Niet alles wat meetbaar is, is relevant en niet alles wat relevant is, is meetbaar. Toetsen in onderwijs en onderzoek, een overzicht.

Ik ben gevraagd om een praatje te houden vandaag over het toetsen van kennis en vaardigheden in een schoolse setting. Zelf ben ik lang docent methoden en technieken van sociaal wetenschappelijk onderzoek geweest aan de Universiteit van Amsterdam en ook alweer 30 jaar onderzoeker bij het Kohnstamm Instituut van dezelfde universiteit. Daarnaast ben ik nu ruim 5 jaar lector bij Kenniscentrum Talentontwikkeling van Hogeschool Rotterdam. Ik zal proberen in het verhaal dat ik hier houd vooral aandacht te geven aan aspecten van metingen in het onderwijs, maar gezien dat veel van de hier aanwezigen ook Minor- en afstudeeronderzoek zullen gaan doen, voel ik me vrij om ook een enkel zijpad te bewandelen dat meer de wetenschappelijke kant van meten en toetsen betreft. Daarnaast hebben beide terreinen veel met elkaar te maken, of beter, kan het schoolse toetsen veel opsteken van de wetenschappelijke kijk op toetsen. Andersom lijkt me dat minder het geval.

Inleiding

Zoals de titel van dit praatje al aangeeft, is niet alles meetbaar. Denk bijvoorbeeld aan toekomstig gedrag of gevoel: hoe proefpersonen zouden omgaan met pijn gerelateerd aan een dodelijke kwaal waar geen genezing voor is, of zij ja dan nee euthanasie zouden laten toepassen, of de mate van verdriet die iemand zou voelen bij het sterven van zijn of haar geliefde. Maar ook zaken die niet in de toekomst liggen, zoals bijvoorbeeld intelligentie, gedefinieerd als het vermogen een nooit eerder gezien probleem op te lossen, is niet 'echt' te meten. Eigenlijk geldt dat elke meting van wat ook, vrijwel nooit perfect weergeeft wat we beogen te meten. Dit geldt uiteraard in grote mate voor min of meer abstracte constructen als genegenheid, angst, intelligentie of kunstreceptie, maar ook aspecten die niet abstract zijn (of lijken) kunnen we vaak niet perfect meten. Leeftijd of lichaamslengte lijken exact te bepalen. Echter, ook bij dit soort metingen treden fouten op. Mensen zijn langer als ze vlak voor de lengtemeting lang op bed hebben gelegen. Kennis die in mijn jeugd dankbaar werd gebruikt door jongens die net wel of net niet de lengte hadden waarboven je voor militaire dienst werd afgekeurd. Daarnaast is de meeteenheid moeilijk te standaardiseren. Om die reden ligt in Parijs ergens een edelmetalen staaf die bij een constante temperatuur en gasdruk weergeeft hoe lang een meter is. En zelfs leeftijd kan moeilijk exact bepaald worden voor iedereen. Waar begint iemands leven? Bij de conceptie? Bij de geboorte? En wat als de geboorte via een keizersnee verliep? Daarnaast worden er fouten gemaakt bij het registreren van de uitkomsten van een meting. Ook ontstaan er fouten doordat de schaal waarop we meten niet oneindig precies is. We kunnen lengte meten in meters, in centimeters, in millimeters, in micrometers, maar hoe fijnmazig we ook gaan, echt continu worden scores nooit. Tevens is bekend dat indien mensen weten dat ze gemeten worden, ze zich anders gaan gedragen. Ook voor zaken die we in de wetenschap proberen te meten als leervermogen, taalvaardigheid, maatschappelijke weerbaarheid of kennis van kunst, geldt dat iedere meting slechts een benadering is van het aspect dat we willen bepalen. En elke onderzoeker weet, of hoort te weten, dat die benadering slechts een meer of minder goede afspiegeling vormt van de beoogde karakteristiek. Ooit, in mijn jeugd, leidde dit besef tot een pleidooi om in wetenschappelijke verslagen niet meer te spreken over een construct dat was gemeten, maar over een specifieke toets die was afgenomen; het operationalisme. Het loslaten van elke pretentie op algemene geldigheid echter, had ook een prijs. Niemand is echt geïnteresseerd in uitslagen op specifieke toetsen. We willen nu juist generaliseren, niet alleen naar populaties, maar ook naar theorieën die niet over specifieke toetsen gaan, maar over theoretische begrippen.
Uit het voorgaande wordt hopelijk duidelijk dat veel aspecten bij het meten voor problemen kunnen zorgen. Om niet door de bomen het bos kwijt te raken, zijn er in de methodeleer een aantal begrippen geïntroduceerd om enige structuur in het probleem van het meten aan te brengen. Een typisch wetenschappelijke benadering dus, reduceren, opknippen in stukjes, zodat we enig begrip kunnen krijgen van de delen, om zo langzaam vanuit de delen richting holistisch begrip te gaan. Daarnaast kunnen we door het meetproces in stukjes te knippen, toetsen of specifieke kenmerken

van ons meetinstrumentarium aan een vooraf beredeneerde gestelde norm voldoet. Dit laatste betreft dus een Popperiaanse of positivistische of kritisch realistische manier van kijken naar hypotheses die we over ons meetinstrument hebben en waarvan we kunnen nagaan wat de kans is dat die hypotheses waar zijn. Om dat te kunnen doen moeten we dan wel data verzamelen, of in dit geval, scores van studenten of leerlingen op de toetsen. Over die normen en het toetsen daarvan later meer.
Welke delen hanteren we in de methodeleer om meettechniek bespreekbaar te maken en de kwaliteit van meetinstrumenten te kunnen evalueren? Enkele begrippen die ons ten dienste kunnen staan bij het denken over de kwaliteit van metingen, betreffen toetstechniek in het algemeen. Andere begrippen zijn meer gericht op het toetsen in het onderwijs.

