Psychometrie: Onderzocht, aangeleerd en toegepast

Psychometrie: Onderzocht, aangeleerd en toegepast

Rede

uitgesproken bij de aanvaarding van het ambt van
bijzonder hoogleraar Kwantitatieve onderzoeksmethodologie
ter bevordering van de academisering van het onderwijs,
de Kohnstammleerstoel
aan de Faculteit der Maatschappij- en Gedragswetenschappen
van de Universiteit van Amsterdam
op 16 januari 2016

door

Andries van der Ark

Dit is oratie 556, verschenen in de oratiereeks van de Universiteit van Amsterdam.

Opmaak: JAPES, Amsterdam
Foto auteur: Dirk Gillissen

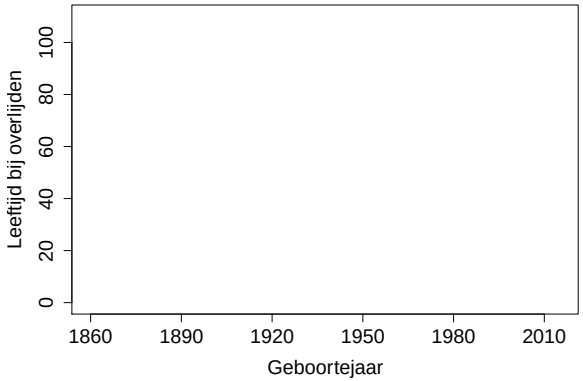
Mevrouw de Rector Magnificus,
Mijnheer de Decaan,
Leden van het Bestuur van de Vereniging ter Bevordering van de Studie der Pedagogiek,
Leden van het curatorium van de Kohnstamm leerstoel,
Zeer gewaardeerde toehoorders,

Met deze oratie geef ik aan mijn benoeming tot bijzonder hoogleraar Kwantitatieve onderzoeksmethodologie ter bevordering van de academisering van het onderwijs (Kohnstamm-leerstoel) te aanvaarden. In deze oratie geef ik mijn visie op het onderzoeken, onderwijzen en toepassen van de psychometrie en de kwantitatieve onderzoeksmethodologie.

Prelude

Maar eerst wil ik u iets vertellen over een beroemd onderzoek van Wainer, Palmer en Bradlow.[1] De onderzoekers gingen naar het kerkhof van Princeton in New Jersey en noteerden bij alle 204 grafstenen de geboortedatum en sterfdatum van de overledene. Vervolgens zetten zij de geboortedatum en de bereikte leeftijd bij overlijden tegen elkaar af in een grafiek om te zien of er een trend is in hoe oud mensen worden. Ik wilde dit onderzoek graag in Amsterdam repliceren en het leek mij een aardig idee als leerkrachten van lagere scholen in het kader van een project over leven en dood, met de leerlingen naar een kerkhof zouden gaan om de geboorte- en sterfdata van de grafstenen te noteren. Ik kan u zeggen dat ik van een koude kermis thuis kwam. Het curriculum in het primair onderwijs is overvol en leerlingen hebben helemaal geen tijd om naast het vastgestelde leerprogramma iets te ondernemen. De Amsterdamse gemeentelijk instelling Onderzoek, Informatie en Statistiek bood hulp.[2] Zij houden sinds 1975 de geboorte- en sterfdata bij van alle Amsterdammers. Op verzoek leverden zij een bestand met de geboorte- en sterfdata van alle Amsterdammers die sinds 1975 zijn overleden.[3] Een enorm virtueel kerkhof. In figuur 1 staat geboortejaar op de horizontale as en de leeftijd bij overlijden op de verticale as. Ik kom later terug op dit onderzoek, maar ik

vraag u nu alvast te bedenken welke relatie u verwacht tussen geboortejaar en gemiddelde leeftijd bij overlijden. Het kan zijn dat u verwacht dat de gemiddelde leeftijd toeneemt, of afneemt, of misschien wel varieert.



Figuur 1: Figuur waarin gemiddelde leeftijd bij overlijden als functie van geboortejaar weergegeven kan worden

Inleiding

Terug naar de Psychometrie. Psychometrie is de wetenschap die zich bezighoudt met de technieken voor het objectief meten van psychologische fenomenen bij een persoon. Het kan hier gaan om kennis (bijvoorbeeld iemands kennis van de Engelse taal), vaardigheden (bijvoorbeeld iemands rekenvaardigheid), attituden (bijvoorbeeld iemands houding ten opzichte van de doodstraf), en persoonskenmerken (bijvoorbeeld iemands mate van extraversie).[4] Het meten van psychologische fenomenen gaat anders dan het meten van meer alledaagse dingen zoals lengte of gewicht. Als een persoon op een betrouwbare en valide weegschaal gaat staan dan kan het gewicht van de persoon direct op de weegschaal afgelezen worden. De lengte van de persoon kan met een deugdelijke rolmaat of meetlat ook direct afgelezen worden. Bij het meten van psychologische fenomenen zijn dergelijk instrumenten niet beschikbaar.

Ik neem als voorbeeld hoofdrekenen. Als ik wil weten wat iemands niveau van hoofdrekenen is kan hem of haar een hoofdrekenopgave voorleggen en vragen om het antwoord te geven zonder verdere hulpmiddelen: 3 – 1 = ? De

persoon krijgt een score 1 als het antwoord goed is en een score 0 als het antwoord niet goed is. Het lijkt aannemelijk dat personen met een score 1 een hogere mate van rekenvaardigheid hebben dan personen met een score 0. Ik ga ervan uit dat iedereen in de zaal een score 1 zou hebben behaald. De meting van hoofdrekenvaardigheid is met slechts één opgave zeer onnauwkeurig. Er wordt slechts onderscheid gemaakt tussen personen die zelfs één van de meest eenvoudige hoofdrekenopgaven niet kunnen maken en alle anderen. Om die reden krijgen personen niet één opgave voorgelegd maar een flink aantal. Een tweede opgave zou bijvoorbeeld kunnen zijn $^{10}\log 25 + {}^{10}\log 4 = ?$ Ook hier krijgt de persoon een score 1 als het antwoord goed is en een score 0 als het antwoord niet goed is. Ik denk dat ik er veilig vanuit kan gaan dat niet iedereen in de zaal een score 1 zou hebben behaald. Overigens is het antwoord bij beide opgaven 2.

Bij twee opgaven kunnen vier scorepatronen onderscheiden worden: (1, 1) beide opgaven goed, (1, 0) alleen de eerste opgave goed, (0, 1) alleen de tweede opgave goed en (0, 0) beide opgaven fout. Het scorepatroon (0, 1) zal zeer weinig voorkomen, maar de mogelijkheid bestaat dat een briljant hoofdrekenaar eventjes afgeleid werd omdat er net iemand binnenkomt en daarom per ongeluk het verkeerde antwoord invult. Naarmate we meer hoofdrekenopgaven toevoegen kunnen we in principe steeds nauwkeuriger meten. Bij een hoofdrekentest met 40 opgaven, wat niet ongewoon is, zijn er ruim een biljoen scorepatronen. Dat zijn ruim 150 keer zoveel scorepatronen als dat er momenteel mensen op aarde zijn.

