1

Inleiding

Aan het construeren van studietoetsen, psychologische tests en andere sociaalwetenschappelijke meetinstrumenten kan een kwalitatieve en een kwantitatieve component onderscheiden worden. Het belangrijkste aspect van de kwalitatieve component betreft het ontwikkelen van de vragen of opdrachten waaruit het meetinstrument bestaat. De kwantitatieve component betreft het analyseren van antwoorden van personen op vragen of opdrachten. De kwantitatieve component van het toetsconstructieproces vormt het aandachtsgebied van de psychometrie. In dit boek wordt beschreven hoe door toepassing van psychometrische theorieën en statistische technieken de kwaliteit van meetinstrumenten beschreven, onderzocht en verbeterd kan worden.

Dit hoofdstuk bestaat uit twee verschillende onderdelen. Het doel van het eerste onderdeel is de bijdrage van de psychometrie voor de testpraktijk aan te geven. Daartoe wordt eerst in paragraaf 1.1 aan de hand van testindelingen een overzicht gegeven van de meetinstrumenten die mede met behulp van de psychometrie ontwikkeld zijn. Vervolgens wordt in paragraaf 1.2 beschreven wat de psychometrie bijdraagt aan de verschillende fasen van het toetsconstructieproces. In het tweede onderdeel van dit hoofdstuk worden de belangrijkste psychometrische aspecten van meetinstrumenten besproken. In paragraaf 1.3 wordt het valideren van meetinstrumenten besproken. In paragraaf 1.4 worden verschillende psychometrische theorieën besproken die bij het construeren van meetinstrumenten worden toegepast.

1.1 Testindelingen

De 'Documentatie van tests en testresearch in Nederland' (Evers, Van Vliet-Mulder, & Ter Laak, 1992) bevat een overzicht van bijna vierhonderd Nederlandstalige psychologische en andere meetinstrumenten en van het onderzoek dat ermee is verricht. Met meetpretentie als indelingsprincipe worden in dat overzicht drie klassen of soorten meetinstrumenten onderscheiden:

De eerste klasse bevat meetinstrumenten die als meetpretentie hebben stabiele persoonlijkheidskenmerken van personen te meten. Het gaat hierbij om kenmerken die zoveel mogelijk onafhankelijk zijn van bijvoorbeeld een arbeids- of opleidingssituatie. Voorbeelden van meetinstrumenten uit deze klasse zijn intelligentietests en persoonlijkheidsvragenlijsten. Ook de verborgen-figurentest die in hoofdstuk 7 besproken wordt, is een meetinstrument uit deze klasse.

De tweede klasse betreft meetinstrumenten die als meetpretentie hebben kenmerken te meten van personen in interactie met een (klasse van) situatie(s). Tot deze klasse behoren meetinstrumenten zoals beroepeninteressevragenlijsten en studietoetsen. In de navolgende hoofdstukken worden met name studietoetsen besproken. Een algemeen bekende Nederlandse studietoets is de Eindtoets Basisonderwijs (Uiterwijk & Engelen, 1993).

Bij de derde klasse gaat het om meetinstrumenten waarmee personen (beoordelaars) een oordeel over bepaalde situaties geven, bijvoorbeeld het oordeel van chefs over taakinhouden en taakkenmerken in de arbeidssituatie. In hoofdstuk 12 worden verschillende beoordelingssituaties besproken waarbij beoordelaars de meetinstrumenten zijn.

In de 'Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen' (1988) wordt een onderscheid gemaakt tussen 'test' en 'studietoets'. De term test wordt gebruikt voor meetinstrumenten die geschiktheid of aanleg meten. In voorgaande indeling behoren deze tests tot de meetinstrumenten uit de eerste klasse. De term studietoets wordt gebruikt voor meetinstrumenten die vaardigheden meten, bijvoorbeeld reken- of leesvaardigheid, die het resultaat zijn van onderwijs, training of instructie.

De indeling op basis van meetpretentie is een van de vele mogelijke indelingsprincipes voor het indelen van meetinstrumenten. Drenth en Sijtsma (1990, p. 36-63) bespreken drie testindelingen. De eerste indeling is gebaseerd op het gedrag van de onderzochte persoon. Hierbij is het belangrijkste onderscheid dat tussen tests voor prestatieniveau, bijvoorbeeld intelligentietests, en tests voor gedragswijze, bijvoorbeeld zelfbeoordelingen. Een tweede testindeling is die op basis van verschillende wijzen van instructie en afname. Twee belangrijke onderscheidingen hierbij zijn die tussen de individuele test en de groepstest en die tussen de snelheidstest ('speed test') en de niveautest ('power test'). De derde testindeling is een indeling die gebaseerd is op de aard van de testvragen, bijvoorbeeld tussen toetsen met gesloten vragen ('multiple choice') en toetsen met open vragen.

In dit boek worden meetinstrumenten onderscheiden op basis van het doel dat met het meetinstrument beoogd wordt. Aangezien dit indelingsprincipe geen inhoudelijke

onderscheidingen tussen meetinstrumenten maakt, heeft daarmee ook een terminologisch onderscheid zoals dat tussen test en toets geen betekenis meer. De termen test en toets worden in dit boek dan ook als synoniem gebruikt. Aangezien het toepassingsgebied van dit boek met name studietoetsen betreft, zal in de meeste gevallen de term toets gebezigd worden.