De begrippen
Bij het meten van theoretische begrippen in de sociale wetenschappen onderscheiden we de begrippen 'construct', 'operationaliseren', 'variabele', 'betrouwbaarheid' en 'validiteit'. Daarnaast onderscheiden we verschillende vormen van toetsen al naar gelang het doel en de wijze waarop getoetst wordt. Een belangrijk onderscheid op het gebied van de doelen betreft het selectief versus diagnostisch toetsen. Kijken we naar toetsvormen, dan kennen we gesloten en open vormen van toetsen, mondelinge en schriftelijke toetsen. Een voorbeeld van een gesloten toets is bijvoorbeeld een meerkeuze toets, een open toets kan een toets zijn met open vragen, maar ook een portfolio of een schrijfopdracht. Ook de wijze waarop een toetsresultaat wordt bepaald, kan variëren, men kan open toetsen beoordelen aan de hand van beoordelingsvoorschriften, aan de hand van voorbeeldprestaties of schalen, met één of meerdere beoordelaars, enzovoorts. In het nu volgende zal ik proberen één en ander te verduidelijken.

Construct
Het eerste methodologische begrip dat ik hierboven noemde, is het begrip 'construct'. Dit begrip staat voor het te meten kenmerk. Deze kenmerken kunnen personen betreffen, maar ook groepen mensen zoals gegroepeerd in scholen, culturen of religies en zelfs kenmerken van objecten kunnen als construct gezien worden, denk bijvoorbeeld aan elasticiteit, weerstand of roestbestendigheid van metaal. Bij het toetsen in het onderwijs, betreffen de constructen in eerste instantie kenmerken van de studenten of leerlingen. Toetsresultaten worden echter ook gebruikt om kenmerken weer te geven van klassen, scholen of landen, denk aan de gemiddelde CITO-scores per school of het percentage studenten dat binnen een bepaalde termijn afstudeert of de gemiddelde score voor leesvaardigheid van Nederlandse 15-jarigen. Een construct kan vele vormen aannemen, het kan een vaardigheid zijn, kennis of een attitude, een mening, een ziekte, een achtergrondkenmerk van een persoon, of welk meetbaar begrip dan ook dat in een theorie wordt gedefinieerd en dat men wil meten. In het onderwijs betreffen de te meten constructen natuurlijk vaak kennis, vaardigheden en attitudes en motivatie.

Operationaliseren
Nadat bepaald is wat men wil meten, moet gekozen worden voor de wijze waarop dat gebeurt. De theoretische definitie van het construct moet worden vertaald naar een operationele definitie, ofwel een omschrijving van hoe het construct kan worden bepaald en welke cijfers of codes moeten worden toegekend aan de uitslagen van de meting, de scores. Een operationalisatie is dus de wijze of het proces waarmee we proberen een theoretisch construct ofwel begrip, meetbaar te maken. Voorafgaand aan een operationalisatie moeten we definiëren wat ons construct of begrip precies is. Naast de theoretische definitie van ons construct via begripsanalyse, gesteld in algemene termen, moeten we dan de stap maken naar een operationele definitie. Operationalisaties kunnen veel verschillende vormen aannemen. We kunnen meten door middel van observaties, interviews, open of gesloten vragenlijsten, reactietijden bijvoorbeeld via de gemeten snelheid waarmee een respondent op een knop drukt, documentanalyses, film- of audio-opnamen, biomedische metingen als de galvanische huidrespons, pupilgrootte, hartslag en nog veel meer. Metingen kunnen verricht

worden zodanig dat de persoon waarbij iets gemeten wordt, niet doorheeft dat er gemeten wordt (unobtrusive measures). Metingen kunnen, zoals gezegd, verricht worden aan personen, maar in de wetenschap kunnen onderzoeksobjecten ook andere eenheden zijn dan personen. Men heeft ooit bijvoorbeeld de populariteit van tentoongestelde objecten in een museum gemeten door de slijtage van de vloertegels rond de objecten te bepalen, een voorbeeld van een unobtrusive measure, de bezoekers van het museum waren zich er immers niet van bewust dat hun voorkeuren werden gemeten. In het onderstaande zal ik het vaak over personen hebben als ik het heb over toetsen of over het meten van constructen. Men moet hierbij dan bedenken dat toetsen ook op andere eenheden dan mensen betrekking kunnen hebben.

