Universiteit Twente
de ondernemende universiteit

De kwaliteit van toetsen

door Prof.dr.ir. T.J.H.M. Eggen

De kwaliteit van toetsen

Rede uitgesproken bij
het aanvaarden van het ambt
van hoogleraar

Psychometrische Aspecten van Examinering

aan de Faculteit Gedragswetenschappen
van de Universiteit Twente
op donderdag 9 april 2009
door

door Prof.dr.ir. T.J.H.M. Eggen

De kwaliteit van toetsen

Meneer de Rector Magnificus, collega's, dames en heren, met deze toespraak geef ik aan de benoeming tot bijzonder hoogleraar Psychometrische Aspecten van Examinering aan de faculteit Gedragswetenschappen aan de Universiteit Twente te willen aanvaarden. Bij deze gelegenheid wil ik U graag duidelijk proberen te maken waarmee ik me in mijn vak, de psychometrie, mee bezig zal gaan houden.

De psychometrie is het wetenschapgebied dat modellen ontwerpt en toepasbaar maakt voor het meten van menselijke eigenschappen. Deze eigenschappen zijn nooit direct waarneembaar. Zo willen we bijvoorbeeld de intelligentie of de rekenvaardigheid van een persoon meten, maar kunnen we dat alleen maar doen op basis van het vaststellen of een aantal opgaven of opdrachten door die persoon goed of fout worden gemaakt. De verzameling opgaven, ook wel items genoemd, vormen het meetinstrument, dat we een toets, een test of een examen noemen.

Mensen die denken dat we met toetsen ooit foutloos een vaardigheid van een persoon kunnen vaststellen, moet ik teleurstellen. Als we twee keer dezelfde eigenschap meten bij een persoon met een toets dan zal daar, in tegenstelling tot bijvoorbeeld bij het meten van lengte, bijna altijd een verschillende uitslag uitkomen. Toetsen gaat altijd gepaard met meetfouten. De belangrijkste oorzaken van deze door toeval beheerste meetfouten liggen in de persoon die gemeten wordt, in de omstandigheden waarin getoetst wordt of in de opgaven of opdrachten die aangeboden worden.

De psychometrie probeert met behulp van statistische modellen grip te krijgen op deze meetfouten. De psychometrie kent meerdere toepassingsterreinen. Te denken valt daarbij aan de sociale wetenschappenen en met name de psychologie. Een ander groot toepassingsterrein is het onderwijs en daarover ga ik het vandaag vooral hebben.

In dit toepassingsgebied zou psychometrie misschien beter Onderwijskundig Meten[1] genoemd kunnen worden of wat in de Engelstalige literatuur wordt aangeduid met Educational Measurement. Onderwijskundig Meten is de wetenschap die van toetsen en examens meetinstrumenten maakt op grond waarvan verantwoorde beslissingen over personen kunnen worden genomen. Onderwijskundig Meten levert onder meer het gereedschap om de kwaliteit van toetsen vast te stellen en te verhogen.

Wat is toetsen eigenlijk precies?

Toetsen is het meten dat leidt tot een score die uitdrukt of een bepaalde kwalificatie wel of niet aan een persoon kan worden toegekend of in welke mate de persoon een bepaalde eigenschap heeft. In het eerste geval willen we classificeren, bijvoorbeeld in geslaagd of gezakt, en in het tweede willen we eigenlijk een wat preciezere uitslag, bijvoorbeeld een score op een examen van 32 punten.

Voor toetsen zijn afhankelijk van het gebied waar ze gebruikt worden en de specifieke doelstellingen veel verschillende benamingen in omloop. Zo kennen we tests, examens, assessments, proeven van bekwaamheid, tentamens, studietoetsen, proefwerken, etc . Als we de toetsvorm erbij betrekken worden er nog veel meer benamingen voor toetsen gebruikt, zoals een multiple choice toets, een computertoets, een gedragsproef, een practicum, een portifolio, een essay toets, een werkstuk, een paper etc.

16 miljoen deskundigen in Nederland

Met zoveel woorden voor in de kern hetzelfde begrip is het niet verwonderlijk dat bijna iedereen veel ervaring heeft met toetsen. En dus dat iedereen wel verstand van toetsen of examens heeft. Iedereen in deze zaal zal een of meerdere toetsen of examens hebben gemaakt. Sommigen onder U hebben alleen examens afgelegd en zijn daarvoor geslaagd of gezakt. Anderen hebben ook toetsen gemaakt in de betekenis dat ze de toetsen hebben samengesteld of zelfs vraag voor vraag hebben geconstrueerd. Dan zijn er onder U mensen die regelmatig de resultaten op toetsen analyseren en dan overzichten van de resultaten van de kandidaten en van de resultaten op de opgaven produceren om op basis daarvan conclusies over de kwaliteit van de toetsen proberen te trekken. Verder zijn er onder U nog personen die met regelmaat toetsen laten maken door kandidaten en anderen die er op toezien dat ze op een adequate manier worden afgenomen. Tenslotte zijn er onder U mensen die beslissen dat toetsen of examens gemaakt moeten worden, in welke betekenis dan ook.

Het feit dat iedereen verstand van ons vak heeft veel voordelen: bijna iedereen kan vrij snel een oordeel hebben over een toets, maar dat geeft soms ook extra moeilijkheden. Voor mensen die in de praktijk met toetsen geconfronteerd worden, zijn de achtergronden of de precieze doelstellingen soms onbekend of onduidelijk. Hun aanvankelijke kritiek of mening blijkt dan bij

volledige kennis van de achtergronden van de toets niet waar of mogelijk te zijn.
Ik wil dit toelichten aan de hand van twee persoonlijke ervaringen die gaan over de Eindtoets Basisonderwijs: de alom bekende Citotoets.

Het was op een zaterdagochtend in februari 1995 toen ik naar een voetbalwedstrijd van mijn kinderen aan het kijken was. Een mevrouw, die ik kende als de moeder van onze linksback sprak me aan. Wat was het geval: haar zoon had de uitslag op de Citotoets ontvangen en had in totaal 200 vragen goed en een score gekregen van 543. Nu waren ze hier best tevreden mee, maar opgevallen was dat een vriendje slechts 195 vragen goed had en een score van 544 had gekregen. Deze mevrouw vond dit niet eerlijk en bovendien twijfelde ze aan de waarde van die Cito score.
Waar deze mevrouw min of meer het slachtoffer van is geworden is dat zij niet precies wist hoe de Cito standaardscore, op een schaal tussen 500 en 550, tot stand komt. De Eindtoets Basisonderwijs bestond in 1995 verplicht voor alle scholen uit 60 opgaven over taal, 60 over informatieverwerking en 60 over rekenen. Daarnaast hadden scholen de keuze om ook eens nog 60 opgaven over wereldoriëntatie af te nemen. Op de school van het kind werden ook de wereldoriëntatie opgaven gemaakt en voor de uitslag van het totaal aantal goed van het kind werden alle goed beantwoorde vragen geteld. Voor de bepaling van de Cito standaardscore werden echter met het oog op de vergelijkbaarheid van deze score van alle Nederlandse leerlingen alleen de resultaten op taal, rekenen en informatieverwerking meegenomen en niet die op wereldoriëntatie[2].
Hierdoor is de door de mevrouw geconstateerde oneerlijkheid wel verklaarbaar. Maar feitelijk worstelt de Citotoets in deze situatie met enigszins strijdige eisen die er aan gesteld worden: een zo breed mogelijke dekking van alle leerstofgebieden, de vrijheid van scholen om te kiezen, de eis van nationale vergelijkbaarheid van de scores en de noodzaak dat het voor iedereen duidelijk is hoe de toetsuitslag tot stand komt.