Met het doel van de toets als indelingsprincipe kunnen drie categorieën toetsen onderscheiden worden. De eerste categorie betreft toetsen waarvan het doel is het leren en onderwijzen in de klas te ondersteunen en te sturen. Deze toetsen geven de docent informatie over de vorderingen van elke leerling waarop de docent zijn onderwijs aan zijn leerlingen kan baseren. In hoofdstuk 10 worden voorbeelden van toetsen uit deze categorie besproken.

De tweede categorie betreft toetsen waarvan het doel is uitspraken te doen over hoe onderwijsprogramma's of onderwijssystemen functioneren. Deze toetsen zijn in de eerste plaats bedoeld om informatie aan bijvoorbeeld leerplanontwikkelaars of beleidmakers te geven. Tot deze categorie behoren de toetsen die onderdeel uitmaken van het peilingsonderzoek dat in hoofdstuk 7 besproken wordt.

De derde categorie betreft toetsen die selectie, plaatsing of certificering van leerlingen tot doel hebben. We spreken van selectie als de toets tot doel heeft leerlingen toe te laten of af te wijzen voor een opleiding. Deze toetsen worden met name gebruikt door opleidingen met een beperkt aantal opleidingsplaatsen. De selectie zal strenger zijn naarmate het aantal opleidingsplaatsen beperkter en de opleiding duurder is, bijvoorbeeld de toelating voor de opleiding tot piloot. Wanneer het doel van de toets is leerlingen naar een bepaald onderwijsprogramma te verwijzen, spreken we van plaatsing of classificatie. Voorbeelden zijn toetsen die gebruikt worden om een leerling naar een school voor speciaal onderwijs te verwijzen, of toetsen die gebruikt worden om te beslissen of een leerling na afsluiting van de brugperiode naar mavo, havo of vwo moet gaan. We spreken van certificering als het doel van de toets is te beslissen of leerlingen de leerinhouden van het onderwijsprogramma waaraan zij hebben deelgenomen wel of niet beheersen. De bekendste voorbeelden zijn de zeer vele examens en tentamens die in alle vormen van onderwijs afgenomen worden. Voor bepaalde opleidingen geldt dat de leerlingen na het behalen van een aantal certificaten in het bezit kunnen komen van een diploma. Beheersingsbeslissingen veronderstellen een zogenaamde drempel of cesuur die aangeeft welke toetsscore als laagste voldoende prestatie aangemerkt kan worden. In hoofdstuk 13 worden methoden voor cesuurbepaling besproken.

1.2 Toetsconstructie

Het constructieproces van een toets kan in een aantal fasen uiteen worden gelegd. Het proces begint met het operationaliseren van de vaardigheid die gemeten wordt en het vaststellen van het gebruiksdoel van de toets en eindigt met het schrijven van de handleiding en de verantwoording van de toets. Tussen de eerste en laatste fase moeten talrijke beslissingen genomen en activiteiten ondernomen worden. In onderstaande beschrijving van het toetsconstructieproces worden acht fasen onderscheiden en toegelicht. Bij deze beschrijving zijn de volgende twee opmerkingen van belang. De eerste opmerking is dat de beschrijving niet geïnterpreteerd moet worden als dat het toetsconstructieproces altijd uit acht fasen zou bestaan. De beschrijving is met name van toepassing op studietoetsen maar zelfs daar kan afhankelijk van de toets het proces uit meer of minder fasen bestaan. De tweede opmerking is dat in de beschrijving het toetsconstructieproces lineair verloopt, terwijl het proces in werkelijkheid eerder iteratief zal zijn. De output van de ene fase is weliswaar de input voor de volgende fase, maar dit betekent niet dat men op beslissingen die in een bepaalde fase genomen zijn niet kan of moet terugkomen.

Fase 1: Doelspecificatie

De eerste fase van het toetsconstructieproces bestaat uit het operationaliseren van de vaardigheid die de toets moet meten en het vaststellen van het gebruiksdoel van de toets. De plaatsingstoetsen Engels voor de brugklas operationaliseren het meten van de vaardigheid Engels als reproduktieve en produktieve aspecten van leesvaardigheid (Sluijter, Boertien, De Klijn, & Van Roosmalen, 1991). Als gebruiksdoel van de plaatsingstoetsen wordt het bepalen van de meest geschikte categorale onderwijsvorm voor leerlingen na afsluiting van de brugperiode genoemd.

Fase 2: Toetsspecificatie

Op basis van de operationalisatie van de te meten vaardigheid en het gebruiksdoel van de toets, worden in deze fase de kenmerken van de toets vastgesteld. Hieronder wordt een niet uitputtende opsomming van vragen gegeven waarmee de toetsconstructeur bij de constructie van een toets te maken kan krijgen (Millman & Greene, 1989, p. 339). De eerste drie vragen betreffen externe randvoorwaarden waarmee de toetsconstructeur rekening moet houden. De vragen daarna hebben betrekking op de kenmerken van de toets waarbij de eerste vraag naar de inhoud van de toets de belangrijkste vraag is.

Bij wie wordt de toets afgenomen?

- Voor het vaststellen van de toetsspecificaties is het noodzakelijk te weten bij welke personen de toets met welk doel wordt afgenomen. Het toetsconstructieproces zal anders verlopen wanneer het een toets betreft voor een heterogene groep personen voor een certificaat, dan wanneer het een toets betreft voor een homogene groep personen met het doel om de meest vaardige personen te selecteren.

Hoeveel toetstijd is er beschikbaar?