Variabelen

Als we een operationele definitie hebben gekozen, een meetinstrument hebben gemaakt en de meting hebben verricht, moeten we de uitslagen van de metingen registreren. In het onderwijs doen we dat vaak in een getallenreeks die loopt van 1 tot 10 of, zoals op de CITO-toets voor groep 8 van 500 tot 550, of via woorden als 'voldoende', 'onvoldoende', 'slecht', 'matig', 'voldoende', 'goed', 'zeer goed', etc. In sociaal wetenschappelijk onderzoek en ook in het onderwijs werken we vaak met getallen om de toetsuitslagen weer te geven. Getallen hebben het voordeel dat we ermee kunnen rekenen. Zo wordt het mogelijk om de uitslag van een toets met bijvoorbeeld 10 vragen in één getal uit te drukken, het gemiddelde, en niet in tien losse oordelen. Daar toetsen bedoeld zijn om verschillen tussen mensen of andere eenheden aan te geven, is een voorwaarde voor een zinvolle toetsing dat niet elke kandidaat exact dezelfde score krijgt. In de statistiek noemen we een kenmerk waarop iedereen dezelfde score krijgt een constante, is er verschil, dan hebben we een variabele ofwel een getal dat varieert.

Het lijkt misschien een simpele oplossing voor een complex probleem, we maken een meetinstrument, bijvoorbeeld een papier met vragen waarop een leerling de antwoorden moet geven, we waarderen de juistheid van die antwoorden, bijvoorbeeld 0 punten voor een fout antwoord en 1 punt voor een goed antwoord, en we sommeren per leerling de punten om de toetsuitslag te bepalen. Een dergelijke simpele oplossing gaat echter uit van een aantal relatief complexe aannamen. Door te sommeren nemen we bijvoorbeeld aan dat elke vraag even zwaar weegt. Immers, iemand die de eerste twee vragen goed doet en alle andere fout, krijgt een score van 2. Iemand die echter de eerste 8 vragen fout doet en de laatste twee goed, krijgt ook een 2. Om nu de verschillende aspecten die een rol spelen bij het verifiëren van de deugdelijkheid van een toets, te kunnen bespreken, moet ik helaas een klein beetje statistiek introduceren. Wel beloof ik dat ik de tekst ook voor mensen met 'dyscalculie' begrijpelijk zal proberen te houden.

De statistiek

Omdat, zoals hiervoor al is aangehaald, elke meting imperfect is, is het in de wetenschap gebruikelijk om de deugdelijkheid van de meting te controleren. Hierbij gaat men uit van een simpel model. Men neemt aan dat een score van een leerling op een meting (toets, vraag, kenmerk etc.) bestaat uit twee delen, een stabiel deel, dat bij herhaalde meting opnieuw zal worden gevonden, en een deel ruis, meetfout of error. Mooi statistisch opgeschreven ziet dat er zo uit: $X_i=W_i+e_i$. Deze formule is de basis waarop de gehele klassieke statistiek is gebouwd. In de formule staat dat de score X van proefpersoon i kan worden opgedeeld in een stukje stabiele trek dat die persoon karakteriseert (W_i) en een stukje ruis dat die persoon bij die meting op dat moment 'overkwam' (e_i). Verder is natuurlijk helder dat niet alle leerlingen dezelfde score krijgen, we hebben geen constante, maar een variabele gemeten. Nu is men in de statistiek gewoon niet alleen te kijken naar de score van individuele personen, maar naar de variatie in de scores van groepen personen. De kwaliteit van een meetinstrument wordt bepaald niet aan de hand van de score van één persoon, maar aan de hand van de scores van een groep. Ook de variatie in de scores wordt, analoog aan de hierboven gegeven formule, gedacht te bestaan uit stabiele variatie, die bij opnieuw meten van dezelfde groep personen opnieuw zal worden gevonden, en foutenvariatie, dus ruis, die bij een nieuwe meting niet bij elke persoon dezelfde grootte zal hebben. De totale variatie in een groep wordt overigens bepaald als de

gesommeerde gekwadrateerde afstanden van alle individuele scores tot het groepsgemiddelde. Met behulp van statistiek kunnen we deze totale variatie opdelen in de twee voornoemde delen, stabiele en foutenvariatie.

De betrouwbaarheid

Nu is het eerste aspect van de deugdelijkheid van een meting de proportie van de variatie in de scores van de gemeten leerlingen dat *niet* uit ruis bestaat en kan dus lopen van 0 tot 1. Men noemt dit de betrouwbaarheid van een meting. Een meetinstrument dat een betrouwbaarheid heeft van .80 levert dus scores die 20% ruis bevatten. De norm voor de betrouwbaarheid die in de APA-standards (standaarden van de American Psychological Association) wordt gesteld voor het doen van metingen waarbij de uitslag consequenties heeft voor het gemeten individu zoals cijfers op schooltoetsen, bedraagt momenteel .80. In mijn jeugd stond die norm nog op .90. maar alles wordt minder, zullen we maar zeggen.