Beoefenaars van de psychometrie, die psychometrici genoemd worden maken wiskundige modellen om van deze enorme hoeveelheden scorepatronen zinvolle meetwaarden te maken. De meest voorkomende manier om met die grote aantallen scorepatronen om te gaan is de scores op te tellen tot een somscore. De somscore kan een zinvolle maat zijn voor het psychologisch fenomeen dat men wilde meten, maar men moet bedenken dat bij het hoofdrekenvoorbeeld het onderscheid verdwijnt tussen personen die alleen de opgave '3-1 = ?' goed hadden en personen die alleen de opgave '$^{10}\log 25 + {}^{10}\log 4 = ?$'goed hadden.[5] Ook is het optellen van de scores mogelijk niet zinvol als er opgaven tussen zitten die geheel of gedeeltelijk andere vaardigheden vereisen. Stel dat we als derde opgave toevoegen: 'Wat is de hoofdstad van Vlaanderen?'[6] Kennis over deze opgave heeft niets met hoofdrekenen te maken en de score op deze opgave zou niet moeten meetellen bij het bepalen van de meetwaarde voor hoofdrekenen. Het is echter niet altijd direct duidelijk of het antwoord op een opgave iets zegt over wat met wil meten. Bijvoorbeeld de opgave 'Hans heeft een nieuwe externe harde schijf gekocht met 5 terabyte geheugen. Hoeveel video's van 5 gigabyte kan hij kwijt op de nieuwe externe

harde schijf?'[7] Deze opgave heeft te maken met hoofdrekenen, maar ook met kennis van computers, en ook met taal. Of deze opgave in de hoofdrekentest opgenomen moet worden is weer een opgave voor psychometrici.

Het bedenken van methoden en technieken om dergelijke opgaven te detecteren zodat ze verwijderd kunnen worden, of te accommoderen zodat ze geen of weinig invloed hebben op de meetwaarde van hoofdrekenen, behoort ook tot de psychometrie. Evenals het ontwikkelen van methoden en technieken om het aantal dimensies in de testdata vast te stellen, om de nauwkeurigheid van de meetwaarden vast te stellen, om meetwaarden van verschillende groepen te vergelijken, om met ontbrekende scores om te gaan, enzovoort. Ik ga verder niet in op al deze methoden en technieken. Ik wil er slechts mee aangeven dat de psychometrie een breed vakgebied is en steeds breder wordt. Door samenwerking met en beïnvloeding door andere vakgebieden zijn de grenzen tussen psychometrie, biometrie, econometrie en statistiek grotendeels vervaagd. Psychometrie is ook al lang niet meer het exclusieve domein van de psychologie; In alle vakgebieden waar mensen het onderwerp van onderzoek zijn – de pedagogiek, onderwijskunde, sociologie, geneeskunde, economie – speelt de psychometrie een belangrijke rol.[8]

Onderzocht

Ik durf te stellen dat de psychometrie van immens belang is geweest voor de maatschappij. In 2005 heb ik samen met collega Patrick Groenen twaalf presidenten van de internationale vereniging voor psychometrie, de Psychometric Society, gevraagd wat de belangrijkste bijdragen van de psychometrie zijn geweest.[9] Hun antwoorden hebben we verdeeld in drie klassen:

Ten eerste heeft psychometrisch onderzoek het idee geïntroduceerd dat het mogelijk is om kennis, vaardigheden, attituden en persoonlijkheidskenmerken te meten. Dit idee heeft een geweldige impact gehad op onze samenleving. Gestandaardiseerde tests zoals de Cito-toets, of persoonlijkheidstests bij pedagogische diagnostiek zijn niet meer weg te denken uit onze maatschappij. Ik heb ook bezwaren gehoord, vooral tegen de grote hoeveelheid tests en het rigide gebruik van tests.[10] Hoewel ik de bezwaren onderken, ga ik in deze rede toch een lans breken voor de psychometrie.

Psychometrisch onderzoek heeft eveneens de methoden en technieken voortgebracht die nodig zijn om een test of vragenlijst te construeren en om de test- en vragenlijstdata te analyseren. Als het mogelijk is om psychologische fenomenen te meten, dan moet er uiteraard ook gereedschap zijn om meetinstrumenten te maken. Voorbeelden van dergelijke methoden zijn

item-responsmodellen, structurele vergelijkingsmodellen, schalingsmodellen en modellen uit de klassieke testtheorie.

Ten slotte heeft psychometrie ideeën aangedragen hoe wij naar data moeten kijken: Wellicht het belangrijkste idee is dat psychologische fenomenen geformaliseerd en gekwantificeerd moeten zijn om te kunnen worden onderzocht. Een ander idee dat uit de psychometrie stamt, is het organiseren van data in een datamatrix met in de rijen de observatie-eenheden en de variabelen in de kolommen.

In mijn eigen onderzoek in de psychometrie is voorzichtigheid een belangrijke drijfveer. Ik houd ervan psychometrische modellen te ontwikkelen of te onderzoeken die zo min mogelijk assumpties veronderstellen. Mijn wetenschappelijke angstdroom wordt beschreven in de strip Dirkjan, waar de hoofdpersoon Dirkjan geheel toevallig op iets interessants stuit, data verzamelt, de data exploreert, een zeer omvangrijk model in elkaar kleit, de resultaten presenteert en daar uiteindelijk veel lof voor oogst.[11] De crux zit uiteraard in het model dat Dirkjan heeft ontwikkeld. Ten behoeve van het model maakt hij allerlei assumpties die niet onderzoekbaar zijn: Hij gaat er bijvoorbeeld vanuit dat er een wezen bij het bot hoort dat grote voeten en een staart heeft, maar hij heeft geen manier om er achter te komen of deze assumpties realistisch zijn. De mathematisch psycholoog Clyde Coombs schreef in zijn boek A theory of data: 'Knowledge is the result of theory – we buy information with assumptions.'[12] Dirkjan gaat hier uiteraard extreem ver in, maar in de psychometrie zijn veel methoden en technieken die alleen geldig zijn onder assumpties die nauwelijks realistisch zijn en mogelijkerwijs een grote invloed hebben op het resultaat van het onderzoek. Ik heb daarom altijd een voorkeur gehad voor methoden en technieken die uitgaan van weinig assumpties. Ik noem twee voorbeelden:

