Meten de maat genomen

Meten de maat genomen

Colofon

Ontwerp en print: Océ Business Services, Maastricht

ISBN: 978-90-5681-270-6
NUR: 870

Meten de maat genomen

Rede

Uitgesproken bij de aanvaarding van het ambt van bijzonder hoogleraar Medisch Onderwijs in het bijzonder op het gebied van Innovatieve Toetsvormen
aan de Faculty of Health, Medicine and Life Sciences van de Universiteit Maastricht

Oratie voor de leerstoel Onderzoek en Ontwikkeling op het gebied van innovatieve toetsvormen

Door
Lambert Schuwirth

Universiteit Maastricht

Geachte rector Magnificus, lieve familieleden en vrienden, beste collega's,

Over een aantal jaren gaat mijn dochter naar de basisschool. Ze zal dan enkele jaren lang, vele uren per dag, meerdere dagen per week geobserveerd worden door verschillende mensen die daar allemaal voor opgeleid zijn. Aan het einde van haar basisschoolcarrière zal een enorme rijkdom aan ervaringen en informatie verzameld zijn over haar kennen, kunnen en haar mogelijkheden. Dat is maar goed ook, want mijn dochter zal niet iedere dag even goed zijn; ze zal op sommige dagen beter presteren en op andere dagen minder goed. Zo zitten mensen nu eenmaal in elkaar. Op basis van al die informatie zou je in principe moeten kunnen voorspellen hoe ze het zal doen op de middelbare school. Ook daar zal ze haar betere en slechtere dagen hebben. Dat is ook niet zo relevant als ze maar uiteindelijk voldoende leert om op een gewenst eindniveau uit te komen.

Tot nog toe vind ik dit zelf best een logisch verhaal en ik hoop u ook.

Toch is er een kans dat het in de praktijk niet zo gaat. Het is alleszins mogelijk dat ze in haar laatste jaar op de basisschool apart getest zal worden en dat ze op basis daarvan een middelbareschooladvies zal krijgen. Als dat het geval is dan zullen alle, in de afgelopen jaren, opgedane ervaringen en informatie ondergeschikt gemaakt worden aan een vorm van hooggestructureerde toetsing, afgenomen op drie dagen in haar laatste jaar; de CITO-toets.

Dat is een gang van zaken waar ik dan moeite mee zal hebben.

Nu moet u wel weten dat ik van mening ben dat het CITO een hele goede organisatie is en dat de CITO-toets in zijn soort een heel goede toets is. Desondanks vind ik het principe vreemd. Hoe kan één enkele toets de informatie vervangen van zes jaar observaties gedurende het onderwijs?

Eerlijk gezegd ben ik van mening dat dat niet kan. Misschien kan die ene toets die informatie wel aanvullen maar zeker niet vervangen.

Dit probleem vormt de kern van deze oratie. In de komende 40 minuten zal ik beargumenteren waarom ik dit vind, wat ik denk dat de beperkingen precies zijn en in welke richting het onderzoek binnen de

leerstoel:"Onderzoek en Ontwikkeling op het gebied van innovatieve toetsvormen" dan zal gaan.

De klassieke psychometrische benadering van toetsing

Het onderzoek en de ontwikkeling van toetsing tot dusverre heeft sterk geleund op de klassieke psychometrie. Psychometrie is een wetenschappelijke, statistische benadering van de kwaliteit van metingen van psychische kenmerken, in dit geval dus de meting van competentie, en in mijn veld de meting van medische competentie. Dit vakgebied heeft een enorme bijdrage geleverd aan ons denken over de kwaliteit van toetsing. Begrippen zoals betrouwbaarheid en validiteit van meting zijn ingeburgerd geraakt, en gelden veelal als de twee belangrijkste kwaliteitscriteria voor toetsing.

Dit past ook heel goed bij de typische benadering van formele toetsing zoals we die allemaal goed kennen en aan den lijve ondervonden hebben, namelijk die van een meting van competentie met als allerbelangrijkste doel te bepalen of een student voldoende competentie heeft of niet; met andere woorden of de student geslaagd is of gezakt. Toetsing is dus geen onderdeel van het onderwijsproces, maar een externe meting van de uitkomst ervan.

Daar is helemaal niets mis mee als je een certificerende instantie bent, zoals bijvoorbeeld het CITO of, in ons veld, een National Board of Medical Examiners in de VS. Het doel van dergelijke organisaties is immers alleen om zo goed mogelijk onderscheid te maken tussen kandidaten; dat wil zeggen zo goed mogelijk de geslaagden van de gezakten te onderscheiden.

In een onderwijssituatie echter heeft deze benadering haar grenzen. Alvorens die met u te bespreken moet ik, denk ik, eerst de centrale begrippen 'betrouwbaarheid' en 'validiteit' kort toelichten.