Er zijn verschillende wijzen om de betrouwbaarheid van een meetinstrument te bepalen. Eén manier is het tweemaal afnemen van een toets en het berekenen van de samenhang of correlatie tussen de beide reeksen scores. Als we aannemen dat de ruis niet samenhangt met de stabiele scores en niet met de ruis op de andere meting, is deze maat op te vatten als schatter van de proportie stabiele variatie in de scores, ofwel de betrouwbaarheid. Nu zult u natuurlijk meteen hier tegenin brengen dat de studenten van de eerste afname kunnen hebben geleerd en de één meer dan de ander, en dat daardoor de correlatie tussen beide afnamen geen goede schatter van de betrouwbaarheid meer is. Ook kunnen studenten zich wellicht hun eerder gegeven antwoorden herinneren, en ook dat maakt deze test-hertestcorrelatie tot een minder goede maat voor het bepalen van de betrouwbaarheid van de meting. Wat nu? Een mogelijkheid is om de test na één afname aselect in twee helften te knippen en de correlatie te berekenen tussen beide testhelften. Nu is de correlatie van een halve test lager dan die van een hele, maar daar is statistisch voor te corrigeren. We gaan er bij deze split-half benadering wel van uit dat beide testhelften dezelfde trek of dezelfde combinatie van trekken meten. Alweer een aanname dus. Nemen we aan dat alle vragen van een test dezelfde trek meten, dus in het geval van een vaardigheidstoets, dat door de studenten bij het beantwoorden van alle vragen een zelfde oplossingsproces gehanteerd wordt, dan is er nog een andere optie. Men kan dan de gemiddelde split half uitrekenen over alle mogelijke helften die er van een test te maken zijn. Men noemt deze maat ook wel de homogeniteit van een test en de statistiek die hierbij hoort heet Cronbach's alpha. Hierbij moet men wel beseffen, dat er deugdelijke somscores denkbaar zijn, die niet homogeen zijn. Als ik bijvoorbeeld een toets maak voor het meten van de vaardigheid in het spellen van werkwoorden of de affectieve attitude ten aanzien religie, verwacht ik dat alle vragen dezelfde trek meten. Meet ik echter gezinsinkomen door de inkomens van alle gezinsleden bij elkaar op te tellen, dan hoeven de vragen niet homogeen te zijn. Immers, indien de ouders veel geld verdienen, is het eerder onwaarschijnlijk dan waarschijnlijk dat de kinderen van die ouders een krantenwijk nemen om wat bij te verdienen. Toch is de som van de inkomens van alle gezinsleden mogelijk een betrouwbare maat voor de variabele 'gezinsinkomen'.

Goed, we weten nu dus wat betrouwbaarheid van een test is en dat deze .80 of hoger moet zijn als we leerlingen toetsen. Onderwijsinstituten zouden dus de betrouwbaarheid van hun toetsen kunnen laten doorrekenen of even kunnen leren hoe men dat zelf kan doen, zo moeilijk is dat niet. Is de betrouwbaarheid te laag, dan zijn er verschillende opties. Men kan de toets langer maken. Indien het aantal vragen van een toets toeneemt, neemt ook de betrouwbaarheid van de somscore over die vragen toe. Ruis per vraag kan dan immers uitmiddelen in somscores. Uiteraard gaat dit alleen op als de vragen onderling positief samenhangen, of beter nog, dezelfde trek meten. Verder moet altijd bedacht worden dat de betrouwbaarheid van een toets afhankelijk is van de groep mensen waarbij de toets wordt afgenomen. Mijn statistiektentamens zijn wellicht betrouwbaar bij mijn statistiekstudenten na afloop van de cursus, bij u is het echter mogelijk dat de scores op mijn statistiektentamen voor 100% uit ruis zullen bestaan. Ook kan men bij een te lage betrouwbaarheid van de toets gaan zoeken naar welke vragen het niet goed doen. Daar kan men van leren om zo tot het formuleren van betere vragen te komen. Een geëigende techniek hiervoor is exploratieve

factoranalyse, maar zelfs het inspecteren van de correlaties tussen elk individueel item en de somscore over alle items, levert hiertoe relevante informatie. Ook kan als de betrouwbaarheid van een toets te laag uitvalt, statistisch bepaald worden hoeveel vragen van gemiddeld dezelfde kwaliteit moeten worden toegevoegd aan de toets om een acceptabele betrouwbaarheid te bereiken. Het moge duidelijk zijn dat de betrouwbaarheid van een toets evenredig is aan de precisie van de individuele scores. Een hogere betrouwbaarheid betekent minder ruis en dus meer precisie in de scores. Met behulp van de betrouwbaarheid kan dan ook bepaald worden met een vooraf gesteld niveau van zekerheid binnen welke grenzen de gemiddelde score van een groep of zelfs de score van een individu ligt.
In het onderwijs gebruiken we naast toetsen ook vaak docentoordelen om studentprestaties te evalueren. Uit onderzoek blijkt nu dat een oordeel van een docent over een opdracht van een leerling vaak een lage betrouwbaarheid heeft. Godshalk, Swineford en Coffmann (1966) lieten in een klassiek onderzoek zien dat de betrouwbaarheid van opstelbeoordelingen door ervaren moedertaaldocenten gemiddeld .20 bedraagt. Inmiddels zijn er wel pogingen ondernomen om hier wat aan te doen, onder andere door het hanteren van beoordelingsschalen met voorbeeldopstellen die passen bij een specifiek cijfer, het sommeren van oordelen van meerdere onafhankelijke beoordelaars al dan niet met gestandaardiseerde beoordelingsvoorschriften, maar door de bank genomen is de betrouwbaarheid van menselijke oordelen laag.
Nog een opmerking over toetsbetrouwbaarheid: in het onderwijs zijn we bij het toetsen vaak meer geïnteresseerd in of een leerling geslaagd of gezakt is, dan in differentiatie boven of onder de zak-slaaggrens. In dat geval is de betrouwbaarheid van de zak-slaagbeslissing relevanter dan de betrouwbaarheid van de gehele scorerange. Deze zak-slaagbetrouwbaarheid wordt groter naarmate er meer vragen qua moeilijkheidsgraad rond dit niveau zitten. Met andere woorden, als we uitgaan van 20% gezakte kandidaten en de zak-slaaggrens zo willen kiezen dat ongeveer 20% van de leerlingen zakt, dan doen we er goed aan alle vragen te kiezen die een p-waarde hebben van .80. De p-waarde is de proportie leerlingen die een vraag goed beantwoordt. Het stellen van heel moeilijke of heel makkelijke vragen zal de betrouwbaarheid van de zak-slaagbeslissing dus niet ten goede komen. Overigens is er wel een andere reden om een toets met enkele makkelijke vragen te beginnen. De leerlingen krijgen er zelfvertrouwen door en raken meer op hun gemak, krijgen minder testangst. En omdat testangst meestal niet is wat we beogen te meten, neemt zo ook de kans toe dat we meten wat we beogen te meten, ofwel de validiteit van de toets.