Ten eerste non-parametrische item-responsmodellen.[13] Dit is een verzameling wiskundige modellen voor de scorepatronen die het resultaat zijn van een testafname bij een grote groep personen. De non-parametrische item-responsmodellen worden vaak afgezet tegen de parametrische item-responsmodellen die meer assumpties hebben. De non-parametrische item-responsmodellen hebben dus relatief weinig assumpties. Ze passen daardoor relatief goed op de testdata, maar leveren relatief minder informatie over de meetwaarde. Enigszins kort door de bocht gezegd kan men met de uit deze non-parametrische item-responsmodellen afgeleide methoden en technieken, die bekend staan onder de naam Mokken schaalanalyse, onderzoeken of de somscore een zinvolle maat is als meetwaarde voor een test.[14] En zo niet, welke opgaven verwijderd moeten worden. Op basis van Mokken schaalanalyse zou men bijvoorbeeld een kunnen bepalen of de opgave 'hoeveel video's van 5 gigabyte

men kwijt kan op een externe harde schijf van 1 terabyte' wel of niet uit de hoofdrekentest verwijderd moet worden? Bij vrijwel de meeste tests en vragenlijsten wordt de somscore gebruikt als meetwaarde. Non-parametrische item-responsemodellen zijn daarom praktisch zeer relevant, en ik ben bezig met het promoten van deze modellen. Enerzijds door deze modellen in gebruikersvriendelijke software onder te brengen[15] en anderzijds door ze te generaliseren naar multilevel data.[16]

Het tweede voorbeeld betreft de afleiding van standaardfouten voor veelgebruikte coëfficiënten in de psychometrie. Een voorbeeld van zo'n coëfficiënt is de percentielscore. Een percentielscore geeft aan welk percentage van de personen een lagere score heeft dan de persoon die gemeten wordt. De percentielscore wordt vaak gebruikt bij het normeren van tests en geeft de meetwaarde betekenis. Als ik een score 37 op een rekentest heb, dan weet ik nog niets. Als ik weet dat 80 procent van de personen een lagere score dan 37 heeft behaald dan is mijn percentielscore 80 en weet ik in ieder geval hoe ik gescoord heb ten opzichte van de overige deelnemers. Deze percentielscore hang echter af van de andere deelnemers. Als de overige deelnemers over het algemeen wat beter hadden kunnen rekenen was mijn percentielscore wellicht 77 of 78 geweest, als de overige deelnemers het over het algemeen wat minder goed hadden kunnen rekenen was mijn percentielscore wellicht 83 of 84 geweest. Percentielscores zijn onderhevig aan steekproeffluctuatie. Opmerkelijk genoeg is dat besef nog niet doorgedrongen tot de praktijk van de testconstructie, want de literatuur over normeren schenkt hier geen aandacht aan: Niet bij percentielscores, maar ook niet bij andere coëfficiënten die bij normeren gebruikt worden zoals percentages, standaardscores, standaarddeviaties en stanines. De onzekerheid door steekproeffluctuatie kan men kwantificeren met zogenaamde standaardfouten. Als er al standaardfouten voor coëfficiënten worden afgeleid dan wordt dat gedaan onder vrij strenge verdelingsassumpties. Er wordt bijvoorbeeld aangenomen dat de testscores normaal verdeeld zijn. Dit is niet altijd realistisch; testscores op veel klinische tests zijn scheef naar rechts verdeeld. Hannah Oosterhuis, een promovenda die ik mede begeleid, heeft echter laten zien dat met minimale verdelingsassumpties vrijwel zuivere standaardfouten afgeleid kunnen worden voor coëfficiënten die bij normeren gebruikt worden.[17] Deze standaardfouten kunnen zowel in het geval van discrete als in het geval van continue testscores gebruikt worden, ongeacht de verdeling van de testscores.

Beide voorbeelden noem ik om mijn hang naar voorzichtigheid te illustreren: Liever minder assumpties om de resultaten zo robuust mogelijk te maken. Mijn hang naar voorzichtigheid, misschien zelfs wel conservatisme, speelt ook een grote rol in mijn visie op onderwijs en toepassingen.

Aangeleerd.

Bij mijn visie op onderwijs en toepassingen trek ik psychometrie breder en spreek ik over kwantitatieve onderzoeksmethoden, wat ook in mijn leerstoelbeschrijving staat. In mijn optiek is het belangrijkste doel van academisch onderwijs het aanleren van een academische, onderzoekende houding. Voldoende kennis van de kwantitatieve onderzoeksmethodologie geeft die academische onderzoekende houding inhoud. Het probleem van het onderwijs in de kwantitatieve onderzoeksmethodologie is in een zin van drie woorden samen te vatten: 'Statistiek is moeilijk.' En hoewel er verschillen in aanleg zijn denk ik dat dit voor iedereen geldt; in ieder geval ook voor een bijzonder hoogleraar in de kwantitatieve onderzoeksmethodologie. Sinds de jaren zeventig van de vorige eeuw, weten we waarom: In een serie onderzoeken lieten de psychologen Tversky en Kahneman zien dat mensen geneigd zijn om verkeerd te redeneren als het om statistiek of kansrekening gaat zodat mensen slecht in staat zijn om de kans op een bepaalde gebeurtenis of de grootte van een bepaald effect in te schatten.[18]

Bij voorbeeld: Iemand gooit een eerlijke munt van €1 tien keer op, en krijgt dus tien keer óf een kop (K) óf een munt (M). Nu is de vraag: 'Welk resultaat is aannemelijker: KMMKKKMMKMK of MMMMMMMMMM?' Mensen zijn geneigd te denken dat het onregelmatige patroon waarschijnlijker is. Echter de kans op beide patronen is even klein: één op 1024 om precies te zijn. Bij de eerste worp is de kans op het gooien van munt 50%. Bij de tweede worp weet het euromuntje niet meer wat de eerste keer gegooid is en is de kans op munt weer 50%. De kans op elke combinatie is 25%. Bij de derde worp weet het euromuntje niet meer wat de voorgaande keren is gegooid. De kans op het gooien van munt is dus weer 50%, en de kans op elke combinatie 12,5%. Dus zowel de kans op het gooien van KMM (de eerste drie worpen in het onregelmatige patroon) als de kans op het gooien van MMM (de eerste drie worpen in het patroon met tien keer munt) is 12,5%. Bij vier, vijf, zes of meer keer gooien blijft dit principe van kracht. Waarom hebben mensen dan de neiging om te denken dat de kans om tien keer munt te gooien zeldzamer is? Tversky en Kahneman noemen dit de gambler's fallacy (de drogreden van de gokker). Mensen zijn geneigd te denken dat na een aantal keer munt gooien, het weer tijd wordt voor een kop, en denken er niet bij na dat het euromuntje geen geheugen heeft.