Het begrip '*betrouwbaarheid*' geeft aan in hoeverre de scores die studenten op een bepaalde toets halen, representatief zijn voor de scores die dezelfde studenten zouden halen als ze een toets zouden hebben gemaakt die bestaat uit alle mogelijke relevante vragen over het vakgebied. De toets is dus de steekproef en alle mogelijke items uit het

vakgebied vormen de populatie. In ons vak wordt voor dit laatste ook wel de term universum gebruikt. Een toets is dus betrouwbaar als de scores op de toets sterk overeenkomen met de universumscores. Natuurlijk zijn de universumscores niet te bepalen, daarom wordt vaak een test-hertest-benadering gehanteerd. Een toets is dan betrouwbaar als de scores die studenten krijgen, overeenkomen met de scores die zij op een vergelijkbare toets (een zogenaamde parallelle toets) zouden hebben gehaald. Ook dit is in de praktijk bij examinering moeilijk te realiseren. Het zou namelijk inhouden dat je studenten na de eerste toets een tweede toets zou moeten voorleggen, maar leereffecten uit de eerste toets en vermoeidheid zullen de scores op de tweede toets ongewenst beïnvloeden. Een alternatief hiervoor is de toets willekeurig in tweeën te delen en de scores op beide subdelen met elkaar te vergelijken. Omdat dit echter ook nog te veel beïnvloed kan worden door toevalsovereenkomst, wordt de toets opgedeeld in even zoveel subtoetsen als dat er items zijn en worden dan de scores met elkaar vergeleken. Daarmee wordt dus in feite bepaald in hoeverre de items het met elkaar eens zijn over de competenties van de kandidaten. Dit wordt dan gerelateerd aan de mate waarin kandidaten van elkaar verschillen. Ik kom hier later nog op terug.

Validiteit gaat over de mate waarin een toets datgene meet wat hij bedoelt te meten. Stel dat ik een nieuwe toets ontwikkel die tot doel heeft medisch inzicht te toetsen, dan moet ik bepalen of datgene wat de toets meet ook daadwerkelijk medisch inzicht is, en niet bijvoorbeeld alleen feitenkennis. Validiteit is echter een behoorlijk ongrijpbaar iets, en er zijn vele verschillende vormen bekend. Ik wil voor dit verhaal twee grote stromingen onderscheiden: de zogenaamde 'indirecte' en 'directe' validiteitsvormen. De indirecte validiteitsvormen gaan uit van toetsscores. De centrale validiteitsvraag hierbij is altijd of de scores zich zo gedragen zoals ze verwacht worden te doen. Een eenvoudig voorbeeld is of de hoogte van de scores op een toets samenhangen met het niveau van expertise. In het geval van mijn nieuwe inzicht-toets is het dus de vraag of medici met enkele jaren ervaring hogere scores halen dan net afgestudeerde basisartsen. Dit ook weer onder de aanname dat de medici met enkele jaren ervaring meer medisch inzicht hebben dan net afgestudeerde basisartsen.

De directe vormen van validiteit daarentegen richten zich meer op een inhoudelijke beoordeling en verdeling van de items in een toets. Een

centrale vraag in deze benadering is bijvoorbeeld of experts het er over eens zijn dat een item inzicht toetst en dat het een relevant item is, en of de verdeling van de onderwerpen binnen de toets evenredig is aan de verdeling van onderwerpen binnen het vakgebied. Of de verdeling representatief is voor het universum. Waarschijnlijk klinkt het u tot dusverre allemaal nog steeds logisch in de oren, en zult u zich afvragen wat er dan anders moet.

Beperkingen van de klassieke psychometrie binnen het onderwijs

Wel, er is een belangrijk verschil tussen deze standaardbenadering en een benadering die binnen het onderwijs gewenst is. Zoals ik al zei, is het doel van de standaard toetsbenadering om zo goed mogelijk onderscheid te maken tussen geslaagden en niet-geslaagden. In het onderwijs is daarnaast echter andere toetsinformatie nodig.

Ik wil dit graag illustreren aan de hand van een analogie met de geneeskunde. De gestandaardiseerde toetsing komt sterk overeen met de benadering van screening in de geneeskunde. In screening wordt gebruik gemaakt van een gestandaardiseerde test, die afgenomen kan worden door iemand met minimale instructie, en die onafhankelijk van individuele patiënten ontwikkeld is. Het enige doel van screening is dan ook een dichotoom resultaat: bijvoorbeeld “ziek - niet ziek”. In de individuele patiëntenzorg vindt echter ook continu toetsing plaats maar op een andere wijze. Al bij het binnenkomen van de patiënt ‘toetst’ een arts wat hij ziet; iedere vraag tijdens de anamnese, iedere handeling tijdens het lichamelijk onderzoek, iedere vorm van aanvullende diagnostiek moet gezien worden als elementen uit een toetsing. Nu is het doel echter niet om alleen maar een dichotoom resultaat te krijgen op iedere toets, maar om al die verzamelde informatie te wegen en te combineren om zo goed mogelijk een geïndividualiseerd plan ten aanzien van verdere diagnostiek of therapeutisch handelen te bepalen. Het is dan ook van het grootste belang dat al die informatie in een status opgeslagen en gebruikt wordt. Screening hanteert een reductionistische informatiearme aanpak, klinische patiëntenzorg hanteert een holistische informatierijke aanpak. De ene vereist geen expertise van de professional de andere vergt juist veel expertise. Beide aanpakken completeren elkaar in een gezondheidszorgsysteem, geen kan de ander vervangen.