Het tweede punt speelde recent. In december van 2008 las ik in krant dat er discussie plaatsvond over het tijdstip van afname van de Eindtoets Basisonderwijs. De toets zou wellicht niet meer in februari maar pas aan het einde van het schooljaar worden afgenomen. Het belangrijkste argument daarvoor was dat scholen meer tijd zouden kunnen gaan besteden aan het aanleren van de o zo belangrijke basisvaardigheden. De Citotoets resultaten zouden dan weliswaar geen rol meer spelen in de advisering over het ver-

volgonderwijs, maar nog wel gebruikt kunnen worden om de kwaliteit van de scholen te beoordelen.
Naar mijn idee had de journalist of zijn bronnen bij dit bericht het belangrijkste doel van het bestaan van de Eindtoets Basisonderwijs niet duidelijk voor ogen. Zo is het in Nederland wettelijk geregeld dat voor de toelating tot het voortgezette onderwijs naast het advies van de school en de mening van de ouders ook een uitslag op een objectieve toets noodzakelijk is.
De Eindtoets Basisonderwijs is met name geconstrueerd voor het doel goede individuele adviezen te geven over de kansen in het vervolgonderwijs. Verplaatsing naar het eind van het schooljaar zou dan heel veel praktische problemen opleveren. Verder werd een secundair doel, de beoordeling van scholen op basis van de Citotoets resultaten van de leerlingen, naar voren geschoven. En de vraag is of een toets, oorspronkelijk bedoeld voor het meten van individuele prestaties, wel even zo geschikt is voor het meten van de kwaliteit van de scholen.
In deze situatie hebben dus we te maken met twee verschillende doelen van dezelfde toets die niet altijd iedereen helder voor ogen heeft.

De kwaliteit van toetsen

Om op basis van toetsen verantwoorde beslissingen over mensen te kunnen nemen dienen deze aan bepaalde kwaliteitscriteria te voldoen.
Wat is eigenlijk de kwaliteit van toetsen? Hoe kunnen we die vaststellen? Als we het ons vak hebben over de kwaliteit van toetsen, dan doen we dat doorgaans op basis van criteria zoals die zijn beschreven in de “Standards for educational and psychological testing”[3] van de Amerikaanse beroepsverenigingen AERA[4], APA en NCME. of in europees verband naar de standaarden van de EFPA. In de richtlijnen die deze beroepsverenigingen en ook de ITC en de AEA-Europe hebben uitgeven gaat het om twee hoofdzaken:

1. de eisen die gesteld kunnen worden aan de personen die met toetsen werken: het gaat hierbij bijvoorbeeld om toetsontwikkelaars of de personen die toetsen afnemen;
2. de eisen die gesteld worden aan het toetsen zelf: aan de procedure en als meetinstrument.

In Nederland is op het gebied van de kwaliteit van toetsen, tests of examens het COTAN systeem[5] prominent. Dit systeem bevat niet alleen criteria voor de kwaliteit, maar is ook een beoordelingsinstrument van de kwaliteit van een toets.

Binnen dit systeem wordt een toets beoordeeld op de volgende vijf hoofdcategorieën

1. Uitgangspunten van de toetsconstructie.
   Hierbij gaat om de vaststelling wat er getoetst wordt en voor wie. Er moet een helder kader zijn waarop de toets gebaseerd is: bij persoonlijkheidseigenschappen is dat vaak de rol die een begrip in een theorie heeft. In het onderwijs wordt zo'n kader gegeven door bijvoorbeeld een verzameling kerndoelen of competenties, door eindtermen, door een kwalificatiedossier of door een examenprogramma. De vertaling van dit kader naar de inhoud van de toets en de keuze van toetsvorm, de inhoudsvaliditeit, wordt beoordeeld.
   Verder valt onder deze hoofdcategorie dat er helderheid wordt verschaft over het gebruiksdoel van (de resultaten van) de toets.
2. Kwaliteit van het testmateriaal en de toetsprocedure of handleiding.
   Onder deze categorie valt de beoordeling of de specifieke deelopdrachten in een toets aan de kwaliteitseisen voldoen. Verder is hier aan de orde dat voor de vergelijkbaarheid van de toetsresultaten er een gestandaardiseerde procedure voor afname en scoring nodig is.
3. Normen
   Een score op een toets krijgt pas betekenis wanneer er een vergelijking mogelijk is met referentiepunten. Om de waarde van een score op een toets te bepalen zijn er twee hoofdwegen mogelijk. Normgeoriënteerd: de score van een persoon wordt dan vergeleken met de scores van een (grote representatieve) groep. De andere hoofdweg is criteriumgeoriënteerd: de score wordt vergeleken met een vooraf vastgesteld criterium.
4. Betrouwbaarheid
   Dit gaat over de nauwkeurigheid waarmee de toetsresultaten worden gerapporteerd.
   Er kunnen verschillende vormen van betrouwbaarheid worden onderscheiden: zoals test-hertestbetrouwbaarheid, interbeoordelaarsbetrouwbaarheid of de nauwkeurigheid van een vaardigheidsschatting.
5. Validiteit
   Bij validiteit gaat om de mate waarin men op basis van de scores op de toets de bedoelde conclusies kan trekken. Er wordt hierbij aangesloten op de klassieke indeling in begripsvaliditeit en criteriumvaliditeit.