Een tweede voorbeeld betreft Christine. Christine is een 58 jarige alleenstaande vrouw, met een hoog IQ en een uitgesproken mening. Ze staat politiek links van het midden, is afgestudeerd filosoof en ze is bezorgd over maatschappelijke ongelijkheid en de vluchtelingenproblematiek. 'Wat acht u meer

aannemelijk: (A) Christine is een bankmedewerker of (B) Christine is een bankmedewerker die tevens vrijwilliger is bij vluchtelingenwerk?' Uit onderzoek blijkt dat veel mensen voor alternatief B kiezen. Echter, A is waarschijnlijker. Als Christine een bankmedewerker én vrijwilliger is, dan is ze per definitie ook een bankmedewerker. Bankmedewerkers die vrijwilliger zijn vormen een subgroep van alle bankmedewerkers. De kans dat Christine bankmedewerker is kan dus niet kleiner zijn dan de kans dat ze zowel bankmedewerker als vrijwilliger is. De reden dat veel mensen toch B kiezen komt doordat de beschrijving van Christine (links van het midden, bezorgd over sociale ongelijkheid en vluchtelingenproblematiek) representatief is voor een vrijwilliger bij vluchtelingenwerk. Tversky en Kahneman noemen dit de conjunction fallacy. De kans op de doorsnede van twee gebeurtenissen wordt groter geacht dan de kans op één van de gebeurtenissen afzonderlijk, en dat is iets wat helemaal niet kan. Ik laat het hierbij, maar er zijn tientallen voorbeelden waaruit blijkt dat mensen geneigd zijn bij het inschatten van kansen of effecten geen logische afweging te maken, maar gebruik maken van hun intuïtie die leidt tot een verkeerde beslissing.[19]

Wat leren we nu van het onderzoek van Tversky en Kahneman, behalve dat statistiek moeilijk is? We leren ook dat statistiek belangrijk is. We kunnen onze intuïtie niet vertrouwen en we kunnen kansen op een bepaalde gebeurtenis niet goed inschatten. Bij het bepalen van gedrag laten we ons leiden door wat we kennen. Als we tien keer een munt opwerpen is onze ervaring dat het patroon onregelmatig is, en dus denken we dat een bepaald onregelmatig patroon vaker voorkomt dan een bepaald regelmatig patroon. Als we iemand beoordelen doen we dat vaak op basis van de stereotype vooroordelen over de groep. Denkt u eens in hoe rampzalig het kan uitpakken als een leraar de rekenvaardigheid van een slimme maar enigszins brutale Marokkaanse leerling alleen op intuïtie beoordeelt. Als we willen voorkomen dat de stereotype vooroordelen een rol spelen in de beoordeling, dan moeten we gebruik maken van objectieve tests. Als we willen voorkomen dat onderzoeksuitkomsten bepaald worden door foute intuïtieve veronderstellingen, moeten we gebruik maken deugdelijke kwantitatieve methoden. Denny Borsboom zei twee maanden geleden bij zijn oratie 'Wie geen belangstelling heeft voor de statistiek heeft geen belangstelling voor de Psychologie.'[20] Hetzelfde kan gezegd worden voor de pedagogiek, en dat doe ik hier ook, al rek ik statistiek op naar kwantitatieve onderzoeksmethodologie: 'Wie geen belangstelling heeft voor kwantitatieve onderzoeksmethodologie heeft geen belangstelling voor de pedagogiek.' Dit heeft een drietal consequenties ons onderwijs.

Ten eerste moeten we met het onderwijs in kwantitatieve onderzoeksmethodologie zoveel mogelijk zien te voorkomen dat mensen in de valkuilen van

hun eigen intuïtie trappen. Dat betekent dat studenten bewust moeten worden van het feit dat hun eigen hersenen niet ingesteld zijn op statistiek, en dat het normaal is dat je een probleem waarin waarschijnlijkheid een rol speelt, in eerste instantie niet begrijpt. Merkwaardig genoeg leren we studenten niet dat ons denken feilbaar is. Mijn vriendin wees mij daar kort geleden op, en ik moest bekennen dat ik daar nooit over had nagedacht. Ik heb direct actie ondernomen en over twee weken moeten de studenten van de eerstejaars cursus Algemene Methodenleer en Statistiek ook bedenken of het waarschijnlijker is dat Christine een bankmedewerker is of een bankmedewerker die vrijwilliger is bij vluchtelingenwerk, en het driedeurenprobleem oplossen.[21] Idealiter raken studenten zich zodanig bewust van het falen van de intuïtie dat zij een soort alarmbellen ontwikkelen die afgaan als zij onderzoekspraktijken tegen komen waarvan ze wel geleerd hebben dat ze niet deugen of waarvan ze niet geleerd dat ze wel deugen. Voorbeelden van praktijken waarvan studenten wel moeten leren dat ze mogelijk niet deugen zijn claims gebaseerd op onderzoek met kleine steekproeven, claims gebaseerd op test- of vragenlijstonderzoek waarbij de tests of vragenlijsten niet vooraf onderzocht en goed bevonden zijn, en claims gebaseerd op onderzoek waarbij men de data dubbel gebruikt heeft. Voorbeelden van praktijken waarvan studenten niet geleerd hebben of ze wel deugen zijn claims gebaseerd op statistische analyses die ze niet begrijpen. In beide gevallen is voorzichtigheid het devies.

Ten tweede moet de docent in staat zijn om de moeilijke materie uit te leggen. De docent moet een vakman of vakvrouw zijn met hart voor de kwantitatieve methoden. Het is van belang dat de docent zelf onderzoek doet in de kwantitatieve onderzoeksmethodologie, zodat hij of zij weet wat er speelt en regelmatig zelf wordt uitgedaagd via het schrijven van artikelen in internationale wetenschappelijke tijdschriften en het houden van lezingen op congressen. Ik pleit daarom voor een gecombineerde onderwijs- en onderzoekstaak voor alle docenten in de kwantitatieve onderzoeksmethodologie, liefst binnen een kwantitatief methodologisch onderzoeksprogramma.

Ten slotte moeten de studenten en docenten beseffen dat een goede beheersing van de kwantitatieve onderzoeksmethodologie veel tijd kost. Uit het beroemde onderzoek van de psycholoog Anders Ericsson bleek dat het kritieke verschil tussen musici op het allerhoogste niveau en het één na hoogte niveau bestond uit de hoeveelheid tijd die aan oefenen werd besteed.[22] Op hun 20e verjaardag hadden de musici op het allerhoogste niveau ongeveer 10000 uur aan piano-oefeningen achter de rug. Musici op het één na hoogte niveau kwamen tot 5000 uur; betere amateurpianisten tot 2000 uur.