- Hoewel door praktische omstandigheden de beschikbare toetstijd vaak beperkt is, moeten leerlingen ruim de tijd krijgen voor het beantwoorden van de toets. Wanneer leerlingen te weinig toetstijd krijgen, dan wordt niet alleen het niveau van de uitvoering maar ook de snelheid van uitvoering beoordeeld. In het laatste geval wordt een andere vaardigheid gemeten dan wanneer alleen het niveau van de uitvoering gemeten wordt. Wanneer de toetstijd te beperkt is, kan dat ook betekenen dat te weinig vragen afgenomen kunnen worden om de vaardigheid van de leerlingen verantwoord te kunnen meten.

Hoe wordt de toets afgenomen?

- Wanneer gekozen kan worden tussen een individuele of groepsgewijze toetsafname, zal om praktische redenen groepsgewijze afname altijd de voorkeur verdienen. Groepsgewijze afname gaat meestal gepaard met schriftelijke toetsen. Hiermee worden toetsen bedoeld waarbij de antwoorden op papier gezet moeten worden. Merk op dat dit laatste ook kan gelden voor toetsen die niet in schriftelijke vorm aangeboden kunnen worden, bijvoorbeeld luistertoetsen. Het is ook mogelijk om de vragen via een beeldscherm te presenteren, de antwoorden in de computer in te voeren en te laten scoren. Door deze mogelijkheid wordt individuele toetsafname niet alleen minder bezwaarlijk maar kan voor bepaalde toepassingen zelfs grote voordelen hebben.

Wat is de inhoud van de toets?

- Het vaststellen van de inhoud van de toets is de belangrijkste toetsspecificatie. Voor deze specificatie wordt bij studietoetsen gebruik gemaakt van een toetsmatrijs die meestal twee-dimensionaal is. Bij de eerder genoemde plaatsingstoetsen Engels bestaat de ene dimensie uit zes inhoudscategorieën die aangeven wat een vraag meet (de betekenis van enkele zinnen, relaties tussen alinea's e.d.). De andere dimensie bestaat uit zes gedragscategorieën die aangeven wat een leerling moet kunnen om het goede antwoord op een vraag te kunnen geven (gegevens combineren en vergelijken, conclusies trekken e.d.). Aan de hand van de toetsmatrijs wordt vastgesteld hoe de vragen uit de toets verdeeld zullen worden over de inhouds- en

gedragscategorieën. De toetsen die op basis van de toetsmatrijs geconstrueerd worden, zijn doorgaans een afspiegeling van hetgeen onderwezen is. Dit laatste kan op verschillende manieren (bijv. curriculum- en functieanalyse) onderzocht worden. In het geval van de plaatsingstoetsen Engels werd aan docenten gevraagd of de vakonderdelen waarop de opgaven betrekking hadden door de docent behandeld waren.

In welke vorm wordt de toets afgenomen?

- Wanneer de vaardigheid met een schriftelijke toets gemeten kan worden, zullen meestal gesloten vragen of open vragen gebruikt worden. Een gesloten vraag is een vraagtype waarbij een persoon uit twee of meer alternatieven of antwoordmogelijkheden het goede antwoord moet kiezen. Vanwege het laatste zou het trouwens juister zijn om de term 'gesloten-antwoord vraag' te gebruiken. De open vraag, ofwel de 'open-antwoord vraag', is een vraagtype waarbij een leerling het antwoord zelf moet formuleren. Studietoetsen, bijvoorbeeld schriftelijke examens, bestaan veelal uit subtoetsen of clusters van vragen die structureel bij elkaar horen. Zo bestaan de schriftelijke examens voor de moderne vreemde talen gewoonlijk uit vijf subtoetsen: vijf teksten waarover tien vragen gesteld worden. In de Engelstalige psychometrische literatuur wordt een subtoets aangeduid met de term 'testlet'. Over de voor- en nadelen van beide vraagtypen is veel gepubliceerd. Als voordelen van gesloten vragen worden genoemd dat men in relatief korte tijd veel vragen kan afnemen en dat die vragen machinaal scoorbaar zijn. Nadelen zouden zijn dat het goede antwoord geraden kan worden en dat de hogere cognitieve vaardigheden niet met gesloten vragen gemeten zouden kunnen worden. Dit laatste zou wel mogelijk zijn met open vragen. Nadelen van open vragen zouden zijn dat er vaak maar weinig vragen voorgelegd kunnen worden en dat de antwoorden beoordeeld moeten worden door beoordelaars die het vaak niet met elkaar eens zijn. Dit laatste komt in hoofdstuk 12 aan de orde bij de bespreking van een toets die slechts uit één open vraag bestaat, namelijk de samenvattingsopdracht. Voor het meten van psychomotorische vaardigheden zoals autorijden, typen en timmeren, kan de motorische component niet met een schriftelijke toets gemeten worden. Bij deze zogenaamde 'performance tests' zal de opdracht of toetsvorm veelal gelijk zijn aan de situatie waarin het geleerde moet worden toegepast.

Hoe worden de vragen of opdrachten gescoord?