Het COTAN beoordelingssysteem is niet onomstreden. Met name bij de beoordeling van toetsen in het onderwijs of bij examens in de reguliere of particuliere sector voert men nogal eens aan dat de eisen die gesteld worden

eigenlijk niet van toepassing zijn op een toets of een examen. Ik zal proberen de belangrijkste discussiepunten, of zo men wil, kritiekpunten aan de orde te stellen:

1. Het COTAN systeem is in eerste instantie ontwikkeld om de kwaliteit van psychologische tests te beoordelen en deze hebben doorgaans als hoofddoel een enkele eigenschap of construct van een persoon te meten en op grond hiervan een uitspraak te doen over dat individu.
   Als we echter naar het gebruik van toetsen in het onderwijs kijken zijn de toetsdoelen vaak meervoudig: zoals in het voorbeeld van de Eindtoets Basisonderwijs waar men met dezelfde toets niet alleen de individuele leerling maar ook de school wil beoordelen.
   Daarnaast zijn de toetsdoelen soms zo uitgebreid dat deze niet met een enkelvoudig instrument getoetst kunnen worden. Een voorbeeld hiervan is de vaststelling of een in het onderwijs te ontwikkelen competentie bij een leerling aanwezig is. Zowel praktische als onderwijskundige redenen zijn er dan om de toetsing met meerdere deelinstrumenten te laten plaatsvinden.
2. Een volgend punt is dat het systeem te weinig rekening houdt met criteria die van belang zijn bij nieuwe moderne toetsvormen. Recent zijn er allerlei nieuwe toetsvormen ontstaan, zoals computertoetsen, simulaties en competente assessment programma's, en voor deze nieuwe toetsvormen zijn de oorspronkelijke criteria niet altijd specifiek genoeg.
3. Verder zouden de criteria niet specifiek genoeg zijn in relatie tot de veranderingen in de toetsdoelen. In een aantal onderwijskundig opvattingen over bijvoorbeeld competentiegericht opleiden en of over samenwerkend leren wordt als het hoofddoel van toetsen eerder gezien dat het mede bijdraagt aan het tot stand brengen van leren, dan dat zou moeten kunnen worden vastgesteld wat er geleerd is. Het gaat om toetsen om te leren, in plaats van toetsen van het geleerde.[6]
4. Een met het vorige punt samenhangend kritiekpunt is dat in het COTAN systeem te weinig gebruik zou worden gemaakt van nieuwe theoretische inzichten op het gebied van de (vaststelling van de) validiteit instrumenten. Het uitgangspunt is dat validiteit een geïntegreerd criterium is waarbij de geschiktheid van de interpretaties die gedaan worden op basis van de scores op de toets gerechtvaardigd dient te worden. Validiteit gaat dus niet over de toets maar over het rechtvaardigen van het gebruik van de scores op de toets. Omdat dit bij elk toetsdoel totaal anders kan zijn, kan de door COTAN gevolgde traditionele indeling in inhoudsvaliditeit, begripsvaliditeit en criteriumvaliditeit daarom soms te beperkt lijken. Door gebruik te

maken van werkwijzen om vast te stellen of aan alle aspecten van validiteit kan worden voldaan, zoals voorgesteld door Messick (1989)[7], en de manier waarop dit kan worden aangetoond, zoals voorgesteld door Kane (1992)[8], zou dat kunnen verbeteren.

5. Een volgend punt is de mening dat de gestelde kwaliteitseisen in de praktijk vaak niet haalbaar zijn. Het is begrijpelijk dat voor bepaalde toetsen, met name in het onderwijs, niet altijd aan in het algemeen gestelde kwaliteitscriteria kan worden voldaan. Het gebrek aan resources,-daarmee bedoelen we dan: geld, deskundigheid en personeel -, zijn hiervan meestal de oorzaak.
   Het onderscheid in het doel van het toetsen en het belang voor kandidaten van het resultaat zijn daarom bij het vaststellen de kwaliteitseisen essentieel. Het moet evident zijn dat waar belangrijke beslissingen over personen op grond van het toetsen worden genomen de kwaliteitseisen aan de toetsen hoog moeten zijn. Daar waar de belangen minder zijn kunnen ook lagere eisen gelden. Dat in de praktijk hierbij soms commerciële overwegingen een rol spelen moet hierbij goed in de gaten gehouden worden. In het algemeen wil ik stellen dat belangrijke beslissingen niet aan amateurs overgelaten moeten worden
6. Een laatste vaker gehoord kritiekpunt is dat de criteria te psychometrisch van aard zijn. Te veel nadruk zou er zijn voor de meetkwaliteit van de toetsen. Gehanteerde argumenten zijn dat andere aspecten van toetsen veel belangrijker zouden zijn. Ik kan het ermee eens zijn dat naast psychometrische criteria ook andere meer kwalitatieve aspecten voor de bepaling van de kwaliteit van toetsen van belang kunnen zijn. Afhankelijk van het doel van het toetsen zullen deze een groter of kleiner gewicht krijgen in de gehele beoordeling. Het kan echter nooit zo zijn dat voor het bepalen van de kwaliteit van toetsen de kwaliteit van het meten geen rol zou spelen. Voor een deel is deze kritiek overigens gebaseerd op onvoldoende kennis van de aanwezige psychometrische criteria. Veelvuldig wordt gedacht dat bij elk toetsdoel naar dezelfde criteria of indices moet worden gekeken. Heel vaak wordt bijvoorbeeld gekeken naar traditionele indices om de betrouwbaarheid van een toets te bepalen zoals een interne consistentie maat als Cronbach's alpha[9]. In veel situaties is deze index echter niet te bepalen of niet geschikt om de nauwkeurigheid van het toetsdoel dat nagestreefd wordt vast te stellen. Als het bijvoorbeeld gaat om het nemen van een slaag/zak beslissing is het veel belangrijker om als betrouwbaarheidscriterium van de toets te kijken naar het percentage te verwachten goede beslissingen.

> In de kritiek die er op de psychometrisch criteria is, wordt dus soms feitelijk bedoeld dat een traditionele criterium niet past bij het nieuwe of alternatieve toetsdoel. Het ontbreken van kennis van betere criteria mag dan evenwel niet leiden tot het afwijzen van het psychometrische criterium. Overigens wordt soms in kritieken op de psychometrische criteria soms zelfs bewust een andere inhoud gegeven aan begrippen als betrouwbaarheid en validiteit, waardoor slechts het label hetzelfde is gebleven maar de inhoud totaal anders. De daardoor ontstane verschillen van mening zijn daardoor feitelijk soms door niet meer dan spraakverwarring veroorzaakt.[10]

Naast het COTAN systeem zijn er diverse andere lijsten met kwaliteitsbeoordelingscriteria van toetsen in omloop. Sommige hebben een officiële status, zoals het systeem dat de Inspectie van het Onderwijs in het mbo toepast[11] en het systeem dat bestaat voor de doorverwijzing naar het leerweg ondersteunend onderwijs (LWOO) en het praktijk onderwijs (PrO)[12]. Andere systemen bevinden zich nog voornamelijk in de (wetenschappelijke) ontwikkelfase. Soms zijn de alternatieven niet meer dan een nadere uitwerking van de oorspronkelijke COTAN criteria, soms zijn dit alternatieven die mede gebaseerd zijn op de eerder geformuleerde kritiekpunten. Een aantal zijn toegesneden op bepaalde toetsvormen, zoals de (extra en of specifieke) eisen die men stelt aan toetsen die met de computer worden afgenomen[13] of aan de competentie assessment programma's[14]. Andere criteria zijn vaak beschreven voor het toetsen in bepaalde onderwijssector, zoals voor het toetsen in het hoger onderwijs.[15] Ten slotte zijn er criteria beschreven die in lijn met de kritiek beter zouden passen bij het toetsen in studentgericht en in competentiegericht onderwijs[16].

