

*Tussen getal en betekenis ligt de brug
van verantwoording*

*Meten is pas relevant wanneer het leidt
tot rechtvaardige beslissingen*

*Niet de uitslag van de toets is beslissend,
maar de interpretatie*



VOORWOORD

Dit is een boek dat ik altijd al eens had willen schrijven. Mijn studie psychologie aan de Universiteit van Amsterdam met de specialisatie onderwijspsychologie, bracht mij in aanraking met twee disciplines: testleer en toetsleer. Die beide disciplines heb ik verder mijn hele werklevens met me meegedragen. En op een bepaald moment ga je je afvragen hoe ze zijn ontwikkeld en wanneer ze met elkaar verweven zijn geraakt. Met andere woorden, je gaat op zoek naar de mijlpalen. Je wilt weten wanneer er belangrijke stappen zijn gezet die voor een kentering hebben gezorgd in de discipline en wie dat hebben gedaan. Daarnaast begin je je af te vragen wat de belangrijkste thema's of onderwerpen zijn binnen de disciplines. Binnen iedere discipline worden voortdurend nieuwigheden aangedragen. Het gros daarvan sterft een zachte dood; slechts enkele overleven. Welke dat zijn en waarom juist deze, zijn voor mij interessante vragen.

Deze overwegingen hebben geleid tot dit boek. Het boek bestaat uit twee delen. Deel 1 bevat portretten van wetenschappers die in mijn ogen een wezenlijke bijdrage hebben geleverd aan de ontwikkeling van (een deel van) de discipline en in deel 2 beschrijf ik de onderwerpen die in de huidige tijd een hoofdrol vertolken in de discipline. Ik noem ze verhalen omdat de beschrijving ervan een bepaalde opbouw hebben, van een wettende start tot aan een robuuste theorie.

Het eerste deel gaat dus over personen die mijn wereld, althans mijn werklevens daaruit, vorm en kleur hebben gegeven. Ik ben ze tegengekomen in mijn studie in de jaren zeventig van de vorige eeuw. Niet in persoon, op een enkele uitzondering na, maar via hun boeken en artikelen. Voor een jong persoon als ik was de inhoud daarvan in eerste instantie erg abstract. Het stond ver van mij af.

Het te bestuderen materiaal werd aangedragen door docenten, die het materiaal kennelijk van belang vonden om van mij een goed (onderwijs)psycholoog te maken. Zelf heb je dat op dat moment niet door. Weer driehonderd bladzijden met moeilijke begrippen en vele formules! De docenten zullen het zich op dat moment niet hebben gerealiseerd, maar het kwartje valt niet altijd op het moment dat je de tekst bestudeert. Dat gebeurt meestal pas later, na het afstuderen wanneer je de studiestof nodig hebt om in de praktijk toe te passen of wanneer je zelf, net als ik, docent bent geworden die het studenten moet uitleggen. Dan begin je de studiestof te begrijpen en ontstaat het inzicht dat de leerstof van toen eigenlijk geniaal was en nog steeds is.

Bij leerstof horen mensen die het hebben geschreven. Wanneer je zelf onderzoeker bent geworden, ga je de drang voelen om te publiceren. Soms komt die drang van buitenaf en soms vanuit een bepaald soort trots die je wilt delen met anderen. Dan ervaar je ook dat het niet makkelijk is je gedachten coherent op papier te zetten. Op dat moment ga je aan die personen denken die de leerstof voor je studie hebben geschreven. Je realiseert je dat ook zij te maken hebben gehad met de dilemma's waar jij nu voor staat: hoe draag je je boodschap uit op papier?

Zij zijn daarin geslaagd en de essentie van de leerstof is bij mij overgekomen. Vervolgens ben ik geïnteresseerd geraakt in deze personen. Wie zijn zij? Wat hebben ze zelf gestudeerd? Hoe zijn hun carrières verlopen? Waar werk(t)en zij en werk(t)en ze samen met anderen of waren het solisten? En van wie hebben zij het vak geleerd, met andere woorden: wie zijn hun leermeesters geweest? En welke thema's in de test- en toetstheorie hebben zij tot de hunne gemaakt? Wat is hun invloed op de test- en toetstheorie en bestaat die nog steeds? Wie hebben hun werk voortgezet? Ik noem ze bruggenbouwers om redenen die ik uitleg in de verantwoording.

De vragen die hiervoor zijn geformuleerd heb ik willen beantwoorden bij het maken van hun portretten. Wat zit er achter een naam van de schrijver van een boek of artikel? Maar ook, welke verhalen hebben zij geschreven om de test- en toetsleer verder te helpen? Dat hebben zij niet alleen gedaan. Het is karakteristiek voor de wetenschap dat het cumulatief is. Dat wil zeggen, wetenschappers sluiten zich aan bij een school, zoeken daar een onderwerp uit dat hen ligt en proberen het verder uit te werken. Zij worden zo een deel van het verhaal. Die verhalen, in alfabetische volgorde, vormen het tweede gedeelte van dit boek. Die gaan over onderwerpen die zich continu verder ontwikkelen. Er is geen begin aan te wijzen en zeker geen einde. Hun verhalen vormen de hoekstenen van de test- en toetsleer en hebben betrekking op belangrijke thema's. Wat behelst het thema? Waarom is het belangrijk? Wie zijn vooraanstaand? Op al die vragen, en meer, probeer ik antwoord te geven.