De validiteit
En daarmee hebben we een bruggetje geslagen naar het volgende begrip dat ik wil bespreken in verband met deugdelijkheid van toetsen; de validiteit. De validiteit betreft de mate waarin de toets meet wat men wil meten. Deze kan dus nooit hoger zijn dan de betrouwbaarheid, want ruis is sowieso niet wat we willen meten. Als we praten over de validiteit van meetinstrumenten (er zijn andere opties) dan onderscheiden we inhoudsvaliditeit, criteriumvaliditeit en constructvaliditeit. Ook bij schools toetsen zijn deze begrippen van belang.
De inhoudsvaliditeit betreft een niet statistische evaluatie van de inhoud van de toets. Men onderscheidt hierbij enerzijds of de toets op het oog meet wat bedoeld wordt aan de hand van de begripsanalyse en operationele definitie en anderzijds of het gehele domein dat onder het te meten begrip valt, wordt bestreken. Dus bij schools toetsen geldt dat als ik een aardrijkskundetoets maak die de stof van hoofdstuk 1 tot en met 5 uit het leerboek moet toetsen en ik stel alle vragen over stof uit hoofdstuk 3, dan ben ik niet goed bezig. En als ik ook punten aftrek indien leerlingen spelfouten maken in hun antwoorden, verknoei ik de validiteit van mijn aardrijkskundetoets. Beter is in het laatste geval een spellingtoets en een aardrijkskundetoets naast elkaar af te nemen, daar ik anders bij het sommeren appels en peren optel. Gebruik ik moeilijke woorden in toetsvragen, dan zal ik bij leerlingen die sommige worden niet kennen, ook woordenschat meten.
Naast inhoudsvaliditeit, kennen we criteriumvaliditeit, een statistisch criterium. Zoals het woord al aangeeft kijken we bij dit soort validiteit naar of de toetsscores samenhangen met metingen van verwante begrippen (convergente validiteit) en niet of laag met niet verwante begrippen (divergente

validiteit). Neem ik bij kinderen onder de 15 de schoenmaat als indicatie van het IQ of de intelligentie, dan is de inhoudsvaliditeit slecht, maar de criteriumvaliditeit kan redelijk zijn. Intelligentiescores op IQ-toetsen en schoenmaten nemen immers beide toe met de leeftijd. Nu is de schoenmaat hierboven een wat ver gezocht voorbeeld, maar als ik wil meten hoe goed een student een kunstwerk in een stroming kan plaatsen, een productieve vaardigheid dus, en ik doe dat door een meerkeuzetoets af te nemen, dan kan de criteriumvaliditeit redelijk zijn, al weten we dat de inhoudsvaliditeit niet helemaal deugt. Immers, meerkeuzetoetsen meten altijd receptief. Het herkennen van een goed antwoord is niet hetzelfde als het produceren van een goed antwoord.
De hoogste vorm van validiteit is de constructvaliditeit. Deze staat voor de mate waarin men echt meet wat men bedoelt te meten. Men kan leerlingen hardop denkend toetsvragen laten beantwoorden om hierover indicaties te verkrijgen, maar ook statistische modellen toetsen die gebaseerd zijn op de theoretische noties die aan het instrument ten grondslag liggen. Zo kan men nagaan of in het instrument onderscheiden aspecten ook daadwerkelijk gescheiden aspecten zijn en of vragen die één aspect moeten meten dat ook daadwerkelijk doen.

Toetsvormen

Naast de bovengenoemde aspecten van het toetsen, kunnen we ook nog kijken naar toetsvormen. We onderscheiden gesloten en open vormen van toetsen. Een voorbeeld van een gesloten toets is bijvoorbeeld een meerkeuze toets, een open toets kan een toets zijn met open vragen, maar ook een portfolio of een schrijfopdracht. Door onderzoek is bekend dat elke wijze van toetsen zijn eigen vorm van ruis creëert. Men noemt dit ook wel methodespecifieke variatie. Stel ik meet spelvaardigheid met een meerkeuzetoets. Dan meet ik ten eerste alleen het herkennen van goed en fout gespelde woorden (receptief) en niet de vaardigheid in het produceren van correct gespelde vormen. Daarnaast meet ik dan enige raadkans, en meer bij hen die de antwoorden minder vaak weten. Ook meet ik de vaardigheid in het omgaan met meerkeuzetoetsen. Enkele aspecten van het laatste zijn te illustreren met wijsheden als; het langste antwoordalternatief is vaak juist, herhaling van woorden uit de vraag in een alternatief duidt vaak op een goed antwoord, vragen met meer alternatieven moeten het eerst gemaakt, daar bij raden door tijdnood de kans op goed gokken bij minder alternatieven groter is en wegstrepen van fout gedachte alternatieven vergroot ook de raadkans. Ook meet een meerkeuzetoets het gemak waarmee een student aan het twijfelen gebracht kan worden. Uit onderzoek blijkt dat taaltoetsen die dezelfde meetmethode hanteren (portfolio, open vragen, meerkeuzevragen enz.) maar verschillende aspecten van taalvaardigheid moeten meten, soms hoger onderling samenhangen dan toetsen die dezelfde vaardigheid beogen te meten met verschillende meetmethoden.
Een geavanceerde vorm van toetsen is het adaptief toetsen. Men laat toetsen maken op een pc en bij het beantwoorden van de vragen rekent een programma continu de betrouwbaarheid van de score uit. Ook worden vragen geboden uit een itembank waarin de karakteristieken van de items verwerkt zitten. Wordt een vraag goed beantwoord, dan volgt een moeilijker vraag, wordt de vraag fout beantwoord, dan volgt een makkelijker vraag. Deze vorm van adaptief toetsen veronderstelt dus wel toetsen op 1 dimensie in de gegeven populatie. Is een vooraf gespecificeerd betrouwbaarheidsniveau bereikt, dan breekt het programma af en wordt de kandidaat gefeliciteerd of sterkte gewenst.