Het is mijn opvatting dat de verschillen tussen de diverse systemen op dit moment te klein zijn om groter te later worden. Ik denk dat in de huidige basis een uitdaging ligt voor het ontwerpen van een algemeen geldend kwaliteitscontrole en beoordelingssysteem voor toetsen in de ruimste zin van het woord.[17] Het moet mogelijk zijn de huidige COTAN indeling op hoofdcategorieën zondanig aan te passen dat deze als generiek kader kan dienen voor de beoordeling van toetsen.

Bij het ontwerpen van dit nieuwe systeem spelen de volgende aspecten een belangrijke rol:

1. Expliciete vaststelling van de doelen van de toets:
   In de basis kan men de volgende hoofddoelen van het toetsen onderscheiden:
   a. De evaluatie van het systeem: scores van individuele personen worden gebruikt om iets te zeggen over de eenheden waar ze deel van uitmaken: bijvoorbeeld de school (in accountability onderzoek) of het schoolsysteem (bij internationale vergelijkingen zoals bij het PISA-onderzoek[18])
   b. De ondersteuning van het leerproces: het toetsdoel is hier vaak het bewaken van de voortgang, zoals we dat kennen in zogenaamde leerlingvolgsystemen. Soms wordt dit aangevuld met het diagnosticeren van zwakke punten in de ontwikkeling om daarna de leerlingen te helpen.
   c. Het verbeteren van het ontwikkel of leerproces. Kenmerkend voor dit toetsdoel is dat door integratie van toetsen en instructie wordt nagestreefd het leerproces te verbeteren of zelfs te optimaliseren. Het toetsdoel is met name ondersteunend te zijn aan de ontwikkeling van de competenties van de lerende. Vooral in het beroepsonderwijs spreekt men dan over ontwikkelingsgericht toetsen en competentiegericht toetsen.
   d. Voor het nemen van een beslissing over de kandidaat.
      Traditioneel onderscheiden we hier selectie, plaatsing en certificering.
      Selectie: Hierbij worden op grond van de toetsuitslag de besten uit een groep beschikbare kandidaten gekozen: kenmerkend is dat er meer kandidaten zijn dan plaatsen
      Plaatsing: Bij een plaatsingsbeslissing is er voor elke kandidaat een plaats, maar op grond van de toets wordt aangegeven welke plaats het beste is: bij welke vervolgopleiding of welke niveau van een cursus het beste past.
      Certificering: Bij het certificeringsdoel geeft de toetsuitslag een erkenning die bijvoorbeeld toegang geeft tot een vervolgopleiding (zoals bij de eindexamen voortgezet onderwijs) of tot het recht voor het uitvoeren van bepaalde handelingen (autorijden) of om een bepaald beroep uit te oefenen (zoals bijvoorbeeld logopedist).
2. Het besef dat bij meerdere doelen van de toets, de kwaliteit van toetsen separaat voor elk doel vastgesteld moet worden.
   Dat betekent naar mijn idee dat als de voor de huidige onderwijspraktijk veelvuldig gepropageerde toetsen ter ondersteuning van het leerproces

ook gebruikt gaan worden voor het nemen van bijvoorbeeld certificeringsbeslissingen, deze toetsen op beide aspecten apart beoordeeld moeten worden. Evenzeer geldt dit voor het meervoudig gebruik van toetsen, waarbij de gegevens zowel voor de individuele leerling als voor de evaluatie van scholen worden gebruikt.

3. Toetsdoel afhankelijke kwaliteitscriteria en standaarden.
   Specifiek voor de verschillende toetsdoelen zullen de criteria nader moeten worden gespecificeerd en zullen voor elk toetsdoel kwaliteitsstandaarden geformuleerd moeten worden. Op grond van deze standaarden zal een beoordeling voor een bepaald gebruik kunnen worden gegeven. Per toetsdoel en ook per toetsvorm zullen de van toepassing zijnde criteria en ook de eisen kunnen verschillen.
4. Het ontwikkelen van criteria voor de kwaliteit van combinaties van toetsen.
   Bij veel toetsvormen, zoals bij portifolio's en proeven van bekwaamheid, bestaat de toets op grond waarvan een beslissing wordt genomen altijd uit meerdere te onderscheiden onderdelen. Voor verschillende relevante aspecten worden doorgaans verschillende deeltoetsen, vaak per persoon verschillende, afgenomen. De vraag is welke kwaliteitseisen men moet stellen, voor een bepaald toetsdoel, aan de combinatie van de deeltoetsen. Is het noodzakelijk is dat alle onderdelen een hoge kwaliteit moeten hebben om te kunnen garanderen dat het geheel een hoge kwaliteit heeft?[19]
5. De validiteit van het toetsen.
   Eerder heb ik aangegeven dat een deel van de kritiek op het huidige COTAN systeem betrekking heeft op de beperkte invalshoek die bij de vaststelling validiteit wordt benadrukt. Met name bij meer complexe toetsvormen is het nodig om na te gaan of het mogelijk is en hoe het zinvol is om de argumentatiebenadering van Kane (1992) voor de bewijsvoering van de validiteit in te zetten. Daarvoor is het ieder geval noodzakelijk, dat een uitwerking van deze algemene benadering voor een aantal concrete toetsvormen, zoals bijvoorbeeld een competentie assessment programma, ontwikkeld en onderzocht wordt.
6. Het ontwerpen van duidelijk omschreven procedures voor de toetspraktijk.
   Met de toetspraktijk bedoel ik hier de uitgevers van toetsen, de exameninstellingen en de toetsdeskundigen en verantwoordelijken in de verschillende onderwijssectoren.
   Als er namelijk algemene kwaliteitscriteria en standaarden voor toetsdoelen zullen zijn ontwikkeld, zullen deze voor de toetspraktijk pas van belang worden als duidelijk is hoe met concrete toetsen kan worden aan-

getoond dat men aan de standaarden voldoet. Uiteraard zal, naarmate de toetsvorm complexer is, hier meer behoefte aan zijn.

Een kwaliteitssysteem met de geschetste kenmerken is mijn idee nodig om bij de huidige toetspraktijk zinvolle uitspraken over de kwaliteit van toetsen te kunnen blijven doen. Een algemeen geaccepteerd systeem voor de vaststelling van de kwaliteit maakt immers ook duidelijk wat de mogelijkheden zijn om te werken aan de verbetering van de kwaliteit van toetsen.

