een weegschaal met twee armen met daaraan gewichten).

Beide aanpakken leiden vooral tot verschillende uitwerkingen in de componentiële IRT-modellen, maar de overeenkomsten zijn verder groter dan de verschillen. Ten eerste hypothetiseren beide een bepaalde cognitieve strategie die beschrijft langs welke weg items worden opgelost. Dit geeft inzicht in het psychologische mechanisme achter de totstandkoming van itemscores en het betekent dat een poging wordt gedaan om de "black box" tussen stimulus en respons te verklaren. Ten tweede heb ik hier niet voor niets steeds het woord hypothetiseren gebruikt: De verklaring dient per item te worden opgesteld en wordt vervolgens vertaald in de parameters van het componentiële IRT-model. Daarna wordt onderzocht of het model de gegevens (itemscores) kan verklaren. Het IRT-model is hier dus de nulhypothese, waarmee wordt getracht iets te leren over de psychologie van het probleemoplossen.

19.2.4 Passen van meetmodel bij data: measurement by implication

Waren de voorgaande drie toepassingen van IRT sterk psychologisch gekleurd, de laatste toepassing is net als de toepassing van de KTT meer

technologisch. De meest algemene toepassing van de IRT is die ten behoeve van test- en vragenlijstconstructie, waarbij het doel is een meetinstrument te maken dat betrouwbaar en valide is. Reeds eerder noemde ik de mogelijkheid om onder een aantal IRT-modellen de informatiefunctie te schatten, die een schatting van de betrouwbaarheid toelaat als functie van de schaal. Ook noemde ik dat elk IRT-model een definitie geeft van de kenmerken waaraan de items dienen te voldoen om samen in een test te kunnen worden toegelaten. Daarmee is een soort van validiteitsdefinitie gegeven.

Anders dan de KTT laat de IRT toe om te onderzoeken of de structuur van het model past bij de verzamelde itemscores. Is dit inderdaad het geval, dan gelden vervolgens bij implicatie de eigenschappen van het IRT-model voor de gevonden empirische schaal. Impliceert het model bijvoorbeeld een ordinale schaal, dan gaan we er bij een passend model vanuit dat we de respondenten mogen ordenen op basis van hun testprestatie. Zo zijn er ook modellen die een interval- of een ratioschaal impliceren. Een dergelijke schaal is vooral in technische zin handig bij het equivaleren en bij toetsselectie en adaptief testen op basis van een itembank. Met dit *meten bij implicatie* staat de IRT in duidelijk contrast tot de KTT, die geen modelpassingsonderzoek kent en waarbij de schaaleigenschappen niet uit een passend model kunnen volgen. Dat men in de praktijk van de KTT doet alsof testscores intervaleigenschappen hebben, is dan ook vooral een praktische keuze, met het oog op de toepassing van de statistiek die zo'n keuze vereist. De KTT leidt dus hooguit tot *meten bij fiat*.

19.3 Tot slot: voor- en nadelen van klassieke testtheorie en item-respons-theorie

Een belangrijk pluspunt van de KTT is dat deze goed aansluit bij de intuïtie. Het idee van meetfouten is goed uit te leggen en maakt aannemelijk dat gestreefd dient te worden naar betrouwbare testscores. Voor het schatten van de betrouwbaarheid zijn goede en praktisch handige methoden bedacht. Een ander principe dat in de KTT goed is uitgewerkt, is dat van de testverlenging: Naarmate de test meer items van een bepaalde kwaliteit bevat, zal de betrouwbaarheid toenemen. Dus, meer informatie betekent naar verhouding meer signaal en minder ruis. In het algemeen geldt dat de KTT niet al te moeilijk is en daardoor goed uit te leggen. Bovendien is er tegenwoordig uiterst gebruikersvriendelijke

software (bijv. SPSS) beschikbaar waarmee een analyse volgens de KTT in een handomdraai (eigenlijk: een muisklik) kan worden uitgevoerd.
Ik hoop dat de voorgaande pagina's duidelijk hebben gemaakt dat de IRT meer te bieden heeft dan de KTT. Betrouwbaarheid als functie van de schaal, meten bij implicatie, de onderzoeken naar vraagonzuiverheid en afwijkende scorepatronen (voor beide toepassingen zijn overigens buiten de context van de IRT ook methoden ontwikkeld), en cognitief modelleren zijn verworvenheden van de IRT, waar de KTT niet aan kan tippen. In een iets ruimere context kunnen daaraan de selectie van tests uit itembanken en het adaptieve testen worden toegevoegd. Bij het zien van deze lijst is het eigenlijk vreemd dat de IRT in de psychologie nog steeds veel minder wordt gebruikt dan de KTT. Ik geef enkele verklaringen.