Toetsdoelen

Toetsen hebben niet alleen verschillende vormen, ze dienen ook verschillende doelen. Zo kunnen we selectief dan wel diagnostisch toetsen. Toetsen we selectief, dan is niet alleen de betrouwbaarheid van de zak/slaagbeslissing van belang, maar ook de wijze waarop we die grens stellen. We onderscheiden norm referenced en criterion referenced toetsen. Bij de eerste vorm wordt het percentage gezakten constant gehouden, min of meer ongeacht de hoogte van de prestatie. Bij de tweede vorm, proberen we een minimaal noodzakelijk prestatieniveau vooraf te bepalen. Een voorbeeld hiervan is het rijexamen, waarbij een aantal handelingen zoals achteruit inparkeren, adequaat moet worden verricht. In de praktijk komen beide vormen ook gemengd voor. De centraal

schriftelijke examens die door het CITO worden gemaakt hebben een norm die niet door een extern criterium wordt bepaald, maar men probeert die norm of het prestatieniveau wel over jaren heen gelijk te houden. Blijken er dan meer dan aanvaardbare percentages leerlingen gezakt, dan stelt men de norm toch bij. Als men in het onderwijs als leerkracht wil proberen enigszins criterion referenced te toetsen, dan moet men op zoek naar een criterium. Men zou kunnen nagaan hoe goed afgestudeerde studenten die op een voldoende niveau hun vak beoefenen scoren op de gemeten trek, men zou ook relevante personen kunnen vragen welk niveau zij wensen, noodzakelijk achten of iets dergelijks. Een probleem dat daarbij optreedt, is dat dergelijke criteria vaak veel variatie vertonen. Op de verplichte rekentoets die afgenomen wordt bij studenten van de PABO heeft men als norm voor rekenen en hoogste niveau gekozen van leerlingen in groep 8 van het primair onderwijs. Duidelijk is dat ook deze norm aanvechtbaar is. Wat als er een of twee kleine Einsteins in groep 8 zitten? Er zijn heuristieken bedacht voor docenten die zelf een norm moeten stellen. Zo kan men vooraf een student in gedachten nemen die zich qua gemeten vaardigheid of kennis precies op het zak/slaagniveau bevindt. Vervolgens schat men dan bij elke vraag hoeveel kans deze student heeft om de vraag goed te maken. De som van de geschatte proporties is dan gelijk aan het aantal vragen dat een student minimaal goed moet beantwoorden om te slagen. In plaats van één student, kan men ook een groep van 100 van deze studenten in gedachten nemen, en schatten hoeveel van hen een vraag goed zullen maken. Als men deze aantallen optelt en door 100 deelt, komt men eveneens op het minimaal aantal goed te maken vragen uit. Een andere weg is meerdere collega's te vragen dit te doen en de uitkomsten te middelen. Een probleem bij deze beide methoden is dat we weten uit onderzoek dat naarmate iemand zelf hoger scoort op de gemeten vaardigheid of kennis, des te meer zal hij of zij de moeilijkheidsgraad van een vraag onderschatten. Overigens, of schoolse toetsen en examens goed kunnen voorspellen in hoeverre iemand in staat is vervolgonderwijs te volgen, kan worden betwijfeld. Het 'cadeau doen' van het middelbare schoolexamen, zoals in mei 1968 in Frankrijk gebeurde, bleek te resulteren in veel meer hoog opgeleide en afgestudeerde studneten in het hoger (tertiair) onderwijs dan bij de lichtingen uit 1967 en 1969. Zelfs de kinderen van de bofkonten van mei '68 blijken hoger opgeleid dan die van vergelijkbare ouders die in 1967 en 1969 hun baccalauréat niet cadeau kregen (Maurin & McNally, 2008).