Zoiets geldt ook binnen het onderwijs. Als een student zijn studie wil optimaliseren zal hij betrouwbare en valide informatie moeten hebben over zijn kennen en kunnen of over zijn sterktes en zwaktes. Dit moet geïndividualiseerd zijn, want het zal voor de ene student een ander studie-advies of andere onderwijsinterventie inhouden dan voor de andere. Dit is te vergelijken met het therapeutisch plan in de gezondheidszorg.

Daarnaast zal in het toetsprogramma van de ene student meer of andere informatie over zijn competentie verzameld moeten worden dan over een andere. Dit is afhankelijk van de informatie die al beschikbaar is. Een student die goed lichamelijk onderzoek kan doen maar moeite heeft met de communicatie met patiënten zou eigenlijk in een vervolgtoets anders getoetst moeten worden dan een student die goed communiceert maar nog steken laat vallen in de uitvoering van het lichamelijk onderzoek. Dit komt overeen met het diagnostische handelen in de geneeskunde

Ik wil nog een keer bevestigen dat natuurlijk ook altijd bepaald moet worden of de student aan de maat is of naar verwachting aan het einde van de onderwijsperiode aan de maat zal zijn, net zoals in de geneeskunde prognose ook een belangrijk aspect is.

In onderwijs hebben we bij toetsing dus ook een informatierijke aanpak nodig om te bepalen welke therapeutische onderwijsinterventies nodig zijn en welke andere diagnostische toetsinformatie geïndiceerd is.

Hierbij heeft de klassieke psychometrie haar grenzen en hebben we behoefte aan een uitbreiding van theorie en praktijk

Zoals alle wetenschappelijke en statistische benaderingen gaat ook de psychometrie uit van een aantal basisaannames. De meest centrale hiervan wil ik vanuit de toetsing van medische competentie graag met u bespreken.

Wanneer betrouwbaarheid bepaald wordt aan de hand van interne consistentie, of zoals ik eerder uitlegde: of de items het eens zijn over de competentie van de kandidaat, dan zit hierachter de aanname dat het universum, ofwel de populatie, homogeen is. Stel dat ik een

bloedmonster van u heb en dat een paar maal goed heen en weer beweeg om het te homogeniseren, dan mag ik aannemen dat het hemoglobine goed homogeen verdeeld is. Als ik nu drie keer met een pipet wat bloed eruit neem en het hemoglobinegehalte bepaal en ik krijg drie verschillende resultaten (dus de 'items' zijn het niet eens met elkaar), dan heb ik gegronde redenen om aan de betrouwbaarheid van het meetinstrument te twijfelen.

Stel nu echter dat ik een BSE-kolom (bloedbezinking) heb met een plasmadeel, een witte-celdeel en een rode bloedceldeel; en als ik nu drie keer met een pipet materiaal opzuig uit die verschillende niveaus en het hemoglobine bepaal en ik krijg drie keer dezelfde uitkomst (de item's zijn het nu wel eens met elkaar) dan heb ik ook gegronde redenen om aan de betrouwbaarheid van het instrument te twijfelen.

Interne consistentie is alleen maar een goede maat voor betrouwbaarheid onder de aanname dat het universum ook homogeen is. Een zeer robuuste bevinding uit psychometrisch onderzoek bij toetsing is echter ook, dat er sprake is van zogenaamde domeinspecificiteit; hoe een student presteert op een casus is een zeer slechte voorspeller voor hoe hij het zal doen op een willekeurige andere casus. Psychometrisch gezien is er meer variantie in scores die niet verklaard kan worden door verschillen tussen studenten dan variantie die wel verklaard kan worden. Met andere woorden: er is meer ruis dan signaal in de meting.

Waarom is dit een probleem?

Dat wil ik graag weer illustreren met een voorbeeld. Stel u gaat naar een arts omdat u ziek bent. U bent echter helemaal niet tevreden met het consult en bespreekt dat ook met hem. Hij bevestigt dit en zegt dat hij ook vindt dat hij tijdens het consult onder de maat heeft gepresteerd, "maar", zo voegt hij eraan toe, "maakt u zich geen zorgen, want bij de volgende patiënt ga ik extra mijn best doen en veel beter presteren, dus gemiddeld zit het vandaag wel goed". U zult hier niet vrolijker van worden, want voor u is de variantie tussen de consulten wel degelijk relevant, in de psychometrische benadering van toetsing wordt het echter in principe als ruis gezien. Dit wordt ook vaak afgedaan als de weerbarstigheid van de werkelijkheid, maar het wekt de schijn op van: nu hebben we zo'n mooi psychometrisch model ontwikkeld en nu wil de

werkelijkheid zich er niet aan houden. Men zou echter ook het standpunt mogen innemen dat dit een zwakte is van de wetenschappelijke theorie om maar zo weinig van de geobserveerde variantie te kunnen verklaren. Nu moet ik eerlijkheidshalve zeggen dat in de psychometrie er ook wel andere theorieën en modellen zijn, zoals generaliseerbaarheidstheorie en item-responsetheroieën, maar deze verschillen op deze punten niet wezenlijk van de klassieke testtheorie.

Ik moet hierbij nog een nuancerende kanttekening maken, anders doe ik de psychometrie zwaar tekort. Diverse wetenschappen hebben te maken met een inductivistisch probleem: vanuit een beperkte set observaties moet een oneindig geldige uitspraak gedaan worden.