Als we dit afzetten tegen de Bachelorprogramma's Pedagogische Wetenschappen en Onderwijskunde hier aan de Universiteit van Amsterdam waar

756 uur (oftewel 15% van het gehele curriculum) aan verplichte cursussen in de kwantitatieve onderzoeksmethodologie wordt besteed, dan steekt dat er nogal schraal bij af. Het is ruim een derde van wat de betere amateurpianist voor zijn of haar 20e verjaardag aan oefening achter de rug heeft. Bij andere universiteiten is dit niet veel anders. Het oefenniveau van een amateurpianist is onhaalbaar binnen 180 studiepunten, want de cursussen in de pedagogiek en onderwijskunde, stages en scripties moeten ook in het programma passen; die zijn ook belangrijk. Het is echter onrealistisch om te denken dat studenten na een Bacheloropleiding voldoende kennis en kunde van kwantitatieve onderzoeksmethodologie hebben om ze zonder verdere nazorg in de maatschappij of de academia los te laten. Studenten die in de praktijk beslissingen moeten nemen waar waarschijnlijkheid een rol speelt, of die een wetenschappelijke carrière nastreven moeten bijgeschoold blijven worden. Het model dat ik zelf voorsta is dat van een leven lang leren. Onderzoekers en praktijkmensen blijven zich ontwikkelen op het gebied van de kwantitatieve onderzoeksmethodologie.

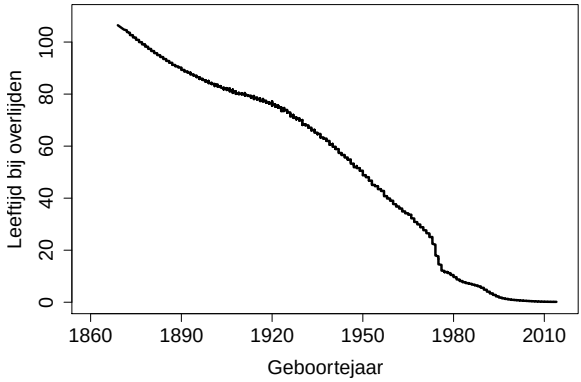
In het kader van mijn leeropdracht ben ik een tweetal projecten gestart om hier een bijdrage aan te leveren. Ten eerste het project Consultatie 2.0. Uit consultatiewerkzaamheden blijkt dat onderzoekers vaak dezelfde vragen hebben over kwantitatieve onderzoeksmethoden. Consultatie 2.0 bestaat uit een serie bijeenkomsten die een dagdeel beslaan met als doel meer en doelgericht te leren over een bepaalde methode. Elke bijeenkomst is als het ware een klassikale consultatie over een onderwerp waar onderzoekers meer over willen leren. De pilotbijeenkomst op 8 april heeft als onderwerp non-parametrische itemresponstheorie en Mokken schaalanalyse. Dit onderwerp heb ik vooral gekozen vanwege mijn eigen expertise. Daarna volgen omgaan met missing data en propensity-score matching, onderwerpen waar veel onderzoekers mee te maken hebben. Consultatie 2.0 bestaat uit een expertlezing, een lezing over software, lezingen door inhoudelijk onderzoekers die met het onderwerp te maken hebben gehad, en een vraag- en antwoordsessie.

Het tweede project betreft het beschikbaar maken van academisch onderwijs voor mensen die daar geen toegang tot hebben, zoals buitenpromovendi en leraren op scholen. Volgende maand start voor het eerst de cursus Ontwerpmethoden bij de Interfacultaire Lerarenopleiding. Mijn doel is dat deze cursus, al dan niet in aangepaste vorm, via webcolleges ook beschikbaar wordt voor leraren op scholen. Mogelijke vervolgprojecten zijn digitaal aangeboden opfriscursussen en deficiëntiecursussen in methoden en technieken, bijvoorbeeld via een verzameling kennisclips (dat zijn korte video's waarin een bepaald probleem uitgelegd wordt) of de meer uitgebreide Massive Open Online Courses (MOOCs): complete cursussen, inclusief tentamen die geheel via

het Internet aangeboden worden. Het is niet zo dat al dit materiaal nog ontwikkeld moet worden. Op het Internet is veel materiaal beschikbaar. Dit materiaal moet alleen zorgvuldig geselecteerd worden en aangeboden aan de doelgroep.

Toegepast

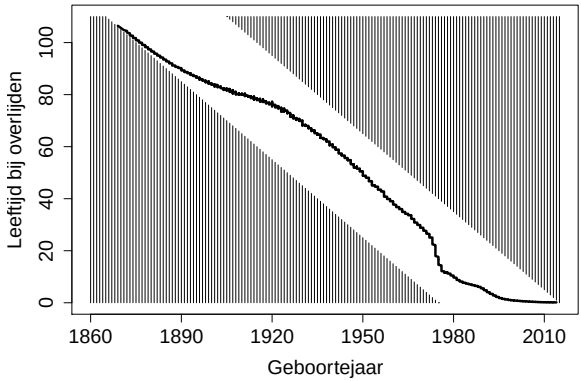
Ik kom nu bij de toepassing van de psychometrie en andere kwantitatieve onderzoeksmethoden. Als eerste kom ik terug op de replicatie van het onderzoek van Wainer, Palmer en Bradlow. De resultaten, weergegeven in Figuur 2, zijn opzienbarend. Het valt immers te verwachten dat de moderne mens steeds ouder wordt, en we hadden daarom een licht stijgende trend verwacht. De figuur laat duidelijk zien dat er een zeer sterk dalende trend is. Ik heb nog even gewacht met het opsturen van de resultaten naar Nature of Science omdat ik als psychometricus geleerd heb om voorzichtigheid te betrachten bij resultaten die ik niet begrijp. Wellicht ziet u wat hier gaande is.



Figuur 2: Gemiddelde leeftijd bij overlijden als functie van geboortejaar

Wat er aan de hand? De gegevens van de overleden Amsterdammers wijken op twee belangrijke punten af van de verwachting. Ten eerste ontbreken de gegevens in de rechter bovenhoek (figuur 3). Deze mensen leven nog. Immers een vrouw die in 2000 geboren is en 80 jaar wordt leeft nog steeds. Zij kan alleen in het register opgenomen zijn als zij niet ouder is geworden dan zestien jaar. Het kan dus niet anders dan dat de gemiddelde leeftijd bij overlijden

in de 21e eeuw erg laag ligt. Ten tweede ontbreken de gegevens in de hoek linksonder (figuur 3). De gegevens worden sinds 1975 bijgehouden. Amsterdammers die voor 1975 zijn gestorven komen niet in het register voor. Bijvoorbeeld, een man die in 1880 geboren is en op 80 jarige leeftijd in 1960 is overleden, is te vroeg overleden om in het register te zijn opgenomen. Hij moet minstens 95 jaar oud zijn geworden om in het register opgenomen te zijn. De gemiddelde leeftijd bij overlijden is aan het eind van de negentiende eeuw dus per definitie erg hoog. Er is hier sprake van een niet-representatieve steekproef: Amsterdammers die nog leven en Amsterdammers die voor 1975 zijn overleden ontbreken, en deze ontbrekende waarnemingen zorgen voor het merkwaardige effect dat de gemiddelde leeftijd afneemt.