Bij diagnostisch toetsen gaat het niet om vaststellen wie boven en wie onder de maat presteert, maar om het vaststellen aan welke kennis of vaardigheden een student moet werken en welke kennis of vaardigheden al beheerst worden. Bij diagnostisch toetsen is het onderscheiden op een bredere range belangrijker dan bij selectief toetsen. Het gaat immers niet om vaststellen of iemand voldoende presteert, maar om te bepalen op welke gebieden een student sterker of zwakker is om vandaaruit vast te stellen aan welke delen van het curriculum de student nu aandacht moet gaan besteden. De toetsuitslag moet dan ook bruikbaar zijn voor het bepalen van het benodigde vervolgonderwijs. Dit betekent dat diagnostische toetsen aparte scores moeten leveren voor aparte vaardigheids- of kennisaspecten. Een toets die meerdere vaardigheids- of kennisaspecten tegelijk meet en in één cijfer uitdrukt, is dus minder geschikt als diagnostische toets. Daarnaast is een diagnostische toets alleen zinvol, als de uitslag gekoppeld kan worden aan vervolgacties en dus aan onderwijs. Een uitslag die alleen aangeeft dat een vaardigheid niet erg goed is, maar waarbij vervolgens niet duidelijk is wat er gedaan moet worden om die uitslag beter te krijgen, is niet zinvol. Uitslagen op diagnostische toetsen moeten dus gekoppeld kunnen worden aan remediërend onderwijs. Daarnaast kunnen ook attituden of motivatie het onderwerp van diagnostisch toetsen zijn. Attituden kunnen immers leerdoelen zijn, denk bijvoorbeeld aan een positieve attitude ten aanzien van lezen in de vrije tijd, maar ook aan attituden ten aanzien van pesten of discriminatie, motivatie om schools te presteren en dergelijke.
Bij diagnostisch toetsen hoeven we overigens niet alleen te denken aan studenten. Diagnostische toetsen kunnen ook gebruikt worden om onderwijs te diagnosticeren. Dus om duidelijk te krijgen wat er wel of niet duidelijk is uitgelegd. Studenten aan de universiteit vullen vragenlijsten in om het onderwijs van hun docent te beoordelen. Deze oordelen worden bij functioneringsgesprekken van docenten gebruikt. In deze vragenlijsten worden allerlei verschillende aspecten van het onderwijs

bevraagd, zodat de scores zich lenen voor gerichte adviezen ter verbetering van het onderwijs van een docent. Overigens is de validiteit van deze studentoordelen aanvechtbaar, zoals ik in een oud onderzoek van mijzelf kon vaststellen (Van Schooten, Eiting & Bechger, 1993).
Momenteel worden ook scholen door de inspectie afgerekend op de prestaties van de gehele schoolpopulatie. De Cito-toetsen worden op leerlingniveau selectief, maar op schoolniveau diagnostisch gebruikt. Overigens gebruikt de inspectie daarbij een wijze van berekenen van schoolkwaliteit die zeer aanvechtbaar is (de 'corrected value added' benadering). Wat ik hiermee wil zeggen is dat ook bij het diagnosticeren van leerkrachten of scholen de validiteit van het gebruikte instrumentarium van belang is om tot de juiste conclusies te komen.

Tot slot; Bias ofwel vraagpartijdigheid
Eén belangrijk aspect van toetsen is in het bovenstaande nog niet genoemd; vraagpartijdigheid. Veel hoeven we hier niet over te zeggen. Leerkrachten moeten echter wel weten dat het bestaat. Vraagpartijdigheid is een kenmerk van een toetsvraag en betekent dat bepaalde leerlingen meer of minder kans hebben een vraag goed te maken, ongeacht de kennis of vaardigheid die de toets beoogt te meten. Ooit deed ik met collega's onderzoek naar toetsen voor het meten van woordkennis (Damhuis, De Glopper, & Van Schooten, 1989a en 1989b). De toets toonde een plaatje en daarnaast drie woorden. De leerling moest aangeven welk woord het best bij het plaatje past. Denk aan een plaatje van een beker met daarnaast de woorden 'beker', 'bakker' en 'boeker'. Nu gingen we uit van een taalkundige theorie. Talen kennen elk hun eigen betekenisonderscheidende klanken, ofwel fonemen. Zo geeft in het Portugees het al dan niet nasaal (zeg maar verkouden) uitspreken van een klinker een andere betekenis. Pau (niet nasaal) betekent 'stok' en Pãu (nasaal) betekent 'brood'. In het Nederlands kennen we het verkouden 'au' dus niet als foneem. Nu kennen we in het Nederlands het verschil tussen lange en korte klinkers (a en aa, o en oo etc.) en dat zijn bij ons fonemen. 'Bal' betekent namelijk niet hetzelfde als 'baal', en ook 'bos' en 'boos' betekenen niet hetzelfde. In het Marokkaans wordt het onderscheid tussen lange en korte klinkers echter niet als foneem gebruikt. De theorie was dan ook dat een leerling die thuis Marokkaans sprak, vaker fouten zou maken die de korte en lange klinkers betreffen. Dus onze hypothese was dat naast een plaatje van een man met een baard en de drie woorden 'bard', 'baard' en 'bord' Marokkaans sprekende kinderen vaker 'bard' als fout zouden kiezen dan Nederlandse leerlingen. Aan mij de eer om uit te zoeken of we deze bias konden vinden. Dat bleek echter totaal niet het geval te zijn. Wel vond ik echter veel items die bias vertoonden voor Marokkaanse leerlingen. Dit bleken alle items te zijn die over typisch Nederlandse zaken gingen (pannenkoeken, boten, molens, klompen, kaas etc.). Later werd in onderzoek gevonden dat een som als 'Jan heeft 2 appels, Kees heeft er 6, hoeveel appels hebben Jan en Kees samen?' bias vertoont in die zin dat Nederlandse kinderen daar hoger op scoren, ongeacht hun optelvaardigheid, dan kinderen met een niet Nederlandse taalachtergrond. Maken we van 'Jan' en 'Kees' 'Achmed' en 'Aisha', dan draait het effect om. Maken we van de appels geweren, dan doen jongens het weer beter en maken we er poppen van, dan zijn meisjes in het voordeel. Kortom, echt perfect gaan we toetsen nooit krijgen. Wel kunnen we proberen de inhoud zo neutraal mogelijk te houden om bias te verminderen.