Als het gaat over zwaartekracht bijvoorbeeld dan geldt, dat hoeveel appels wij ook uit bomen omlaag zien vallen, we daaruit nooit mogen concluderen dat met zekerheid alle appels altijd omlaag zullen vallen. In de natuurwetenschappen wordt uit deze beperkte set observaties een wetmatigheid gedestilleerd die in kritische experimenten onderzocht wordt. Als een goed uitgevoerd kritisch experiment een repliceerbare tegengestelde bevinding oplevert (bijvoorbeeld dat appels omhoog vallen) zal de theorie herzien moeten worden. Ook binnen toetsing hebben we te maken met dit generalisatieprobleem, daar valt niet onderuit te komen. Binnen toetsing wordt ook een beperkte set observaties (bijvoorbeeld een toets) gedaan en moet een oneindige uitspraak gedaan worden: "In hoeverre beheerst de student het hele vakgebied". Ik wil hier dan ook absoluut niet de noodzaak tot generalisatie ontkennen, maar wel de eenzijdige wijze waarop deze generaliseerbaarheid binnen toetsing benaderd wordt.

We generaliseren namelijk door aan te nemen dat datgene wat we meten als een construct behandeld moet worden. Een construct is een psychische eigenschap, waarvan het bestaan weliswaar vermoed wordt, maar die niet direct waar te nemen is. Bekende voorbeelden zijn intelligentie of extraversie. We nemen allemaal aan dat er zoiets bestaat, maar intelligentie en extraversie als eigenschap zijn niet rechtstreeks waar te nemen, zoals bijvoorbeeld lichaamslengte dat wel is. Ze zijn natuurlijk wel af te leiden uit gedrag.

Van constructen wordt aangenomen, dat ze op het moment van meting stabiel zijn (dus niet variërend in tijd) en dat ze onafhankelijk van

elkaar zijn: iemand kan extravert en dom zijn of iemand kan extravert en slim zijn. Het bekendste voorbeeld van de constructbenadering in de definitie van wat medische competentie is, is die van de constructen: 'kennis', 'vaardigheden', 'probleemoplossend vermogen' en 'attitude'. Dit is echter een obsolete benadering.

In het onderwijs is "competenties" tegenwoordig een beetje het modewoord. En hoewel de competentiegerichte benadering een andere onderwijsbenadering is dan die die uitgaat van constructen, wordt het denken van toetsing nog steeds heel sterk door een constructbenadering beheerst. Nog steeds worden de competentiegebieden in het denken over toetsing min of meer als constructen behandeld.

De constructbenadering is niet per se fout. Ze is alleen erg beperkt, en ze kan niet zomaar op alle vormen van toetsing toegepast worden, want ze heeft nogal wat forse implicaties.

Een eerste implicatie is dat een reductionistisch perspectief gehanteerd moet worden. Zelfs wanneer ik een groep studenten een 100-item multiplechoicetoets voorleg, krijg ik antwoorden waaruit ik heel wat kan afleiden. Ik kan per student bepalen welke vragen hij goed had en welke hij fout had, maar ook welke foute antwoorden hij gegeven heeft. Van de groep studenten kan ik bepalen of er veelvoorkomende misconcepties bestaan. Dit wordt echter gereduceerd naar een goed-foutbeslissing per vraag. Nu weet ik alleen nog maar welke vragen goed en welke vragen fout beantwoord zijn, maar niet meer welke fouten zijn gemaakt. Vervolgens wordt dit omgezet naar een procentuele score. Nu weet ik alleen nog maar hoeveel vragen goed en fout waren, maar niet meer welke. Vervolgens wordt dit vergeleken met een zak-slaaggrens. Uiteindelijk weet ik dus alleen nog maar of er genoeg vragen goed beantwoord waren of niet. Veel van de toetsstatistiek en de discussies over zak-slaaggrenzen zijn erop gericht zo goed mogelijk informatie weg te gooien.

Daarom kunnen, en moeten vaak zelfs, individuele items ook als betekenisloos behandeld worden. Als ik een toets interne geneeskunde produceer dan kan dit item (figuur 1):

> Bij een vrouw van 72 jaar met angina pectoris wordt bij herhaling een bloeddruk gemeten van 170/100.
> Welk antihypertensivum heeft bij deze patiënte de voorkeur?
> a) captopril.
> b) chloorthalidon.
> c) metoprolol.

Figuur 1.

even goed vervangen worden door dit item (figuur 2):

> Meneer Jansen, 35 jaar, bezoekt zijn huisarts met de klacht van pijn op de borst. Zonder verdere gegevens van meneer Jansen is de kans het grootst dat de pijn op de borst voorkomt uit:
> a) de borstwand;
> b) de longen;
> c) het myocard;
> d) de slokdarm.

Figuur 2.

zonder dat iemand zal vinden dat het een groot gemis is als een van beide niet in de toets voorkomt. Van belang is alleen dat een item in de toets bijdraagt aan de meting van het construct, daarvoor hoeft het zelf niet intrinsiek betekenisvol te zijn. Dit is een notie die komt uit de testpsychologie. In een bekende persoonlijkheidsvragenlijst, de MMPI, staat bijvoorbeeld een vraag die luidt:

"Als ik er zeker van was onopgemerkt zonder te betalen in een bioscoop te kunnen komen, zou ik het waarschijnlijk doen".