Figuur 3: Verklaring van de gemiddelde leeftijd bij overlijden als functie van geboortejaar

Ook in de pedagogiek liggen niet-representatieve steekproeven op de loer. Bijvoorbeeld de steekproeven die gebruikt worden om de onderwijskwaliteit aan de UvA te meten. Op de UvA wordt een gestandaardiseerd meetinstrument gebruikt, UvAQ genaamd, om de kwaliteit van de cursussen te evalueren.[23] UvAQ bevat onder andere een aantal gesloten vragen over een bepaalde cursus; bijvoorbeeld: 'De hoorcolleges hielpen mij de stof te begrijpen'. De studenten antwoorden op een vijfpuntschaal waarbij een 1 betekent 'zeer oneens' en een 5 'zeer eens'. Na afloop van de cursus ontvangen de docent en de onderwijsdirecteur een rapportje met de gemiddelde scores. Die scores worden gebruikt om de kwaliteit van de cursus te beoordelen, hoewel UvAQ daar geen speciale richtlijnen voor geeft. Ik heb het niet speciaal gemunt op UvAQ,

en ik zie ook de voordelen van UvAQ, maar UvAQ is een mooi voorbeeld hier omdat iedereen op de UvA ermee te maken heeft.

Het probleem waar ik aandacht aan wil besteden is dat het studenten vrij staat om UvAQ al dan niet in te vullen. Het percentage studenten dat de UvAQ invult varieert per cursus; binnen mijn eigen cursussen tussen de 60% tot 103%, maar er zijn cursussen waarbij het percentage aanzienlijk lager is. Volgens de handleiding zijn voor cursussen met minimaal 50 geregistreerde studenten de evaluatieresultaten betrouwbaar als 30% van de studenten UvAQ invult. Voor kleinere cursussen ligt het percentage hoger.[24] Wat betrouwbaar in deze context betekent is niet helemaal duidelijk, maar bij grote cursussen worden de resultaten in ieder geval niet in twijfel getrokken als minimaal 30% van de geregistreerde studenten UvAQ heeft ingevuld.[25] Bij mijn cursus Testconstructie en onderzoeksverslaglegging heeft 60% UvAQ ingevuld. Omdat UvAQ geheel anoniem is, is het onduidelijk welke studenten de vragenlijst wél en welke studenten de vragenlijst níet hebben ingevuld. Zijn deze 60% de ontevreden studenten die hun grieven kwijt willen? In dat geval is de cursusevaluatie mogelijk te negatief. Zijn deze 60% de studenten die braaf hebben meegedaan met de cursus en daarom ook braaf de cursusevaluatie invullen? Dat geval zijn de resultaten wellicht veel te rooskleurig. Net als bij de overleden Amsterdammers ontbreken gegevens waardoor de resultaten niet zondermeer geïnterpreteerd kunnen worden. Bij de overleden Amsterdammers was achteraf te verklaren welke personen wel en niet in het databestand zaten zodat de resultaten achteraf te verklaard kunnen worden. Bij UvAQ is dat niet het geval. Als het UvAQ-rapport op basis van 60% respons gemiddeld 3 rapporteert op de vraag 'Ik kreeg voldoende feedback op mijn werk', dan kan bij een niet-representatieve deelname het gemiddelde in werkelijkheid tussen de 2.2 – Wat als zeer slecht gezien wordt – en de 3.8 – wat als goed gezien wordt – liggen. Als het gemiddelde gerapporteerd wordt op basis van basis van 30% respons (wat UvAQ betrouwbaar acht) dan kan bij een niet-representatieve steekproef het gemiddelde zelfs tussen de 1.6 en 4.4 liggen, vrijwel de gehele schaal. Bij de cursus Toegepaste Methodenleer en Statistiek hadden 95 studenten het evaluatieformulier ingevuld, terwijl er slechts 92 geregistreerd stonden. Uit mijn eigen gegevens bleek dat 150 studenten aan het tentamen hadden meegedaan. Een grote groep studenten stond dus blijkbaar niet geregistreerd. Hier ontbreken dus zowel studentantwoorden als studentregistraties, wat de interpretatie van de resultaten een zeer bemoeilijkt.

Nu denkt een aantal van u wellicht: Is het dan weer niet goed? Eerst propageert hij het aanleren en het gebruik van kwantitatieve onderzoeksmethoden in de pedagogiek en vervolgens doet hij een poging om een toepassing daarvan – een gestandaardiseerde vragenlijst nog wel – neer te sabelen. Het ant-

woord op deze retorische vraag is: 'Dat hangt hoe de resultaten geïnterpreteerd worden'. Allereerst moeten bij een deelname van 103% alle alarmbellen afgaan. Maar, alleen al op basis van de ontbrekende gegevens, zouden de resultaten van UvAQ zeer voorzichtig geïnterpreteerd moeten worden. UvAQ zou hoogstens een signaalfunctie mogen hebben van mogelijke misstanden. Ik vind het acceptabel als de onderwijsdirecteur om de tafel gaat zitten met een docent die een lage gemiddelde beoordeelding heeft gehad, om in alle openheid over zijn of haar onderwijs te praten. UvAQ heeft dan alleen een signaalfunctie. Ik vind het onacceptabel als het onderwijs van een docent met een lage beoordeling zonder verder onderzoek als onvoldoende bestempeld wordt. Behalve niet-representatieve steekproeven zijn er overigens meer redenen waarom men voorzichtig moet zijn met het interpreteren van de resultaten. Het kan zijn dat er irrelevante vragen in staan, of dat er wel relevante vragen niet staan.

Wat ik met het UvAQ-voorbeeld wil zeggen is dat het gebruik van kwantitatieve onderzoeksmethoden slechts de eerste stap is, het begrijpen van de kwantitatieve onderzoeksmethoden, het correct gebruik van de kwantitatieve onderzoeksmethoden en het begrip van zaken die mogelijk mis kunnen gaan, zijn vervolgstappen die noodzakelijk zijn voor juiste conclusies. Het blind toepassen van welke methodologie dan ook zie ik als potentieel gevaarlijk.

Gegeven de complexiteit van kwantitatieve onderzoeksmethoden kan consultatie een oplossing zijn om een kwantitatieve methode of techniek zo ver onder de knie te krijgen dat deze toegepast kan worden. Psychometrici adviseren vaak bij inhoudelijk onderzoek. De bijdrage kan variëren van een consult van 30 minuten tot een zeer substantiële bijdrage, zoals doen van alle analyses en het schrijven van de methodesectie. Laat ik voorop stellen dat ik consultatie en samenwerking enorm belangrijk vind. Binnen de programmagroep Methoden en Technieken is tien procent van de tijd gereserveerd voor consultatie, en ik coördineer deze consultaties ook. Ik heb gemerkt dat de bijeenkomst soms wat ongemakkelijk is. De volgende vergelijking beschrijft het gevoel van de samenwerking vanuit het oogpunt van de inhoudelijk onderzoeker.[26]

'Er zijn mensen die feestjes geven. Ze steken vaak veel tijd in de organisatie van die feestjes, ze halen verschillende soorten drank in huis, ze zijn vaak dagen aan het koken om lekkere hapjes voor te bereiden, ze sturen uitnodigingen, ze zorgen dat er dansmuziek is, ze versieren de woonkamer en proberen ervoor te zorgen dat iedereen in een goede stemming is. Er zijn ook mensen die nooit zelf een feestje geven, maar wel op de feestjes van anderen komen. Als ze daar eenmaal zijn beginnen ze direct met klagen: Er had veel meer bier gehaald moeten worden, het licht is te fel, de wijn is te zuur, de

koekjes hadden anders gebakken moeten worden, de muziek staat niet hard genoeg, en de kleur van de versiering had niet rood moeten zijn, maar blauw.'