Kwaliteitsverbetering

In ons vak is er terecht veel aandacht voor de verbetering van de kwaliteit van het toetsen. Voorbeelden zijn de verbetering van de toetsprocedures, het verbeteren van het toezicht, en de totstandkoming van en de discussie over het certificeren van personen die beroepsmatig met toetsen werken. Ik wil drie zaken noemen die naar mijn idee van groot belang zijn voor de kwaliteitsverbetering.

1. Onbekendheid van de kwaliteit.
   Helaas moet er geconstateerd worden dat over veel toetsen die worden gebruikt de kwaliteit eigenlijk onbekend is. Soms zijn alleen procedurele aspecten van de kwaliteit bekend, bij andere toetsen alleen inhoudelijke kwaliteitsaspecten, bij een aantal zijn er psychometrische aspecten bekend. Om een eindoordeel te hebben over de kwaliteit van een toets voor een bepaald toetsdoel is het absoluut noodzakelijk over alle drie aspecten informatie te hebben. Dit is zeker het geval als op basis van de toets belangrijke beslissingen worden genomen over een individuele kandidaat, zoals bij alle examens ter afsluiting van een onderwijsprogramma.
2. De zwakste schakel.
   Bij veel toetsen is het aangetoond dat het beoordelen door menselijke beoordelaars een zeer moeilijke zaak is.[20] Dit is een grote bedreiging voor de kwaliteit van de toets, omdat de beoordelaar immers deel uitmaakt van de toets. Zonder beoordelaar is er geen score en dus geen uitslag van de toets mogelijk. Het is een feit dat verschillende mensen een identieke situatie heel vaak anders hebben gezien. Zelfs als in complexe toetssituaties alle beoordelingsaspecten in detail zijn uitgewerkt, blijkt de overeenstemming tussen verschillende beoordelaars vaak gering. De afhankelijkheid van de specifieke beoordelaar voor het resultaat van de kandidaat is zeker

bij belangrijke toetsdoelen eigenlijk niet te accepteren. Dus elke inspanning om deze zwakke schakel sterker te maken verdient veel prioriteit. Te denken valt daarbij aan het psychometrisch modeleren van de verschillen die er zijn tussen de beoordelaars[21], de training van beoordelaars, en daar waar mogelijk door keuze van de toetsvorm, de vraagvorm en het beoordelingsvoorschrift, de beoordeling zoveel mogelijk te objectief te maken.

3. Computergestuurd toetsen
   Dit brengt mij bij het laatste punt dat ik aan de orde wil stellen. Ik wil het met U hebben over de kwaliteitsverbetering die te bereiken is met het inzetten van de computer bij het toetsen.

Computergestuurd toetsen

Alhoewel er nog enige hapering is en de ontwikkelingen niet zo snel zijn gegaan als men twintig jaar geleden dacht, ben ik van mening dat het computergestuurd toetsen in de nabije toekomst steeds meer zal toenemen. De belangrijkste redenen hiervoor zijn[22]:

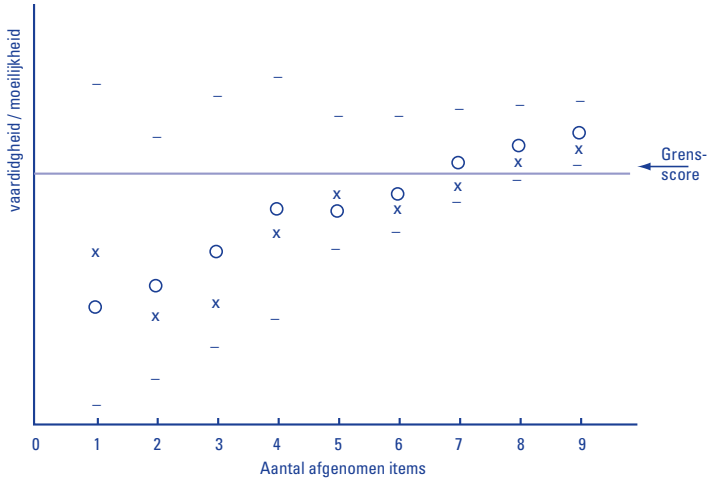
1. Met de computer zijn vaardigheden te toetsen die op papier niet getoetst zouden kunnen worden. De inhoudelijke meerwaarde is op diverse plaatsen aangetoond.[23]
2. Voor iedereen in ons land, maar vooral voor de jeugd die getoetst wordt, is de computer inmiddels een zo normaal onderdeel van hun leven dat zij de computer meer een standaard modaliteit vinden om zaken te doen, en dus ook om getoetst te worden, dan pen en papier.
3. De met de individualisering van onze maatschappij en in het bijzonder ook van ons onderwijs samenhangende eisen met betrekking tot de flexibiliteit van toetsen of examens kunnen bijna alleen worden bereikt als we op grotere schaal met de computer gaan toetsen. Het gaat hierbij om flexibiliteit ten aanzien van de inhoud, de plaats en het afnametijdstip van de toets.
4. Door de ontwikkelingen in de ICT is de routinematige ontwikkeling en afname nu op grote schaal mogelijk, zodat initiële investeringen sneller worden terugverdiend.
5. Het is te verwachten dat de kosten van het uitvoeren en van de logistiek van pen en papier toetsen van met name landelijke toetsprogramma's zullen toenemen, terwijl de kosten voor computergestuurd toetsen, ook door de groeiende mogelijkheden voor geautomatiseerde beoordeling, waarschijnlijk minder snel zullen stijgen.

Er bestaan twee hoofdvormen van computergestuurd toetsen: lineaire en adaptieve toetsen. In lineaire computergestuurde toetsen worden de individuele opgaven in een van te voren vastgestelde volgorde aan de kandidaat voorgelegd; een basisvorm hiervan is een bestaande papieren toets op het beeldscherm aan de kandidaat aan te bieden. Bij een adaptieve toets wordt de inhoud van de toets tijdens het toetsen aangepast aan de prestaties van de kandidaat.

Op computergestuurd adaptief toetsen (cat) ga ik nader in.
Wat zijn de belangrijkste kenmerken van een cat?

1. Bij een cat zijn de toetsafname en de toetssamenstelling computergestuurd en geïndividualiseerd.
2. Kandidaten krijgen niet allemaal dezelfde toets maar elke kandidaat krijgt de voor haar of hem best passende toets. De specifieke opgaven die iemand krijgt worden mede bepaald op basis van de antwoorden op eerdere aangeboden opgaven.
3. Een adaptieve toets wordt psychometrisch optimaal samengesteld onder praktische randvoorwaarden met betrekking tot de inhoud.
4. De nauwkeurigheid waarmee gemeten wordt met een cat wordt van te voren ingesteld.

Schematisch is de toetsing met een cat weergegeven in onderstaande figuur.