- We kunnen bij het scoren van vragen een onderscheid maken tussen dichotome en polytome scoring. Bij dichotome scoring wordt uitsluitend aan het goede antwoord een puntenaantal, meestal één scorepunt, toegekend. Bij polytome scoring wordt ook aan een antwoord dat gedeeltelijk goed is een puntenaantal toegekend. Bij de

beoordeling van de antwoorden op open vragen en opdrachten wordt veelal gebruik gemaakt van een antwoordmodel dat de antwoorden en de bij de verschillende antwoorden behorende aantallen scorepunten bevat. Een antwoordmodel is bedoeld om tot een objectieve beoordeling te komen, dat wil zeggen een beoordeling waarbij het aantal toegekende scorepunten onafhankelijk is van de persoon die beoordeelt. In hoofdstuk 12 wordt beschreven hoe de objectiviteit van een antwoordmodel onderzocht kan worden.

Hoeveel items moeten geconstrueerd worden?

- Ook het antwoord op deze vraag is van een groot aantal factoren afhankelijk. In welke mate wil men dat de onderscheiden categorieën uit de toetsmatrijs bevraagd worden? Hoeveel vragen blijven bij een bepaald vak gewoonlijk over na een proeftoets? Hoeveel toetsversies moeten er geconstrueerd worden?

Wat zijn de gewenste psychometrische kenmerken van de items en de toets?

- Afhankelijk van het doel van de toets zullen de items en de bijbehorende toets andere kenmerken dienen te hebben. Aan toetsen die bedoeld zijn om de docent te informeren over de voortgang van de leerlingen zullen andere eisen gesteld worden dan aan toetsen die bedoeld zijn om beleidmakers te informeren over stand van zaken in het basisonderwijs. Wanneer de toets bedoeld is voor het selecteren van goede leerlingen, zal de toets moeilijker items moeten bevatten dan wanneer de toets bedoeld is voor het selecteren van zwakke leerlingen. In verschillende hoofdstukken wordt uitgebreid ingegaan op de relatie tussen toetsdoel en kenmerken van items en toetsen.

Fase 3: Itemconstructie

Vragen en opdrachten worden ontwikkeld door teams van vakinhoudelijke deskundigen. Daarbij kan het zo zijn dat er één persoon is die de itemspecificaties formuleert, terwijl anderen de items feitelijk schrijven. Recepten voor hoe itemschrijvers goede items kunnen maken bestaan er niet. De verwachting is dat als gevolg van de toegenomen mogelijkheden op automatiseringsgebied het ambachtelijke karakter van dit aspect van het constructieproces in de toekomst zal veranderen.

Fase 4: Toetsafname

We moeten bij toetsafname een onderscheid maken tussen een try-out of proefname en de definitieve toetsafname. Een proefafname is bedoeld om een indruk te krijgen van hoe de items inhoudelijk en psychometrisch functioneren bij de leerlingen waarvoor de definitieve toets bedoeld is. Op basis van de resultaten van de proefafname zullen sommige items verwijderd of gereviseerd worden. Na revisie zal er opnieuw een proefafname moeten plaatsvinden. Het aantal leerlingen waaraan de toets voorgelegd wordt, is bij een proefafname kleiner dan bij een definitieve toetsafname. Voor toetsen die voor onderzoeksdoeleinden gebruikt worden, bijvoorbeeld peilingsonderzoek, laat men om praktische redenen de proefafname soms achterwege en vindt er alleen een definitieve toetsafname plaats. Dit laatste betekent wel dat de toetsafname zeer goed voorbereid dient te worden.

Het is essentieel belang dat de toets onder gestandaardiseerde condities afgenomen wordt. Standaardisatie houdt in dat de toets door alle leerlingen onder gelijke omstandigheden uitgevoerd wordt. Alleen dan is het mogelijk de toetsprestaties van leerlingen met elkaar te vergelijken. Wanneer in dit boek over toetsen gesproken wordt, worden altijd gestandaardiseerde toetsen of meetinstrumenten bedoeld.

Fase 5: Itemevaluatie

Methoden voor het evalueren van items kunnen in twee categorieën verdeeld worden. De eerste categorie bestaat uit kwalitatieve methoden voor het evalueren van de inhoud van items. De Groot en van Naerssen (1973, p. 69) bespreken zes eisen waaraan gesloten vragen moeten voldoen. Gesloten vragen moeten objectief zijn, wat inhoudt dat verschillende vakdeskundigen hetzelfde alternatief als het juiste aanwijzen. Een andere eis is die van specificiteit. Een vraag is specifiek voor een bepaalde leerstof wanneer alleen leerlingen die de leerstof bestudeerd hebben de vraag kunnen oplossen. Kwantitatieve methoden voor het analyseren van antwoorden op items, bijvoorbeeld voor het bepalen van hoe moeilijk een item is, worden met name in de hoofdstukken 3, 4 en 5 behandeld.

Fase 6: Toetssamenstelling

Voor het kunnen selecteren van vragen is het nodig dat zowel kwalitatieve kenmerken, bijvoorbeeld leerstofcategorieën, als kwantitatieve kenmerken, bijvoorbeeld moeilijkheidsgraad, van de items bekend zijn. De mogelijkheden voor selectie worden uiteraard

bepaald door de omvang van de verzameling items. Wanneer de verzameling uit een groot aantal items bestaat die van kwalitatieve en kwantitatieve kenmerken voorzien zijn, spreekt men van een itembank. Itembanken zijn vaak onderdeel van een zogenaamd toetsservicesysteem, een geautomatiseerd stelsel van voorzieningen voor het opslaan, terugzoeken en selecteren van items, het samenstellen van toetsen en het analyseren van toetsresultaten. Methoden voor het selecteren van items gegeven de kenmerken waaraan de toets moet voldoen, worden in hoofdstuk 11 besproken.