De inhoudelijk onderzoeker moet met zijn of haar mooie en vaak ambitieuze onderzoek, waar zoveel tijd in heeft gezeten, met de billen bloot bij de psychometricus. De psychometricus heeft geleerd om zeer voorzichtig te zijn met het interpreteren van resultaten en om expliciet te zoeken naar aspecten in het onderzoek die mis kunnen gaan. Die verschillende insteken kan soms wel een botsen. De consultatie kan soepeler verlopen als we ons aan een paar stelregels houden.

Mijn eerste advies is om een psychometricus van begin af aan bij het onderzoek te betrekken en niet pas als de data verzameld zijn. Dit is al een oud advies. De Britse statisticus Sir Ronald Fisher zei: 'To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.'[27] De belangrijkste beslissingen van een onderzoek worden in de beginfase gemaakt. De onderzoeksvraag, het onderzoeksdesign, en de meetinstrumenten. Een cruciale fout hier kan nooit meer recht worden gezet als de data eenmaal verzameld zijn. Haal de psychometricus er van meet af aan bij. Het mes snijdt aan twee kanten. Enerzijds wordt de kans op cruciale fouten verminderd, en anderzijds is er geen geklaag meer op feestjes, want de psychometricus heeft zelf meegeholpen met het uitzoeken van de versiering.

Mijn tweede advies is om alleen resultaten te publiceren die de inhoudelijk onderzoeker zelf begrijpt, zelfs als de psychometricus coauteur is. Bij het consult kan de psychometricus suggesties doen, de methodesectie schrijven, eventueel syntax schrijven, maar het is de inhoudelijk onderzoeker die verantwoordelijk is voor de resultaten. Als de inhoudelijk onderzoeker nog niet begrijpt wat er gebeurt tijdens de kwantitatieve analyse, dan is het consult nog niet afgelopen. De meest verstandige stap is dan dat de psychometricus literatuur verstrekt zodat de inhoudelijk onderzoeker zichzelf kan bekwamen in de methode. In feite sluit dit aan op wat ik eerder zei over een leven lang leren, en de invoering van klassikale consultaties voor veelvoorkomende onderwerpen.

Mijn derde advies is het ontwikkelen van alarmbellen voor praktijken die mogelijk tot problemen leiden. Ook bij onderwijs sprak ik hier al over. In het boek Statistical rules of thumb beschrijft Gerald van Belle[28] een groot aantal do's en don'ts in onderzoek, bijvoorbeeld met betrekking tot steekproefgrootte, effectgrootte en onderzoeksdesign die kunnen helpen bij het ontwikkelen van deze alarmbellen. Als zowel de inhoudelijk onderzoeker als de psychometricus hier kennis van hebben genomen zal dit de samenwerking ten goede komen.

Ten slotte. Aan het einde van het consult moet duidelijkheid zijn over de vier belangrijkste vragen van het onderzoek: 1. Wat is de onderzoeksvraag? 2. Hoe wordt er gemeten? 3. Hoe worden de data verzameld? 4. Wat gaan de data vertellen? Goed nadenken over deze vragen voorafgaand aan het consult, en uiteraard voorafgaand aan het onderzoek zelf bespoedigen het consult.

In het kader van mijn leeropdracht werk ik naast het eerder genoemde Consultatie 2.0 ook met zogenaamde Academische werkplaatsen. Dit project staat nog in de kinderschoenen maar beoogt samenwerking op het gebied van onderwijsonderzoek van mensen uit beleid, universiteit, HBO en praktijk. Naar mijn overtuiging hangt het succes van Academische werkplaats af van de kwantitatieve onderzoeksmethodologie. Enerzijds het opleiden van de deelnemers met nog niet zo veel kennis van de kwantitatieve onderzoeksmethodologie, en anderzijds als een kwaliteitsgarantie bij het nieuw op te zetten onderwijsonderzoek.

Dankwoord

Aan het eind van deze rede wil ik graag enkele woorden van dank uitspreken. Geachte toehoorders, ik dank u allen dat u de moeite hebt genomen om hier naar mijn rede te luisteren. Ik weet dat een aantal van u een flinke reis hebben moeten maken om hier te komen. Ik hoop dat ik mijn visie op het onderzoek, onderwijs, en toepassing van de kwantitatieve onderzoeksmethodologie en van de psychometrie in het bijzonder goed heb kunnen overdragen. Een college, een vereniging en aantal mensen wil ik persoonlijk bedanken.

Ik dank het College van Bestuur van de Universiteit van Amsterdam voor het mogelijk maken van mijn benoeming. Ik dank Frans Oort, Geert-Jan Stams en Ton Notten, die samen het curatorium van de Kohnstamm-leerstoel vormen, voor hun vertrouwen en hun inzet tijdens de benoemingsprocedure. Ik dank het bestuur van de Vereniging ter Bevordering van de Studie der Pedagogiek voor het in mij gestelde vertrouwen. Jullie hebben een moedige keuze gemaakt om een bijzonder hoogleraar kwantitatieve onderzoeksmethodologie aan te stellen. Een keuze die door sommige leden wellicht met enige scepsis zal zijn ontvangen. Ik zal jullie vertrouwen niet beschamen en ik zie uit naar een vruchtbare samenwerking.

Ik dank mijn leermeesters: Harry Vorst, je hebt mij destijds als student psychologie aan de UvA opgemerkt, hard aan het werk gezet, en mij ervan overtuigd ik verder moest in de wetenschap. Ik ben je hiervoor zeer erkentelijk. Mijn promotores Peter van der Heijden en Ab Mooijaart: Onder jullie vakkundige leiding heb ik statistiek geleerd en ben ik zelfstandig geworden.

Jeroen Vermunt: Ik heb veel van je geleerd over dingen slim aanpakken op het gebied van zowel onderzoek en als bestuur. Ten slotte Klaas Sijtsma: 36 artikelen samen gepubliceerd zegt misschien wel genoeg. Ik ben je zeer erkentelijk voor de adviezen, bespiegelingen en uitstekende samenwerking de afgelopen zeventien jaar.