In de figuur zien we dat er één schaal is waarop zowel de moeilijkheid van de opgaven als de vaardigheid van de getoetste persoon kan worden afgebeeld. Een adaptieve toets start bijvoorbeeld met een opgave van gemiddelde moeilijkheid (x). Als deze opgave fout wordt gemaakt dan komt de geschatte vaardigheid van de persoon na deze eerste opgave (o) iets beneden de moeilijkheid te liggen. De onnauwkeurigheid van deze vaardigheidschatting is groot zoals weergegeven door de breedte van het betrouwbaarheidinterval. De tweede opgave wordt zo goed mogelijk passend bij de eerste vaardigheidsschatting gezocht. Als deze opgave goed wordt gemaakt wordt de nieuwe vaardigheidschatting groter en neemt de onzekerheid daaromtrent af, dat wil zeggen een kleiner interval. Indien, zoals in de figuur, er een duidelijke grensvaardigheidsscore tussen bijvoorbeeld voldoende en onvoldoende is, gaan we op deze wijze verder tot we met de vooraf ingestelde nauwkeurigheid de beslissing kunnen nemen.

Wat zijn nu de kwaliteitsverbeterende voordelen van een cat?[24]

1. De inhoud van de toets is optimaal afgestemd op de persoon: iedereen krijgt een andere toets die het best past bij zijn vaardigheidsniveau. Er is maximale flexibiliteit ten aanzien van de inhoud mogelijk.
2. De efficiëntie waarmee gemeten wordt met de toets is veel groter in vergelijking met een reguliere, lineaire toets. Onderzoek heeft uitgewezen dat met een cat slechts ongeveer van de helft opgaven nodig is om met dezelfde nauwkeurigheid uitspraken over de kandidaat te kunnen doen. Deze winst in efficiëntie is met name heel groot de bij extreem laag vaardige en hoog vaardige kandidaten.
3. Met een cat wordt er beter gemeten. Dit is het geval omdat er een op geldigheid getoetst psychometrisch model ten grondslag ligt aan de metingen. Bovendien kan de te bereiken nauwkeurigheid van de cat al voor daadwerkelijke toetsafname worden ingesteld.
4. Met lineaire computertoetsen delen cats de voordelen
   a. dat de toets onder gestandaardiseerde condities wordt afgenomen
   b. dat er objectief gescoord wordt;
   b. dat er inhoudelijk rijkere toetsmogelijkheden zijn; en
   c. dat er een grote flexibiliteit ten aanzien van de tijd en plaats van de toetsafname mogelijk is.

Voor een cat hebben we een verzameling opgaven nodig, die met allerlei kenmerken en op basis van empirisch onderzoek verkregen meeteigenschappen, worden opgenomen in een zogenaamde gekalibreerde itembank. Heel essentieel is verder dat we een psychometrische theorie hebben die ons

in staat stelt de kenmerken van opgaven zoals de moeilijkheid te beschrijven, het gereedschap biedt om opgaven tijdens het toetsen te selecteren en om de vaardigheid van de getoetste vast te stellen. Deze theorie is de itemresponstheorie[25] (IRT).
De basispsychometrie nodig voor cats werd al aan het eind van de jaren 80 van de vorige eeuw vastgesteld. Bij eenvoudige IRT modellen werden algoritmen ontwikkeld die gericht waren op het zo snel en zo nauwkeurig mogelijk schatten van de vaardigheid van een persoon. De schatting van de vaardigheid werd gebaseerd op statistische aannemelijkheidsmethoden, en de selectie van de opgaven op de maximale Fisher informatie bij de lopende vaardigheidsschatting. Voor de gehanteerde stopregel keek men naar de standaardfout van de schatter en het maximaal aantal af te nemen opgaven. Uiteraard werden er ook Bayesiaanse varianten ontwikkeld.
De toen in de praktijk ingevoerde cats waren sterk psychometrisch georiënteerd. Naarmate de gebruiksmogelijkheden van computertoetsen en de toepassing ervan toenam, is het psychometrisch onderzoek voor een groot deel gevoed door wensen die in het praktisch gebruik van de toetsen naar voren kwamen. Ook de in het Nederlandse onderwijs in gebruik zijnde cats hebben algoritmen die voldoen aan door de praktijk gegeven randvoorwaarden. Zo vindt bijvoorbeeld in de toetsen voor kleuters uit het leerlingvolgsysteem van Cito de opgavenselectie psychometrisch niet helemaal optimaal plaats, maar op een zodanige manier dat de succeskans op de opgaven voor deze jonge kinderen groot is. In de WISCAT-pabo, die toets waarvoor alle instromende eerstejaars pabostudenten een voldoende moeten halen, zijn meerdere praktische randvoorwaarden ten aanzien van de inhoud en het gebruik van de opgaven ingebouwd.
Helaas heb ik geen tijd meer om in detail op het huidige en toekomstige onderzoek in te gaan, maar enkele richtingen wil ik toch kort aangeven.

1. Algoritmen voor classificatie doelenden.
   In traditionele cats zijn de algoritmen gericht op de efficiënte schatting van de vaardigheid en wordt daarvoor gebruik gemaakt van statistische schattingsmethoden. Als het doel van de toets de juiste classificatie is van een persoon in één van een beperkt aantal categorieën, kan ook met succes gebruik gemaakt worden van sequentiële statistische toetsingsmethoden. Bij toepassing van de combinaties van zogenaamde, TSPRTs kan voldaan worden aan van te voren aan te geven maximale kansen op foute beslissingen.[26] Recent onderzoek heeft uitgewezen dat deze procedures in sommige gevallen door toevoeging van extra stopregels aan het algoritme van een cat nog aanmerkelijk verbeterd kunnen worden. Stochastisch afknot-

ting[27] voegt het basisidee toe dat je kunt stoppen met toetsen (in twee betekenissen nu) op het moment dat de kans dat je nog van beslissing verandert heel erg klein is. De mogelijke extra winst die door stochastisch afknotting van de SPRTs en door andere sequentiële statistische methoden mogelijk is in verschillende toepassingssituaties van adaptief toetsen zijn op dit moment onderwerp van onderzoek.

2. Itemselectie

   Het itemselectie algoritme is de basis voor de efficiëntie voordelen van elke cat toepassing. De laatste jaren zijn veel resultaten bereikt in het onderzoek naar selectie van de opgaven in cats, die voldoen aan allerlei praktische randvoorwaarden zonder dat dit grote verliezen heeft voor de efficiëntie van het toetsen.

   De twee belangrijkste zijn inhoudscontrole en afnamecontrole.

   Bij inhoudscontrole wordt er bijvoorbeeld voor gezorgd dat onderwerpen in een bepaalde volgorde in secties aan de orde komen en dat er binnen een sectie van een toets opgaven over deeldomeinen in de gewenste verhouding voorkomen.

   Bij afnamecontrole wordt ervoor gezorgd dat opgaven niet te vaak of te weinig worden geselecteerd.

   De beheersing van het gebruik van de opgaven in de beschikbare opgavenbanken is een onderzoeksterrein dat nog volop aandacht vraagt. De mogelijkheden voor de duur van de geheimhouding van opgaven bij op flexibele tijden en plaatsen aangeboden toetsafnames speelt hierbij een belangrijke rol. Duidelijk is dat computergestuurd adaptief toetsen hierbij voordelen heeft boven andere toetsvormen.