Fase 7: Referentiekader

In deze fase wordt de wijze van rapporteren van de scores vastgesteld. De scores die op een toets behaald worden, hebben op zichzelf geen betekenis. De score die een leerling behaalt, krijgt pas betekenis wanneer die score vergeleken wordt met een bepaalde standaard of met de scores die andere leerlingen behaald hebben. De rapportage van scores wordt in hoofdstuk 13 behandeld.

Fase 8: Handleiding en verantwoording

Deze laatste fase bestaat uit het maken van handleiding en instructies voor de diverse categorieën personen die bij de toetsing betrokken zijn. Ten behoeve van de opdrachtgever en het wetenschappelijk forum dient een verantwoording geschreven te worden. In de eerder genoemde Richtlijnen en de Documentatie staan de eisen beschreven waarop toetsmateriaal, handleiding en verantwoording beoordeeld worden.

1.3 Het valideren van meetinstrumenten

Het hoofdstuk over validiteit in de Richtlijnen (1988), een vertaling van de Amerikaanse 'Standards for educational and psychological testing' (1985), nemen we als uitgangspunt voor onze bespreking van validiteit. Het hoofdstuk opent met "Bij de beoordeling van een test verdient de validiteit de meeste aandacht. Validiteit heeft te maken met de betekenis ('meaningfulness'), de bruikbaarheid ('usefulness') en de juistheid ('appropriateness') van de conclusies ('inferences') die uit testskores worden getrokken. Het valideren van een test is het verzamelen van gegevens met de bedoeling

na te gaan of deze conclusies juist zijn. Uit de testskores kunnen verschillende soorten conclusies worden getrokken en er bestaan veel manieren om informatie te verzamelen ter ondersteuning van elke gevolgtrekking. Validiteit is een overkoepelend begrip (’unitary concept’) dat in deze grote verscheidenheid struktuur aanbrengt. De gevolgtrekkingen (’consequences’) bij een specifieke toepassing worden gevalideerd, niet de test" (p. 11). Merk op dat we om de rest van deze paragraaf beter te kunnen begrijpen, bij een aantal begrippen de oorspronkelijke Engelse termen achter de Nederlandse vertaling vermeld hebben.

Over het inzicht dat in de laatste zin van het citaat staat en dat we te danken hebben aan Cronbach (1971, p. 447) bestaat algemeen consensus. Drenth en Sijtsma (1990) bijvoorbeeld omschrijven de validiteit van een test als "...de mate waarin de test aan zijn doel beantwoordt" (p. 173). Om het belang van dit inzicht nog eens te benadrukken geven we de omschrijving van De Groot en van Naerssen (1973): "De validiteitsvraag heeft altijd -bij definitie - betrekking op de mate waarin dat instrument beantwoordt aan het doel waarvoor het wordt gebruikt. Bij studietoetsen is dat doel in het algemeen: bepalen, ’meten’, van de stand van zaken van kennis en inzicht van leerlingen, op een bepaald gebied" (p. 30). Uit het voorgaande en de rest van het citaat uit de Richtlijnen kunnen we twee conclusies trekken.

De eerste conclusie is dat we niet kunnen spreken van de validiteit van een test, maar dat afhankelijk van het doel van de toets, de toets meer of minder valide kan zijn. De tweede conclusie is dat we voor het onderbouwen van de validiteit gegevens dienen te verzamelen. In de Richtlijnen worden drie manieren voor de onderbouwing van de validiteit van een toets onderscheiden: inhoudsvaliditeit, criteriumvaliditeit en begripsvaliditeit. In de Standards worden deze begrippen respectievelijk aangeduid met ’content-related’, ’criterion-related’ en ’construct-related evidence of validity’.

De belangrijkste theoretici op het gebied van validiteit, Cronbach (1971) en Messick (1989), zijn evenals de Richtlijnen van mening dat "Validiteit is een overkoepelend begrip dat in deze grote verscheidenheid struktuur aanbrengt", maar hebben kritiek op de wijze waarop de Richtlijnen daar vervolgens invulling aan geeft door drie soorten validiteit te onderscheiden. Aanleiding voor de kritiek was de toelichting bij de eerste richtlijn. Deze toelichting (Richtlijnen, 1988) luidt: "Het hangt van de aard van de vraagstelling, de kontext en de omvang van eerder verkregen bewijsmateriaal af of één of meer soorten validiteitsgegevens vereist zijn " (p. 19). De bezwaren van onder andere Messick (1988) vloeien voort uit zijn opvatting van validiteit die hij aldus verwoord heeft: "The heart of the unified view of validity is that appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force is empirically grounded construct interpretation. Thus from the

perspective of validity as a unified concept, all educational and psychological measurement should be construct-referenced because construct interpretation undergirds all score-based inferences - not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores. As a consequence, although construct-related evidence may not be the whole of validity, there can be no validity without it. That is, there is no way to judge responsibly the appropriateness, meaningfulness, and usefulness of score inferences in the absence of evidence to what the scores mean" (p. 35). Als gevolg van de toelichting bij de eerste richtlijn vreest Messick (1988) dat: "But the comment also leaves the door open for an interpretation that there exist circumstances under which one kind of validity evidence - be it content-related, for example, or criterion-related - may be adequate and fitting for an applied purpose" (p. 35).