Deze vraag heeft niet tot doel het bioscoopgedrag van mensen te onderzoeken, maar draagt bij aan de bepaling van een factor, in dit geval de L-schaal (leugenschaal). Ze had net zo goed vervangen kunnen worden door eenzelfde vraag naar het theater of zelfs naar wat de proefpersoon doet als hij teveel wisselgeld terugkrijgt in een winkel. De vraag is

inhoudelijk niet direct relevant, belangrijk is dat ze 'laadt' op een bepaalde factor. Hetzelfde zal dus gelden in een toets, wanneer het doel van de toets is om een construct te meten. Toch wordt dit lastig bij een medische vaardigheidstoets? Twee medische vaardigheden zijn het verrichten van een reanimatie of het beoordelen van een bloeduitstrijkje. Nu is het wat lastiger om eenvoudigweg te zeggen dat de 'items' in zichzelf betekenisloos zijn als ze maar laden op het construct vaardigheden. De klassieke psychometrie gaat hier echter van uit.

Dit inhoudelijkheidsprobleem komt nog nadrukkelijker naar voren in de paradox tussen inhoudsvaliditeit en betrouwbaarheid. Een voorwaarde voor een inhoudsvalide toets is dat een blauwdruk gemaakt wordt. Dit is een tabel waarin gespecificeerd staat hoeveel vragen over ieder onderwerp in een toets moeten zitten. Dit is van belang om te voorkomen dat erg eenzijdig naar één bepaald onderdeel gevraagd wordt of dat een bepaald onderdeel te weinig aan bod komt. Dit is in tegenspraak met de andere aannames. Betrouwbaarheid gaat uit van een homogeen universum, maar de noodzaak tot blauwdrukken impliceert een heterogeen universum. Los van wat waar is, moet gesteld worden dat of de ene aanname waar moet zijn of de andere, beide tegelijk kan niet. Bij veel psychologische testen is dit minder een probleem omdat ze veel meer gericht zijn op profielscores.

Binnen de toetspsychometrie is ook naar oplossingen gezocht, bijvoorbeeld door een aangepaste betrouwbaarheidsmaat, de zogenaamde gestratificeerde alpha. Hierin wordt rekening gehouden met verschillen tussen de homogeniteit binnen en tussen de blauwdrukcategorieën. Toch lost ook dit het probleem niet essentieel op, want binnen de blauwdrukcategorieën blijft de domeinspecificiteit bestaan en bovendien is het verschil in homogeniteit tussen en binnen categorieën niet zo bijster indrukwekkend. Een soort Babushkapop-oplossing dus.

Een andere implicatie die ik met u wil bespreken illustreert het beste het verschil in opvatting tussen certificerende toetsing en toetsing binnen onderwijs. Alweer aan de hand van de bepaling van de betrouwbaarheid. Kijkt u eens naar deze matrix die de scores weergeeft van drie studenten op een toets met drie items (tabel 1):

Tabel 1.

	item 1	item 2	item 3	score
student A	1	0,5	0	*1,5*
student B	1	0,5	0	*1,5*
student C	1	0,5	0	*1,5*
itemmoeilijkheid	1	*0,5*	0	

Hier ziet u een volledig onbetrouwbare toets. Alle vragen verschillen in hun oordeel over de competentie van de student. Item 1 denkt van iedere student dat hij heel goed is, item 2 vindt iedere student maar matig en item drie vindt iedere student slecht. De toets maakt verder geen onderscheid tussen studenten en is dus niet in staat de goede van de slechte te onderscheiden.

De volgende matrix (in tabel 2) geeft precies het omgekeerde weer.

Tabel 2.

	item 1	item 2	item 3	score
student A	1	1	1	*3*
student B	0,5	0,5	0,5	*1,5*
student C	0	0	0	*0*
itemmoeilijkheid	*0,5*	*0,5*	*0,5*	

De items zijn het bij alle studenten eens over de competentie, en er is maximaal verschil tussen de studenten: student A wordt door alle items goed gevonden, B matig en C slecht. Dit is een perfect betrouwbare toets.

Maar wat nu bij deze matrix in tabel 3:

Tabel 3.

	item 1	item 2	item 3	score
student A	1	1	1	*3*
student B	1	1	1	*3*
student C	1	1	1	*3*
itemmoeilijkheid	*1*	*1*	*1*	

Nu is de betrouwbaarheid niet meer te meten, er is namelijk geen verschil tussen studenten. Een slechte toets dus.

Maar stel nu eens, dat dit een cursus EHBO was, en dat de drie items luidden:
- demonstreer een reanimatie;
- demonstreer hoe je een drukverband aanlegt bij een slagaderlijke bloeding;
- demonstreer hoe je een wonddekverband aanlegt.

Zou het dan niet ideaal zijn als alle cursisten alle drie vaardigheden aan het einde van de cursus perfect beheersten? Toch, als dat het geval is, dus als het onderwijs perfect zou zijn geweest, kunnen we het volgens de klassieke testtheorie niet meer betrouwbaar en dus ook niet meer valide meten.