Ik kom nu aan bij mijn collega's. Allereerst een speciaal dank aan mijn coauteurs en promovendi voor de succesvolle en prettige samenwerking. Voorts dank ik mijn ex-collega's van het departement Methoden en Technieken van Onderzoek van Tilburg University voor de fijne collegiale samenwerking die ik daar vijftien jaar heb ondervonden, en mijn nieuwe collega's bij de programmagroep Methoden en Technieken voor het warme welkom. Ik ben blij dat wij echt een groep zijn die gezamenlijk de uitdagingen in onderwijs en onderzoek aan gaan, en ik ben ook blij dat we daar ook best succesvol in zijn. Ook de samenwerking met de andere collega's bij zowel Pedagogiek en Onderwijskunde als bij de Lerarenopleiding ervaar ik als zeer prettig.

Ik kom nu bij mijn familie en vrienden. Voor een gelukkig leven zijn vrienden onmisbaar. Hoewel ik niet iedereen meer zoveel zie als vroeger heb ik iedereen nog even lief als vroeger. Ik dank ieder van jullie dat jullie mijn leven gelukkig maken. In tegenstelling tot vrienden kun je familie niet kiezen. Ik heb met mijn familie daarom simpelweg geluk gehad. Een warm nest, steun, goede opleiding, de vrijheid om op mijn achttiende in mijn eentje liftend de wereld rond te trekken, maar met de verantwoordelijkheid dat volledig zelf te bekostigen, en met de zekerheid dat er altijd een thuis is. Door de decennia heen is daarin eigenlijk niets veranderd. Vader, moeder, jullie hebben het geweldig gedaan. Michiel, Thijs, Judith, Sarah en Fenne, jullie bijdrage hieraan kan niet overschat worden. Veel dank.

Tsja, nu is het gebruikelijk om de levenspartner te danken voor de opofferingen aan het thuisfront en vergiffenis te vragen voor al die tijd ik aan het werk ben geweest. Lieve Romke, dit gaat voor jou niet op. Het is allemaal wat onorthodoxer: Ik dank je liever voor je creatieve ideeën die ik graag van je pik om in mijn oratie te gebruiken; de leuke schoonfamilie die je meebracht; het samen lekker aan het werk zijn, in bed op zondagochtend, ieder met zijn eigen computer; de klop aan de deur om middernacht die voorbode is van nog een wijntje en een videootje. En voor talloze andere dingen. Heel veel dank.

Ik heb gezegd.

Noten

1. Wainer, H., Palmer, S., & Bradlow E.T., 'A selection of selection anomalies.' In: Chance, 11, p. 3-7, 1998
2. Met dank aan Dr. Annika Smits van Onderzoek Informatie en Statistiek Amsterdam.
3. 311027
4. https://nl.wikipedia.org/wiki/Psychometrie_(psychologie)
5. Antwoord: $^{10}\log 25 + {}^{10}\log 4 = {}^{10}\log (25 \cdot 4) = {}^{10}\log (100) = 2$.
6. Antwoord: Brussel
7. Antwoord 1000. Voor opslag op de harde schijf wordt de standaard gehanteerd dat 1 TB gelijk is aan 1000 MB. Voor processing wordt ervan uitgegaan dat 1 TB gelijk is aan $2^{10} = 1024$ MB.
8. In het onderwijs verzorgen de psychometrici vaak statistiekvakken en vakken in onderzoeksmethodologie. Bij consultatie beantwoorden psychometrici meestal ook vragen van statistische of algemeen methodologische aard. Het is dan ook niet verwonderlijk dat de termen psychometricus, statisticus en methodoloog binnen de sociale wetenschappen bijna inwisselbaar zijn. In mijn oratie zal ik aan deze drie taken onderzoek, onderwijs, en consultatie bespreken. Gezien mijn leeropdracht besteed ik relatief veel tijd aan onderwijs en consultatie.
9. Groenen, P.J.F. & Van der Ark, L.A., 'Visions of 70 years of psychometrics: The past, present, and future.' Statistica Neerlandica, 60, p. 135-144, 2006
10. De staat van de Pedagogiek. Congres van de vereniging ter Bevordering van de Studie der pedagogiek. Zeist, 5-6 oktober, 2015.
11. Retera, M. 2009. Dirkjan 15. Uitgever Mandarijn, p. 28, 2009.
12. Coombs, C.H., A theory of data. Wiley, New York, 1964, p. 5
13. Voor een inleiding zie Sijtsma, K., & Molenaar, I. W. Introduction to nonparametric item response theory. Sage, Thousand Oaks, 2002
14. Mokken, R.J., A theory and procedure of scale analysis. De Gruyter, Berlijn, 1969.
15. Van der Ark, L.A., 'New developments in Mokken scale analysis in R.' In: Journal of Statistical Software, 48(5), p. 1-27, 2012
16. Bijvoorbeeld Crisan, D.R., Van de Pol, J.E., & Van der Ark, L.A., 'Scalability coefficients for two-level polytomous item scores: An introduction and an application.' In: L.A. van der Ark, D.M. Bolt, W.-C. Wang, J.A. Douglas, & M. Wiberg (red.), Quantitative psychology research. Springer, New York, 2016
17. Oosterhuis, H.E.M., van der Ark, L.A., & Sijtsma, K., 'Standard errors and confidence intervals of norm statistics for educational and psychological tests'. Manuscript ter publicatie aangeboden.
18. Tversky, A., & Kahneman, D., ' Judgment under uncertainty: Heuristics and biases.' In: Science, 185, p. 1124-1131, 1974
19. Zie bijvoorbeeld https://en.wikipedia.org/wiki/List_of_cognitive_biases
20. Borsboom, D., Bodemschatten: Wat grondslagenonderzoek de psychologie te bieden heeft. Oratie uitgesproken op 25 november 2015.
21. https://nl.wikipedia.org/wiki/Driedeurenprobleem

22. Ericsson, K.A., Krampe, R.Th. & Tesch-Römer, C., ' The role of deliberate practice in the acquisition of expert performance. 'In: Psychological Review, 100, p. 363-406, 1993
23. http://wiki.uva.nl/uva-q/index.php/Hoofdpagina
24. Minder dan vijftien studenten: 70%; 16-35 studenten: 50%; 36-50 studenten: 35%; meer dan 50 studenten: 30% (zie http://wiki.uva.nl/uva-q/index.php/Betrouwbaarheid_evaluatierapport)
25. In feite zijn hier twee problemen. Ten eerste de grootte van de steekproef. Als de steekproef te klein is dan is de berekende gemiddelde tevredenheid instabiel. De term 'betrouwbaar' die UvAQ hanteert heeft hier waarschijnlijk mee te maken. Ten tweede de representativiteit van de steekproef. Dit is mogelijk een veel ernstiger probleem.
26. Met dank aan R. Rouw
27. R.A. Fisher, Presidential Address to the First Indian Statistical Congress, 1938
28. Van Belle, G., Statistical rules of thumb (2e ed.), Wiley, Hoboken, NJ, 2008