Wat de Richtlijnen onder inhoudsvaliditeit en criteriumvaliditeit verstaan en waarom deze onvoldoende zijn voor het valideren van meetinstrumenten lichten we nu toe. Voor het onderbouwen van de inhoudsvaliditeit van een toets zijn volgens de Richtlijnen gegevens nodig die aantonen dat de steekproef van vragen waaruit de toets bestaat representatief is voor wat men wil toetsen. Zoals we eerder zagen was die onderbouwing bij de plaatsingstoetsen Engels gebaseerd op het oordeel van docenten. Een analyse van de inhoud alleen is volgens Shepard (1993, p. 414) echter onvoldoende om daarmee de validiteit van een toets te verdedigen, omdat er altijd onverwachte effecten zijn die de bedoelde relatie tussen testscore en het begrip of construct kunnen verstoren. Zij geeft een voorbeeld dat ontleend is aan onderzoek met betrekking tot plaatsingstoetsen. De inhoud van deze toetsen was gebaseerd op zorgvuldige curriculum specificaties. Empirisch onderzoek liet echter zien dat er aanzienlijke sexe-verschillen waren. De subtoetsen die uit meerkeuzevragen bestonden waren relatief gemakkelijker voor de mannen terwijl de subtoetsen die uit open vragen bestonden relatief gemakkelijker waren voor de vrouwen. Dit betekent dat onderdelen van de toetsen bij mannen een andere vaardigheid meten dan bij vrouwen en men moet zich dan ook de vraag stellen of de validiteit van die toets nog wel verdedigbaar is. Voornoemde opvatting van inhoudsvaliditeit wijkt nogal af van die van Ebel (1983) die van mening is dat inhoudsvaliditeit de enige validiteit is voor toetsen die na afloop van onderwijs of training afgenomen worden.

Voor het onderbouwen van de criteriumvaliditeit van een toets zijn volgens de Richtlijnen gegevens nodig die de samenhang aantonen tussen de testscores met een criterium. Criteriumvaliditeit is vooral belangrijk voor toetsen bedoeld voor selectie- en plaatsingsbeslissingen, omdat die beslissingen expliciet gebaseerd zijn op de relatie tussen de prestatie op de toets en de prestatie op het criterium. De criteriumvaliditeit

van bijvoorbeeld een plaatsingstoets moet dan ook onderbouwd worden door het aantonen van een empirische relatie tussen de scores op de plaatsingstoets en het succes van de plaatsingsbeslissingen. Afgezien van het feit dat het grootste probleem bij het onderzoek naar de criteriumvaliditeit van toetsen paradoxaal genoeg het ontbreken van valide criteria is, zijn empirische relaties met externe criteria noodzakelijk maar niet voldoende voor het onderbouwen van de validiteit van een toets (Shepard, 1993, p. 411). De hedendaagse opvatting van validiteit (= begripsvaliditeit), vereist dat niet alleen de relevantie en de integriteit van de criteriummaten geëvalueerd wordt, maar dat de voorspellingen zelf ook verdedigd worden. Toetsconstructeurs moeten kunnen verklaren waarom de toets voorspelt en waarom we op die relatie kunnen vertrouwen bij het nemen van beslissingen.

Voor het onderbouwen van de begripsvaliditeit zijn volgens de Richtlijnen gegevens nodig die de betekenis van de testscore duidelijk maken. Voor een toets tekstbegrip kan die onderbouwing bijvoorbeeld bestaan uit empirisch vastgestelde relaties met andere relevante meetinstrumenten, een zogenaamd nomologisch netwerk (Cronbach & Meehl, 1955), dat de betekenis of begripsvaliditeit van de toets duidelijk maakt. Dit is het geval wanneer de toets hoog correleert met soortgelijke toetsen (soortgenootvaliditeit) maar laag correleert met andere toetsen. Bij hoge correlaties spreken we van confirmerende validiteit en bij lage correlaties van discriminante validiteit.

Begripsvaliditeit kan op vele manieren (bijv. logische en empirische analyse, correlationeel en experimenteel onderzoek) en met vele analysetechnieken (bijv. multivariate analyse) onderzocht worden. Voor een overzicht van die manieren en technieken verwijzen we naar Messick (1989, p. 49 e.v.). Hier volstaan we met het noemen van twee analysetechnieken. De eerste is de multitrek-multimethode-benadering van Campbell en Fiske (1959). De tweede analysetechniek betreft psychometrische modellen waarmee de interne structuur of dimensionaliteit van toetsen onderzocht kan worden. In hoofdstuk 5 worden een aantal mogelijke modellen besproken.

Hoewel enerzijds iedereen de opvatting deelt dat bij een beoordeling van een test de validiteit de meeste aandacht verdient, moet anderzijds ook geconstateerd worden dat begripsvalidatie van toetsen op de manier zoals hiervoor en bij Shepard (1993, p. 432 e.v.) beschreven is, in de praktijk niet of nauwelijks voorkomt. Shepard (1993, p. 407) spreekt zelfs van een kloof tussen validiteitstheorie en toetspraktijk. Deze kloof is volgens Kane (1992) te wijten aan het ontbreken van praktische richtlijnen voor het valideren van toetsscores. Hij stelt de ’argument-based approach to validity’ voor en licht deze benadering toe met een plaatsingstoets wiskunde. Op deze benadering gaan we hier verder niet in.