Hoe heeft de klassieke psychometrie ons denken beïnvloed?

Dit denken heeft een grote invloed gehad op de toetsontwikkeling tot nu toe.

De belangrijkste uitvloeisels daarvan zijn:
- we zijn eraan gewend geraakt dat een toets altijd tot doel heeft te bepalen of een student voldoende of onvoldoende is. Het toetsmoment is dus altijd ook beslismoment (bijvoorbeeld ten aanzien van al dan niet herkansen of het overdoen van een onderwijsdeel);
- voor ieder 'construct' is gezocht naar de ene beste toetsvorm. De stationstoets wordt bijvoorbeeld nu gezien als de gouden standaard voor medische vaardigheden, en de nieuwe ster aan het firmament, het portfolio, is *het* instrument voor het meten van reflectie;
- de aanname dat de toetsvorm in hoge mate bepaalt wat de toets meet en niet de toetsinhoud ('open vragen zijn beter dan gesloten vragen');
- de aanname dat alleen objectieve toetsing betrouwbaar kan zijn en subjectieve toetsing altijd onbetrouwbaar; en
- de aanname dat summatieve (toetsing om de voortgang van studenten te bepalen) en formatieve toetsing (met als doel feedback te geven aan de student) van elkaar gescheiden moeten zijn.

Onderzoek en heel veel ervaring met toetsen leiden echter tot héél andere conclusies:

- niet iedere toets moet per se een beslismoment zijn. Te veel beslismomenten leiden tot ongewenste studievertraging en studieuitval, en daarmee tot grote financiële, maatschappelijke en motivationele verliezen. Dit is te vergelijken met zogenaamde fout-positieve resultaten op diagnostiek, op basis van een afwijkende testuitslag wordt ten onrechte geconcludeerd dat de patiënt ziek is. Veel beter is het om toetsmomenten en beslismomenten uit elkaar te halen en het aantal beslismomenten per jaar te beperken;
- er is niet één panacee, niet één enkele superieure toetsvorm, alle toetsvormen hebben hun voor- en nadelen, hun indicaties en contra-indicaties. Een goed toetsprogramma zal dus (net als een therapeutisch spectrum) verschillende methoden bevatten die alle op het juiste moment en voor het juiste doel zijn gebruikt. Ook zal in dit programma de informatie uit verschillende toetsen en onderdelen van toetsen op een andere manier gecombineerd moeten worden. In plaats van onderdelen binnen een zelfde toetsmoment per definitie bij elkaar op te tellen zullen toetsdeelresultaten die gelijksoortige informatie opleveren met elkaar gecombineerd moeten worden, ook al komen ze van verschillende toetsen of toetsvormen;
- de vorm bepaalt nauwelijks wat er getoetst wordt, de inhoud des te meer. Toen ik trouwde (met de mooiste vrouw van de hele wereld) had de ambtenaar van de burgerlijke stand mij kunnen vragen:"Wat is de kleur van de jurk van de persoon rechts naast u?". Dat was een open vraag geweest, maar op dat moment een triviale beslissing. Mij werd daarentegen gevraagd:"Wilt u deze vrouw tot uw wettige echtgenote nemen?". Dit was een twee-optie multiple-choice , maar ik verzeker u dat het geen triviale beslissing was. (Dit voorbeeld dank ik overigens aan mijn gewaarde collega Van Berkel). Dit wordt gestaafd door de bevinding, dat variaties binnen toetsen groter zijn dan tussen toetsen. Op een stationstoets is het verband tussen de scores op een station 'buikonderzoek' en 'neurologisch onderzoek' veel kleiner dan het verband tussen een station 'buikonderzoek' en vragen in een schriftelijke toets over buikanatomie en -aandoeningen. Toch sommeren we het 'buikonderzoek' met het 'neurologisch onderzoek' omdat het dezelfde toets(vorm) betreft;
- betrouwbaarheid is niet zozeer een gevolg van objectiviteit als wel van steekproefneming. Een groot aantal subjectieve oordelen zijn in principe betrouwbaarder dan een enkele objectieve meting. Stel ik

schrijf 10 stukjes muziek en ik neem tien stukjes muziek van Ludwig van Beethoven, en ik leg beide sets voor aan een panel van experts. Ik vraag hun dan te beoordelen welke muziek een grotere innerlijke kracht heeft, wie de grotere kunstenaar is, welke muziek een interessantere harmonie heeft, wie het ambacht van componist beheerst, kortom allerlei subjectieve zaken. Hoogstwaarschijnlijk zullen de experts tot het oordeel komen dat Beethoven het wat beter doet dan ik, in alle opzichten. Ook als ik al mijn eigen stukjes pak, en alle werken van Beethoven of alle experts in de wereld vraag, het oordeel zal niet veranderen. Het is dus perfect betrouwbaar. Maar het oordeel is en blijft subjectief. Het omgekeerde geldt voor een multiple-choice toets over interne geneeskunde die bestaat uit slechts één vraag. Dit is een zogenaamde objectieve toets maar een veel te kleine steekproef om betrouwbaar te zijn.