   Als laatste onderzoeksrichting bij itemselectie wil ik noemen dat het mogelijk is alternatieven voor itemselectie te ontwikkelen die niet per definitie uitgaan van het psychometrisch optimaal meten. Een mogelijkheid is de al genoemde sub-optimale selectie bij een gewenste moeilijkheidgraad. Daarnaast zou men hierbij ook kunnen denken aan het ontwikkelen van selectiecriteria bij toetsen, waarbij het feitelijke hoofddoel de ontwikkeling van de getoetste persoon is en niet de efficiëntie schatting van de vaardigheid. Selectiealgoritmen, waarbij de controle deels ook door de getoetste persoon zelf kan plaatsvinden, vormen een nog onontgonnen onderzoeksterrein.

Indien toepasbaar, is computergestuurd adaptief toetsen een mogelijkheid om de kwaliteit van bestaand en nieuwe toetsen te beheersen en te verbeteren. Naar mijn opvatting zou bij twee actuele toepassingsvelden in het

onderwijs de mogelijkheid voor het gebruik van computergestuurd adaptief toetsen moeten worden overwogen. De eerste is bij de vroegtijdige signalering van kinderen die mogelijk dyslectisch zijn[28]. Daarnaast zie ik toepassingsmogelijkheden bij het vaststellen van het bereiken van de referentieniveaus zoals gedefinieerd door de commissie Meijerink. in de doorlopende leerlijnen voor rekenen en taal van primair tot hoger onderwijs.[29] Recente psychometrische bevindingen, zoals bijvoorbeeld op het gebied van het klonen van opgaven[30] en van het gebruik van responstijden[31], zullen de toepassingsmogelijkheden alleen maar vergroten.

Ten slotte

In het voorgaande heb ik proberen aan te geven hoe naar de kwaliteit van toetsen zou kunnen worden gekeken en ben ik ook ingegaan op een aantal mogelijkheden om de kwaliteit te verbeteren. Ik heb het idee willen overbrengen dat je pas goed kunt werken aan de verbetering van de kwaliteit van toetsen, als je eenduidig hebt vastgelegd wat je met kwaliteit bedoeld.
In dit verband wil ik echter niet nalaten met U te delen, dat ik me ook wel eens zorgen maak over dit kwaliteitsdenken, dat niet alleen bij toetsen, maar bij veel zaken in onze maatschappij, een belangrijke rol speelt. Bij de toepassing van kwaliteitscontrole systemen bestaat namelijk het gevaar dat we de kwaliteit gaan nastreven van de instantie of de persoon die de kwaliteit controleert.
Zo wordt soms bij toetsen niet de maximaal haalbare kwaliteit nagestreefd, maar slechts die kwaliteit waarop de toets beoordeeld zal gaan worden. Vaak spelen hierbij economische motieven rol.
Ook in ons wetenschappelijk bedrijf zien we hetzelfde gevaar soms opdoemen: het onderzoek wordt afgesloten op het moment dat het geschreven verslag goed genoeg is om door een redactie van een tijdschrift geaccepteerd te worden voor publicatie.

Het streven naar hogere kwaliteit, zeker in de wetenschap, mag naar mijn idee niet ophouden bij de kwaliteit waar we zelf op beoordeeld worden.

De boodschap: “Als het niet kan zoals het moet, moet het maar zoals het kan” moet wat mij betreft veel vaker plaats maken voor “Als het niet beter moet dan het kan, dan moet het maar niet kunnen” .

Dankwoord

Beste mensen, graag wil ik U allen hartelijk bedanken voor de belangstelling die U heeft getoond voor deze bijeenkomst en voor mijn toespraak.
Ik weet dat velen van U er een flinke reis voor hebben moeten maken.
Verder wil ik graag even aandacht geven aan de mensen die een bijzondere rol spelen of hebben gespeeld in het feit dat ik hier vandaag sta. Mijn ouders en schoonouders konden deze gebeurtenis helaas niet meer bijwonen.
Yvonne, Frank, Wiesje en Meike wil ik laten weten dat ik me realiseer dat het feit dat ik hier sta soms ook ten koste van hun tijd is gegaan.

Graag noem ik hier professor Terpstra. Onder zijn verantwoordelijkheid volgde ik aan deze universiteit de studie Toegepaste Wiskunde, die achtertaf gelukkig, een flink mathematisch karakter had.
Pas veel jaren later ben ik dankzij hem gaan inzien dat procedures voor kwaliteitscontrole die gebruikt werden in de textielindustrie ook nuttig kunnen zijn bij het maken van goede toetsen en examens.

Dit brengt me bij mijn collega's. Ik heb er gelukkig veel. Tegen mijn Twentse collega's wil ik zeggen dat ik me hier direct weer welkom voelde. Ik waardeer het zeer hoe ik hier mijn gang kan gaan. Ik hoop dat we veel kunnen samenwerken en voor elkaar kunnen betekenen. Dat doen zeker mijn Cito collega's. Ik denk dat er geen betere manier is waarop ik al veel jaren de collegialiteit en samenwerking ervaar. Ik heb heel veel van mijn collega's geleerd en ben blij met de altijd open deuren van iedereen om met elkaar discussies te voeren, elkaar te helpen en problemen op te lossen. In het bijzonder wil ik Norman Verhelst bedanken. Ik denk dat ik van hem geleerd heb wat psychometrie is. Vooral de samenwerking in het PPON project, toen de eerste versie van OPLM door hem werd gemaakt, en tijdens het werken aan mijn proefschrift, onder zijn begeleiding, zijn voor mij zeer waardevol geweest.
Bijna een Cito leven lang is Piet Sanders mijn direct leidinggevende geweest. Van hem leerde ik vooral veel van wat psychometrie in de praktijk is. Graag wil ik hem bedanken omdat hij mij steeds gestimuleerd heeft en me heeft doen geloven dat ik ooit een oratie zou kunnen houden.

Ik ben de Universiteit Twente zeer erkentelijk dat ik weer een deel van mijn tijd op deze mooie universiteit kan doorbrengen. Ten slotte wil ik mijn werkgever Cito bedanken voor mijn benoeming en het daarmee in mij gestelde vertouwen.

Ik heb gezegd.

Noten

1 De laatste jaren is in de onderwijskunde ook het woord edumetrie geïntroduceerd. Toetsen heeft bij auteurs die deze term gebruiken doorgaans een heel specifiek doel. De algemene term Onderwijskundig Meten verdient daarom de voorkeur.