Welke tips kunnen we nu geven voor het maken van schoolse toetsen?
- Definieer de te meten trek (vaardigheid/kennis etc.) zoals bedoeld (theoretische definitie).
- Maak een operationele definitie die zoveel mogelijk past bij de theoretische definitie (productief vs receptief) en besef waar de twee definities van elkaar afwijken (wat meten we niet dat er wel bij hoort en wat wel dat er niet bijhoort). Denk bijvoorbeeld aan het meewegen van spelfouten bij een stelopdracht.
- Bestrijk het gehele te meten domein.
- Laat een collega beoordelen of alle vragen het bedoelde aspect meten en niet iets anders.
- Probeer de variantie op vaardigheden die je niet wilt meten zo klein mogelijk te houden. Dus als je een rekentoets maakt, zorg dan dat de minst taalvaardige leerling alle opgaven goed begrijpt, anders

meet je bij deze leerling ook taalvaardigheid. Een rekentoets met veel taal erin zal bij laag taalvaardige leerlingen ook taal meten.
- De wijze waarop een toetsresultaat wordt bepaald, is mede bepalend voor de betrouwbaarheid en de validiteit van de scores. De betrouwbaarheid wordt groter naarmate men meer vragen opneemt, naarmate de oordelen over de juistheid door meer beoordelaars gegeven worden en naarmate men de oordelen meer standaardiseert (beoordelingsvoorschriften, voorbeeldprestaties of schalen).
- Gesloten toetsen meten kennis en vaardigheden receptief, elke methode kent eigen methode specifieke variatie. Operationaliseren op verschillende wijzen, doet methodespecifieke variatie afnemen (door uitmiddelen). Dit kan door meerdere vragen per aspect, of verschillende wijzen van toetsen.
- Bepaal het doel van de toets. Selectief, streef dan hoge betrouwbaarheid rond de zak/slaaggrens na, diagnostisch, zorg dat die kenmerken gemeten worden waar onderwijs in gegeven kan worden, zodat een didactisch handelingsplan is op te stellen op grond van de resultaten, kies dan ook vragen op alle vaardigheidsniveaus.
- Bepaal de zak/slaaggrens door bij alle vragen aan te geven hoeveel kans een net geslaagde leerling moet hebben om die vraag goed te beantwoorden en tel vervolgens de proporties op om tot de minimaal te halen score te komen die voldoende is.
- Meet onderscheiden aspecten apart, maak indien gewenst een beredeneerde weging van subscores tot één eindscore.
- Stel voldoende vragen, bestrijk het gehele domein.
- Geef duidelijke instructies.
- Maak helder aan studenten wat en hoe er getoetst gaat worden aan het begin van de cursus. Laat studenten weten hoe ze getoetst worden en wat ze moeten kennen en kunnen.
- Leg geen tijddruk op als presteren onder tijdsdruk geen onderdeel is van de theoretische definitie (dus vrijwel nooit, behalve bij bv technisch leesvaardigheid).
- Verifieer of studenten de vragen opvatten zoals bedoeld.
- Controleer na afname de dimensionaliteit (validiteit) en betrouwbaarheid van de toets.
– Het gebruik van een itembank biedt mogelijkheid tot perfectioneren, maar is wel fraudegevoelig door het bekend raken van vragen.
- Bij selectief toetsen (zak/slaagbeslissing) vooral vragen opnemen die qua moeilijkheid rond de zak/slaaggrens liggen.
- Stel nooit strikvragen.
- Starten met enkele makkelijke vragen voor het zelfvertrouwen en om testangst tegen te gaan.
- Gebruik zo mogelijk bij docentbeoordelingen meerdere onafhankelijk werkende beoordelaars per te beoordelen aspect.

Literatuur:

Damhuis, R., De Glopper, K. & Van Schooten, E. (1989a). Leesvaardigheid in het Nederlands van allochtone en Nederlandse leerlingen in groep drie van het basisonderwijs. *Pedagogische Studiën ,66 (*4), 158-171.
Damhuis, R., De Glopper, K. & Van Schooten, E. (1989b). Leesprestaties van allochtone leerlingen in groep drie en onderwijs-, leerling- en achtergrondkenmerken. *Pedagogische Studiën, 66*(6), 256-267.
Godshalk, F.I., Swineford, F., & Coffmann, W.E., (1966). *The measurement of writing ability*. New York: College entrance examination board.
Maurin, E. & McNally, S., (2008). Vive la Révolution! Long Term Educational Returns of 1968 to the Angry Students. *Journal of Labor Economics, 26*(1), 1-33.
Van Schooten, E., Eiting, M. & Bechger, T.M. (1993). Een onderzoek naar de betrouwbaarheid en validiteit van studentoordelen over de didactische vaardigheid van universitair docenten. *Tijdschrift voor Onderwijsresearch, 18*(1), 11-25.