- Het onderscheid tussen summatief en formatief is niet erg zinvol. Een puur summatieve toets zou de student geen informatie verschaffen, alleen laten weten of hij geslaagd of gezakt is. Een puur formatieve toets zou veel informatie geven maar te weinig gewicht in de schaal leggen om serieus genomen te worden. Als het in onderwijs belangrijk is dat toetsing studenten helpt om beter te worden, dan zal een student precies te weten moeten komen wat hij dan moet doen om beter te worden, ***en*** die informatie moet gewicht in de schaal leggen. Als een tennisser een wedstrijd met 3 x 6-0 verliest, weet hij heel zeker dat hij op dat moment slechter was dan zijn tegenstander. Dat is echter vrij nutteloze kennis voor als hij een volgende keer wel wil winnen. Daarvoor heeft hij informatie nodig over wat er *goed* ging en wat er *fout* ging; hij heeft informatie nodig over hoe de fouten te remediëren en hij moet de mogelijkheid hebben om die fouten te remediëren. Alleen maar telkens te horen dat hij niet goed genoeg is maakt hem geen toptennisser.

Nu wil ik even met u teruggaan naar het begin van deze rede. De CITO-toets waar ik over sprak is een eenmalige meting (weliswaar verdeeld over drie dagen) die alleen maar aangeeft of mijn dochter voor een bepaald niveau geschikt is of niet. Het toetsmoment is in feite tegelijk beslismoment, het is bedoeld als de ene superieure toets die alle aspecten van mijn dochter's competentie zou moeten bepalen, er wordt gekapitaliseerd op één zogenaamd objectieve toetsvorm met name vanwege de logistiek en de interne consistentie van de toets, de toets heeft geen formatieve waarde (mijn dochter zal er niet uit leren waar ze

in het vervolgonderwijs op moet letten en wat haar sterktes en zwaktes zijn), het is een momentopname met dus slechte steekproefkwaliteiten ten aanzien van de tijdsperiode (heeft ze nu een betere dag of een slechtere dag?), niettemin zal de interne consistentie, de schatting van de betrouwbaarheid, hoog zijn..

Dit zijn de achterliggende argumenten van mijn probleem met een dergelijke gang van zaken rond de CITO-toetsing.

Inhoud van de leerstoel

Dat brengt mij tenslotte bij de inhoud van de leerstoel Onderzoek en Ontwikkeling op het gebied van innovatieve toetsvormen. Hierin zullen twee deelgebieden onderzocht worden die ik nog kort voor u wil ontvouwen.

Deelgebied 1: Programmatische aanpak van toetsing.

Als het zo is dat niet één enkel toetsinstrument voldoende is en dat een variatie aan methoden op de juiste wijze gecombineerd zal moeten worden, zullen we meer te weten moeten komen over hoe zo'n toetsprogramma eruit moet zien. Zoals een toets meer moet zijn dan een losse verzameling items, zal een toetsprogramma meer moeten zijn dan een losse verzameling toetsen. Verbazingwekkend genoeg is er op dit terrein praktisch geen onderzoek en slechts mondjesmaat literatuur te vinden.

Dit deel van het onderzoeksprogramma zal daarom antwoord moeten geven op de vraag wat dan hoog-kwalitatieve toetsprogramma's zijn en wat niet. Daarvoor zijn we bezig een model te ontwikkelen, dat op verschillende wijzen getoetst zal worden. Binnen dit model definiëren we criteria voor kwaliteit van toetsprogramma's. In een eerste stap zullen model en criteria getoetst worden met behulp van afstemmings- en consensusmeetings met panels van internationale en nationale experts. Dit zal eventueel aangevuld worden met een WIKI-gebaseerde constructie van aanvullende elementen en criteria. Vervolgens gaan we het model toetsen aan de hand van een bestaand toetsprogramma dat onafhankelijk van het model opgezet is, en daarna aan de hand van een toetsprogramma dat met inachtneming van de criteria ontwikkeld

is. Dit zal ons uiteindelijk inzicht moeten verschaffen in welke criteria zinvol en bruikbaar zijn voor de doelstellingen van een toetsprogramma en de wijze waarop in zo'n programma informatie over de studenten verzameld moet worden. We zullen meer moeten weten over de wijze waarop deze informatie gecombineerd moet worden, de manier waarop beslissingen op basis van de resultaten genomen kunnen worden en hoe deze teruggekoppeld gaan worden aan studenten, docenten, onderwijsorganisatie en maatschappij ter verbetering van de kwaliteit. Dit alles zal niet in een vacuüm gebeuren, er zal sterk gekeken worden naar andere wetenschappelijke en maatschappelijke domeinen om daaruit lessen te trekken en te voorkomen dat we het wiel opnieuw uitvinden.

Deelgebied 2: Aanpassing van de wetenschappelijke en statistische modellen.