2 Tabel met (fictieve) resultaten op Eindtoets Basisonderwijs Cito

	Totaal aantal goed	Cito standaardscore	Aantal goed op taal, rekenen en informatie
Jeroen	200	543	150
Sjoerd	195	544	153

3 American Psychological Association, American Educational Research Association en National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington: American Psychological Association (APA).
De AERA, APA en NCME werken vanaf najaar 2008 gezamenlijk aan herziening van de Standards.

4 De gebuikte afkortingen staan voor:

AERA	American Educational Research Association
APA	American Psychological Association
NCME	National Council on Measurement in Education
ITC	International Test Commission
AEA	Association for Educational Assessment -Europe
EFPA	European Federation of Psychologist's Associations
COTAN	Commissie Test Aangelegenheden Nederland

5 COTAN (2004). *Documentatie van Tests & Testresearch in Nederland. Aanvulling 2004/01*. Amsterdam: Boom test uitgevers.

6 Zie bijvoorbeed in Dochy, F. & McDowell, L (1997). Assessment a tool for learning. *Studies in Educational Evaluation*, 23, 279-298.

7 Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational Measurement*. Third Edition. New York: Mc Millan

8 Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

9 Zie bijvoorbeeld Drenth, P.J.D.en Sijtsma,K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu van Loghum

10 Deze indruk kreeg ik toen ik hoofdstukken bestudeerde uit het boek Dochy, F., Heylen, L. & Van de Mosselaer, H. (Red.) (2002).*Assessment in onderwijs. Nieuwe toetsvormen en examinering in studentgericht onderwijs en competentiegericht onderwijs.* Utrecht: Uitgeverij Lemma BV.

11 Inspectie van het onderwijs (2007). Toezichtskader Examinering BVE 2007-2008.

12 Ministerie van het onderwijs (2008): Regeling screenings- en testinstrumenten leerwegondersteunend onderwijs (LWOO) en praktijkonderwijs (PrO) schooljaar 2009-2010.

13 Keuning, J. (2004). *De ontwikkeling van een beoordelingssysteem voor het beoordelen van "Computer Based Tests".* POK Memorandum 2004-1. Citogroep: Arnhem.

14 Zie Baartman, L.K.J. (2008). *Assessing the assessment. Development and use of quality criteria for Competence Assessment Programs.* Academisch proefschrift. Universiteit van Utrecht.en ook Wools, S.(2007). *Evaluatie van een instrument voor kwaliteits-beoordeling van competentie-assessments.* Utrecht/Arnhem:Universiteit van Utrecht/Cito.

15 Berkel, H.J.M. van, Bax A.E. (2006). *Toetsen in het hoger onderwijs.* Houten: Bohn Stafleu van Loghum.

16 Dochy, F., L Heylen & H. Van de Mosselaer (red.) (2002). *Assessment in onderwijs.* Lemma, Utrecht.

17 De werkzaamheden die op dit moment voor het herzien van het COTAN systeem plaatsvinden, kunnen daarvoor een goede start zijn.

18 PISA staat voor Programme for International Student Assessment. Om de drie jaar wordt onderzoek gedaan naar de kennis en vaardigheden van 15-jarigen in de belangrijkste geïndustrialiseerde landen.

19 Het aggregatie probleem, dat we op basis van meerdere metingen van een persoon tot een besluit moeten komen voor die persoon, heeft eigenlijk altijd al bestaan voor alle toetsen. Bijvoorbeeld bij eindexamens in het voortgezet onderwijs.
Bij toetsen waarop een slaag/zak beslissing wordt genomen op basis van de uitslagen van meerdere (deel)toetsen wordt dan het onderscheid gemaakt tussen conjuctieve, complementaire en compensatorsiche regels. Dit onderscheid in type regels die men bij combinatie van toetsuitslagen hanteert, zou ook richtinggevend kunnen zijn in de combinatie van de kwaliteitscriteria die moeten gelden voor een toets, die bestaat uit meerdere onderdelen.

20 Zie bijvoorbeeld de boeken van Hofstee, W.K.B. (1999) *Principes van de beoordeling. Methodiek en ethiek van selectie, examinering en evaluatie.* Lisse: Swets en Zeitlinger. en van Hendriks, P., Schoonman, W. (2006). *Handboek Assessment deel 1, gedragsproeven Ontwikkeling, implementatie en evaluatie.* Assen: Van Gorcum.

21 Zie bijvoorbeeld, Maris, G., Bechger, T.M. (2007). Scoring open ended questions. In C.R. Rao & S. Sinharay (Eds), *HandBook of Statistics.* Vol. 26., Ch. 20, p. 663-680. Elsevier: Amsterdam.

22 Zie ook: Brennan, R.H. (2006) (Ed.) *Educational Measurement , Fourth Edition.* Westport: American council on education and Praeger Publishers.

23 Zie bijvoorbeeld in het boek Cynthia G. Parshall,C.G., Spray, J.A. Kalohn, J.C. and Davey, T. (2002). *Practical Considerations in Computer-Based Testing.* New York: Springer-Verlag.

24 Goede overzichten van (de psychometrie van) computergestuurde adaptieve toetsen zijn te vinden in: Van der Linden, W.J. & Glas, C.A.W. (Eds.) (2000). *Computerized adaptive testing. Theory and practice.* Dordrecht: Kluwer Academic Publishers. en in Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer.* Second edition. Hilsdale, NJ: Lawrence Erlbaum Associates.

25 Zie bijvoorbeeld: Van der Linden, W.J. & Hambleton, R.K. (Eds.) (1996). *Handbook of modern item response theory.* New-York: Springer-Verlag.

26 TSPRTs zijn Truncated Sequential Probability Ratio tests In mijn proefschrift staat hier veel over geschreven: Eggen, T.J.H.M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* Arnhem: Citogroep.

27 Zie Finkelman, M . (2008). On Using Stochastic Curtailment to Shorten the SPRT in Sequential Mastery Testing. *Journal of Educational and Behavioral Statistics,* 33, 4, 442–463 en ook het afstudeerwerk van J.T. Wouda: *Computerized classification testing in more than two categories by using stochastic curtailment.* Unpublished master's thesis, University of Amsterdam.

28 De recente proefschriften van J. Keuning en J. Vloedgraven maken duidelijk dat daar inhoudelijk en psychometrisch mogelijkheden voor zijn.
Keuning, J.(2008) *Modelling growth in reading and Spelling.* Proefschrift Radboud Universiteit Nijmegen.
Vloedgraven, J (2009). *Development of phonological awareness in relation to literacy* Proefschrift Radboud Universiteit Nijmegen.

29 In 2008 rapporteerde de door de staatssecretaris van het Ministerie van OCW, ingestelde Expertgroep Doorlopende Leerlijnen Taal en Rekenen (Commissie Meijerink): *'Over de drempels met taal en rekenen'*

30 Glas, C.A.W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. Applied Psychological Measurement, 27, 249-263.

31 Klein Entink, R. (2009).Statistical models for responses and response times Proefschrift Universteit Twente

Universiteit Twente
de ondernemende universiteit