Aan het eind van deze paragraaf willen we toelichten waarom in dit boek geen afzonderlijk hoofdstuk aan validiteit gewijd is. Zoals de bespreking van validiteit heeft laten zien, wordt onderzoek naar validiteit in het algemeen uitgevoerd met in de sociale wetenschappen algemeen bekende onderzoeksmethoden en analysetechnieken. Die methoden en technieken worden in vele uitstekende boeken meer uitgebreid behandeld dan in het kader van dit boek mogelijk geweest zou zijn. Van een behandeling van die methoden en technieken is dan ook afgezien. In dit boek beperkt validiteitsonderzoek zich tot onderzoek waarbij psychometrische technieken een rol spelen. Met name in de hoofdstukken 5 en 9 komen psychometrische modellen en technieken voor validiteitsonderzoek aan de orde.

1.4 Psychometrie in de praktijk

Het meest essentiële kenmerk van een toets als meetinstrument is dat het resultaat van de meting feilbaar is. De resultaten op toetsen zijn, zoals iedereen wel eens ervaren zal hebben, onderhevig aan allerlei toevalsfactoren. Een agglomeraat van toevalsfactoren in de condities waaronder getoetst wordt, in de persoon die getoetst wordt en ook in het meetinstrument zelf, maakt dat de metingen met toetsen nooit exact zullen kunnen zijn. Het zal ook duidelijk zijn dat de waarde van de informatie, die gebaseerd is op resultaten gemeten met deze instrumenten, en de rol die deze informatie kan spelen in het eerder beschreven toetsconstructieproces staat of valt met de nauwkeurigheid hiervan. Het aandachtsgebied van de psychometrie als toegepaste wetenschap is altijd geweest aan de gebruiker van meetinstrumenten de nauwkeurigheid van metingen zichtbaar te maken en die gebruiker methoden aan te bieden om de kwaliteit van meetinstrumenten te beoordelen. Vaardigheden die niet nauwkeurig gemeten worden, kunnen ook niet valide zijn. Dat wil niet zeggen dat nauwkeurige metingen ook valide metingen zijn. Meetnauwkeurigheid is een noodzakelijke maar geen voldoende voorwaarde voor validiteit.

Zoals we reeds eerder opmerkten richt de psychometrie zich op die aspecten van het toetsconstructieproces waarbij gebruik gemaakt wordt van empirische gegevens. In hoofdstuk 2 wordt een aantal algemene begrippen besproken die bij het verzamelen van deze gegevens een rol speelt. In de psychometrie bestaan die empirische gegevens in ieder geval uit kwantificeringen van kenmerken van personen die op zijn minst de aan- of afwezigheid van het kenmerk indiceren. Doorgaans zijn de te analyseren gegevens echter veel rijker. Bij toetsscores duidt de hoogte van de score op zijn minst ook de mate van aanwezigheid van het kenmerk van de persoon aan. De kenmerken die we

willen bestuderen, zijn doorgaans niet direct waarneembaar. De variabelen waarin we feitelijk geïnteresseerd zijn noemen we latent. De theorieën in de psychometrie leggen relaties tussen latente variabelen en geobserveerde variabelen. De rekenvaardigheid van een leerling kunnen we slechts proberen vast te stellen door de antwoorden op waarneembare indicatoren van dit kenmerk, bijvoorbeeld rekenopgaven, te beschouwen. De notie dat de observaties nooit een exacte weergave zullen zijn van de werkelijke aanwezigheid van een kenmerk, maakt dat psychometrische theorieën zich bedienen van formele beschrijvingssystemen die rekening houden met toevalsfactoren. De gebruikte modellen zijn dan ook probabilistische of stochastische modellen. De methoden en technieken die bij de ontwikkeling van modellen en bij het analyseren van gegevens worden gebruikt en die we in dit boek zullen beschrijven, maken deel uit van wat in de wiskunde bekend staat als de toegepaste statistiek.

De psychometrie bestond tot halverwege deze eeuw alleen uit de klassieke testtheorie. Een eerste volledige behandeling is te vinden in Gulliksen (1950). Een formeel volledige beschrijving en een aantal uitbreidingen vinden we in het boek van Lord en Novick (1968) dat nu nog steeds het standaardwerk van deze theorie is. Het uitgangspunt van de theorie is dat de geobserveerde score van een persoon op een toets de som is van een ware score, de waarde van een niet waarneembare variabele waarin we geïnteresseerd zijn, en een niet systematische, niet controleerbare meetfout. In de theorie worden deze begrippen preciezer gedefinieerd en veronderstellingen gedaan omtrent het stochastische karakter van de meetfout. In het werken met het klassieke testmodel hebben we uiteraard altijd te maken met toetsscores van meerdere personen, waarvan dan aangenomen wordt dat deze aselect getrokken zijn uit een of andere populatie. De statistiek die we in deze theorie gebruiken, generaliseert dan naar deze populatie van personen. Het primaire doel van de klassieke testtheorie is een beschrijving te geven van de nauwkeurigheid van de metingen. In de klassieke testtheorie staan daarvoor de begrippen betrouwbaarheid en standaardmeetfout centraal. Na Lord en Novick (1968) is de formele klassieke testtheorie nog nauwelijks uitgebreid. Ingegeven door de theoretisch enigszins magere fundering van het klassieke testmodel, maar ook door zijn inherente beperkingen en praktische problemen, kwam de moderne testtheorie, genaamd itemresponstheorie of latente trek theorie, tot ontwikkeling. Dat wil echter niet zeggen dat de klassieke testtheorie inmiddels volledig vervangen is door deze moderne theorie. De klassieke testtheorie heeft zoveel bruikbare methoden en technieken opgeleverd die kunnen bijdragen aan de kwaliteitsbeheersing van toetsen, dat met name in de tegenwoordige psychometrische praktijk nog veelvuldig gebruik gemaakt wordt van de klassieke testtheorie. Deze

theorie zal daarom in hoofdstuk 3 worden behandeld en ook in verschillende andere hoofdstukken ruime aandacht krijgen.