Dat laatste zal in nog sterkere mate gelden voor het tweede deelgebied. Hierin zullen vier aspecten centraal staan:

- De vraag welke mogelijkheden er bestaan voor de optimalisatie van de informatie uit een variëteit aan informatiebronnen, zowel kwantitatief als kwalitatief. Hoe gebeurt dit in andere wetenschappelijke en maatschappelijke gebieden? Wat kunnen we leren van de besliskunde, actuariële methoden, beoordelingen, de business literatuur, de rechtspraak, etc over benaderingen van betrouwbaarheid, geldigheid, verdedigbaarheid en eerlijkheid van informatierijke beslissingen? Het centrale vraagstuk is welke generieke elementen hieruit te halen zijn, en welke we kunnen decontextualiseren en vervolgens gebruiken binnen toetsing.
- Hoe gaan mensen te werk bij het nemen van beslissingen, wat zijn precies hun feilbaarheden? Welke interne cognitieve factoren zijn hierop van invloed, welke procesfactoren en welke organisationele factoren beïnvloeden de uitkomst? Welke elementen hieruit zijn toepasbaar op toetsing?
- Hoe kunnen de twee voorgaande punten bij elkaar gebracht worden; hoe kunnen oordelen en optimalisatie van informatieprocessen geïntegreerd worden om zo goed mogelijk uit een veelheid van verschillende bronnen tot gedifferentieerde beslissingen te komen? Wat is nodig voor een diagnostische beslissing (wat gaat goed wat gaat niet goed), wat voor een therapeutische beslissing (wat moet in het onderwijs meer aandacht krijgen en wat minder) en wat is

nodig voor een prognostische beslissing (hoe goed is de student op dit moment en naar verwachting op het einde)?
- Welke rol kunnen algoritmes, heuristieken c.q. decision support systemen hierin spelen?

Ik vind deze twee onderwerpen een uitermate boeiend terrein en ik ben heel benieuwd naar de antwoorden die ons onderzoek zal opleveren. Ik hoop dat ik iets van mijn enthousiasme voor dit vak heb kunnen overbrengen.

Rest mij woorden van dank uit te spreken. Dat wil ik niet lichtvaardig of plichtmatig doen, maar ik verzeker u dat ze recht uit mijn hart gegrepen zijn.

Moeder, vader, ik hoop (en weet ook wel) dat jullie hier trots zitten te wezen. Heerlijk, maar ben vooral trots op jezelf. Je hebt, ook in moeilijke tijden, je kinderen de mogelijkheid gegeven zich zo te ontwikkelen als ze wilden. Steun en ondersteuning, motivering en medeleven hebben jullie me gegeven als het nodig was. Zonder jullie had ik nooit hier gestaan. Natuurlijk had ik zonder jullie nergens gestaan, maar zeker niet hier. Dank je wel voor alles.

Het zelfde geldt voor Francien, mijn zus, en Gerard en Wim, mijn broers.

Bedanken wil ik ook mijn andere moeder en vader, mijn schoonouders. Jullie hebben net zo zeer meegeleefd met me, me van hulp en suggesties voorzien alsof ik jullie eigen zoon was. Het doet me dan ook het grootste plezier te zien dat jullie erbij zijn en te zien dat jullie ervan genieten.

Candida, Marko en Johannes, wat fijn dat jullie er vandaag bij zijn. Nu moet je in een van onze volgende discussies alles direct geloven wat ik zeg, want ik draag nu een mooie toga, nietwaar?

Professor dr. Scherpbier en Professor dr. Hillen, Albert en Harry, wil ik graag hier noemen. Beiden hebben jullie telkens tijd en aandacht genomen me te ondersteunen als ik aan mezelf twijfelde en bij het mij op weg te helpen bij het begin van deze moeilijke klus. Heel veel dank hiervoor.

Dat brengt me bij Onderwijsontwikkling en Onderwijsresearch, de capaciteitsgroep waar ik sinds 1991 werkzaam ben. Bij alle collega's kan ik altijd binnen lopen als ik een vraag of verzoek heb; de deur is nooit dicht en samenwerken is altijd de norm. Dank jullie wel voor jullie collegialiteit.

Eén iemand wil ik echter heel speciaal noemen, met name omdat hij van onschatbare waarde is geweest voor mijn academische vorming. Cees van der Vleuten, baas, mentor, vriend. Heel veel heb ik van je geleerd, van je mogen leren. Inhoudelijk en als mens. Een van de eigenschappen die ik het liefst van je afkijk is je vermogen anderen in hun waarde te laten, anderen te laten groeien, anderen zelfs te stimuleren boven jou uit te groeien. Een zeldzame en mooie eigenschap, die ik hoop van je over te kunnen nemen Ik besef echter wel terdege dat je mij ook op dit terrein geval toch altijd een stap voor blijft. Maar dat is ook goed zo.

Iris, tot voor twee jaar geleden centrum in mijn universum, en nu een van de twee centra samen met Charlotte en binnenkort een van de drie centra in mijn leven; jouw rol is zo essentieel dat ik die hier bijna niet kort kan beschrijven, maar ik probeer het toch. Luisteren naar mijn verhalen, die soms veel te technisch waren (leuke hé, generaliseerbaarheidstheorie), me oppeppen in moeilijke tijden, de zaak thuis opvangen als ik weer eens voor het werk in het buitenland was, me helpen bij de moeilijke balans tussen thuisverplichtingen en werkverplichtingen, en ga zo maar door. Onschatbaar ben je en toch een schat.

Allen die ik hier nu niet kan noemen, maar die ik wel zou moeten nemen, weten hopelijk dat ik hun dankbaar ben en hen mee wil laten delen in het plezier en de feestelijkheid van vandaag.

Ik heb gezegd.