Ten eerste is de IRT moeilijker te begrijpen dan de KTT. Kan men voor een elementair begrip van de KTT nog toe met kennis van correlatierekening, voor de studie van de IRT is statistische kennis nodig die vaak niet tot de basisstof van de psychologieopleiding behoort. Wel leert de psychologiestudent tegenwoordig de beginselen van de IRT (zie, bijvoorbeeld, Hoofdstuk 6 uit Drenth & Sijtsma, 1990; zie tevens Van den Brink & Mellenbergh, 1998), maar de ervaring leert dat de IRT een zware kluif is. Het enige dat dan echt helpt, naast allerlei didactische trucs, is er meer aandacht aan besteden. Deze grotere aandacht wordt gerechtvaardigd door de superioriteit van de methode.

Ten tweede zijn IRT-methoden nog niet in pakketten als SPSS opgenomen, maar is er sprake van een versnippering van de diverse methoden over nog meer stand-alone programma's. Diverse van deze programma's zijn zeer de moeite waard, maar het zijn soms wel programma's door psychometrici voor psychometrici. Dit blijkt nog niet eens alleen uit de aansturing (men moet soms wel heel veel beslissingen nemen en specificaties opgeven), maar bijvoorbeeld ook uit de technische keuzes die de gebruiker dient te maken (bijvoorbeeld, de schattingsmethode) en waar hij/zij geen verstand van heeft. Zoiets intimideert en schrikt af. Naast de goede "high brow" versies die er zijn, zouden er dus meer expliciete gebruikersversies moeten komen.

Ten derde moet de tijd gewoon zijn werk doen, de IRT is tenslotte nog jong en wat goed is komt toch wel boven drijven. Maar zonder de nodige aandacht in de basisopleiding en zonder eenvoudige computerprogramma's zal de IRT niet volledig doordringen. Vooralsnog wordt de KTT dus het meest gebruikt en echt erg is dat niet. De KTT legt

de nadruk op de betrouwbaarheid van de meting en dat is een goede zaak. Ook in combinatie met de IRT kan de KTT een nuttige rol vervullen, bijvoorbeeld in het vooronderzoek, waarin de onderzoeker de experimentele test eerst op een kleine steekproef uitprobeert. Voor kleine steekproeven is de KTT vanwege de grotere statistische eenvoud geschikter dan de IRT en leidt zij snel tot een eerste indicatie van de kwaliteit van individuele items en van de betrouwbaarheid van de testscore. Zo bezien hoeft de opkomst van de ene methode niet automatisch de ondergang van de andere methode te betekenen.

Literatuur

Drenth, P.J.D. (1965a). *De psychologische test*. Arnhem: Van Loghum Slaterus.

Drenth, P.J.D. (1965b). *Test voor niet-verbale abstractie*. Amsterdam: Swets & Zeitlinger.

Drenth, P.J.D., & Hoolwerf, G. (1970). *Numerieke aanleg test, 1970 (NAT '70)*. Amsterdam: Swets & Zeitlinger.

Drenth, P.J.D., & Sijtsma, K. (1990). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu Van Loghum.

Drenth, P.J.D., & Van Wieringen, P.C.W. (1969). *Verbale aanleg testserie 1969 (VAT '69)*. Amsterdam: Swets & Zeitlinger.

Evers, A., Van Vliet-Mulder, J.C., & Ter Laak, J. (1992). *Documentatie van tests en testresearch in Nederland*. Assen: Van Gorcum.

Fischer, G.H. (1995). The linear logistic test model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131-155). New York: Springer-Verlag.

Holland, P.W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kelderman, H., & Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Kok, F.G. (1988). *Vraagpartijdigheid*. Amsterdam: Universiteit van Amsterdam (dissertatie).

Laros, J.A., & Tellegen, P.J. (1991). *Construction and validation of the SON-R 5½-17, the Snijders-Oomen non-verbal intelligence test*. Groningen: Wolters-Noordhoff.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Meijer, R.R., & Sijtsma, K. (in press). A review of methods for evaluating the fit of item score patterns on a test. *Applied Psychological Measurement*.

Rijkes, C.P.M. (1996). *Testing hypotheses on cognitive processes using IRT models*. Enschede: Universiteit Twente (dissertatie).

Schoonman, W. (1989). *An applied study on computerized adaptive testing*. Amsterdam: Swets & Zeitlinger.

Van den Brink, W.F., & Mellenbergh, G.J. (Eds.), (1998). *Testleer en testconstructie*. Meppel: Boom.

Van der Linden, W.J. (1983). *Van standaardtest naar itembank*. Enschede: Universiteit Twente (oratie).

Van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Van Maanen, L., Been, P.H., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E.E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267-287). New York: Springer-Verlag.