Alvorens in te gaan op de moderne testtheorie staan we even stil bij theorieën die we kunnen beschouwen als belangrijke uitbreidingen van de klassieke testtheorie. Op de eerste plaats is dat de generaliseerbaarheidstheorie (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In tegenstelling tot de klassieke testtheorie kunnen in de generaliseerbaarheidstheorie verschillende foutenbronnen onderscheiden worden. De generaliseerbaarheidstheorie biedt dan ook de mogelijkheid verschillende 'betrouwbaarheden' te schatten. De theorie wordt in hoofdstuk 3 behandeld en in hoofdstuk 11 toegepast.

Andere uitbreidingen van de klassieke testtheorie zijn modellen waarbij er sterkere aannames over de meetfouten worden gedaan dan in het klassieke testmodel. Bekende modellen die met een gespecificeerde verdeling van de meetfouten werken zijn het binomiale-foutenmodel en het poisson-foutenmodel. Deze modellen die onder andere in Lord en Novick (1968) beschreven worden, zullen we in dit boek niet behandelen omdat de toepassing in de huidige psychometrische praktijk slechts incidenteel is.

In de moderne testtheorie met als startpunten Lord (1952) en Rasch (1960) wordt niet de score op een toets, samengesteld uit de scores op de items, gemodelleerd, maar wordt een expliciet model aangenomen voor de respons op elk afzonderlijk item. De kans dat een persoon een bepaalde respons op een item geeft, is een gespecificeerde functie van de te meten latente variabele van de persoon, de vaardigheidsparameter, en één of meerdere itemparameters. De itemresponstheorie heeft veel van de bezwaren van de klassieke testtheorie weggenomen. In de itemresponstheorie bestaat, in tegenstelling tot de klassieke testtheorie, de mogelijkheid de geldigheid van het aangenomen model expliciet te toetsen. Daarnaast zijn de itemkarakteristieken onafhankelijk van de specifieke toets waarin de items zitten. Bovendien levert de theorie methoden en technieken die nieuwe toepassingen van de psychometrie mogelijk maken. Was de klassieke testtheorie volledig geconcentreerd op het resultaat van de meting, in de itemresponstheorie zijn er veel meer mogelijkheden om te onderzoeken hoe dit resultaat tot stand is gekomen.

De toepassingsmogelijkheden van de eerste itemresponsmodellen zijn beperkt. Het zijn modellen die uitgaan van dichotoom gescoorde items en die zulke strenge eisen aan de responsen opleggen, dat in veel praktijkgevallen het model als ongeldig moest worden verklaard. Heden ten dage echter zijn de modellen op allerlei manieren uitgebreid. Er zijn modellen met meer itemparameters en de beperking tot dichotoom gescoorde items is vervallen. Daar komt bij dat de analyses in de itemresponstheorie hogere statistische en rekentechnische eisen stellen dan de analyses in de klassieke

testtheorie. Pas na enkele decennia werk van een groot aantal psychometrici en door de enorme ontwikkelingen op computergebied, heeft de itemresponstheorie ook een zeer belangrijke plaats in de psychometrische praktijk gekregen. Een verschuiving van wat Van der Linden (1983) noemt het klassieke complex, het werken met gestandaardiseerde toetsen en de klassieke testtheorie, naar het moderne complex, het werken met itembanken en itemresponstheorie, is waar te nemen.

In hoofdstuk 4 zal een uitvoerige inleiding worden gegeven in de basisconcepten en de schattings- en toetsingsmethoden in de itemresponstheorie. Dit zal worden besproken aan de hand van het model van Rasch (1960). In hoofdstuk 5 wordt een overzicht gegeven van uitbreidingen van het Raschmodel en andere itemresponsmodellen. Aparte aandacht krijgt, met name vanwege het grote belang voor de praktijk, de itemresponstheorie in zogenaamde onvolledige gegevensverzamelingen. Enkele concrete toepassingen van itemresponstheorie worden in hoofdstuk 7 behandeld.

Omdat toetsen vaak gebruikt worden om beslissingen te nemen over personen kan een besliskundige benadering van de psychometrie ook zeer vruchtbaar zijn. Wij zullen om praktische redenen deze benadering niet expliciet behandelen. Voor een overzicht van de besliskundige testtheorie verwijzen wij naar Van der Linden (1985).

In hoofdstuk 8 tot en met 10 worden problemen uit de praktijk besproken die met behulp van de itemresponstheorie worden opgelost. Achtereenvolgens komen daarbij de volgende onderwerpen aan de orde: het equivaleren van toetsen, vraagonzuiverheid en het meten van veranderingen. Hierbij worden, evenals in het volgende hoofdstuk, zowel oplossingen met behulp van de klassieke testtheorie als de itemresponstheorie besproken. Hoofdstuk gaat over het samenstellen van optimale toetsen met behulp van mathematische programmering. De beoordeling van niet zonder meer objectief scoorbare toetsen of opdrachten is het onderwerp van hoofdstuk 12. Zoals elk toetsconstructieproces, en trouwens ook elke toets, wordt dit boek afgesloten met een behandeling van de rapportage van de toetsresultaten.