Al gaandeweg het schrijven begon mij wat op te vallen. Vooral de latere bruggenbouwers, te beginnen bij Cronbach, Meehl en Messick, kregen interesse in het onderwerp validiteit. De tot dan toe heersende opvatting dat er drie soorten validiteit te onderscheiden zijn, inhouds-, begrips, en criteriumvaliditeit, werd gaandeweg als te strikt gezien en daarom verlaten. Zij, maar ook anderen, gingen er voortaan van uit dat alle "soorten" validiteit onder één noemer zijn te brengen: *constructvalidity*. Het gaat er niet meer om om aan de test zelf bepaalde eigenschappen te geven, maar om het accent te leggen op de interpretatie van de scores en op de beslissingen die op basis van de scores worden genomen en hoe rechtvaardig die gevolgen zijn. Kane's argument-gebaseerde benadering vraagt om expliciete onderbouwing: welke conclusies trekken we uit scores en welk bewijs ondersteunt dat. Dus de nadruk is komen te liggen op het aanreiken van bewijzen voor de rechtvaardigheid van beslissingen die op basis van de testscore worden genomen. Vandaar de ondertitel van het boek: Van meten naar rechtvaardig beslissen. De vroege bruggenbouwers deden erg hun best om de nieuwe begrippen in de test- en toetsleer zo goed mogelijk te operationaliseren, te meten en te berekenen. Daarna kwamen bruggenbouwers die de uitkomsten van die berekeningen als startpunt namen. Zij gingen eisen stellen aan de gevolgen die de uitkomsten van metingen teweegbrachten.

Dit boek is geschreven voor toetsdeskundigen, psychometrici, psychologen, onderwijskundigen en studenten die de testleer, de toetsleer en de psychologie aan het bestuderen zijn. Voor hen die willen begrijpen waar hun vak vandaan komt en waar het naartoe gaat. Het boek is een eerbetoon aan de bruggenbouwers die *meten* veranderden in *begrijpen*.

Ik dank Anneke Bax voor het kritisch lezen van het manuscript én voor haar toestemming om een van haar schilderstukken¹ te gebruiken voor de buitenkant van dit boek.

Henk van Berkel
Maastricht
Voorjaar 2026

¹De naam van het doek is Winter en het is het vierde schilderij in de reeks De Vier Jaargetijden uit 2022.



INHOUD

VOORWOORD	4
VERANTWOORDING	8
INLEIDING	11

DEEL 1

DE BRUGGENBOUWERS	16
1. W. H. Angoff	17
2. R.L. Brennan	24
3. L.J. Cronbach	29
4. R.L. Ebel	35
5. L. Guttman	40
6. R.K. Hambleton	46
7. M.T. Kane	52
8. F.J. Kelly	56
9. F.M. Lord	61
10. P.E. Meehl	67
11. W.A. Mehrens	72
12. S.J. Messick	77
13. R.J. Mislevy	84
14. G.W. Rasch	91
15. C.E. Spearman	97
16. R.L. Thorndike	102
17. W.J. van der Linden	109
18. A.A. von Davier	116
19. H. Wainer	121
20. D.J. Weiss	126
21. Twintig bruggenbouwers: wat hebben ze gebracht?	131



DEEL 2

DE VERHALEN 136

22. Beleid en Ethiek	137
23. Betrouwbaarheidstheorie.....	141
24. Computerized adaptive testing (CAT).....	147
25. Differential item functioning (DIF)	152
26. Generaliseerbaarheidstheorie (G-theorie)	156
27. Itembanking	160
28. Itemresponstheorie (IRT)	164
29. Klassieke testtheorie (KTT)	169
30. Meerkeuzevragen	174
31. Normeren en Cijfergeven	179
32. Onderwijsevaluatie.....	187
33. Test equating.....	190
34. Toetsconstructie.....	194
35. Validiteitstheorie	199
36. De verhalen: van meten naar rechtvaardigen	204

REFERENTIES208

AFKORTINGEN221

BEGRIPPENLIJST222



VERANTWOORDING

De selectie van twintig bruggenbouwers

De selectie van twintig bruggenbouwers in dit boek berust op een weloverwogen inhoudelijke en conceptuele afweging. Het doel van het boek is niet het bieden van een uitputtend overzicht van alle invloedrijke onderzoekers binnen de psychometrie, toetsleer of testleer, maar het presenteren van een representatieve en samenhangende selectie van denkers die bepalend zijn geweest voor de ontwikkeling van het vakgebied en die elk op eigen wijze een verbindende rol hebben gespeeld tussen theorie, methodologie en praktijk.

De geselecteerden vertegenwoordigen gezamenlijk de historische ontwikkeling van de test- en toetsleer vanaf het begin van de twintigste eeuw tot heden. Zij markeren sleutelperioden en keerpunten in deze ontwikkeling, waaronder de formalisering van de klassieke testtheorie, de introductie en uitwerking van factoranalyse en itemresponstheorie, de ontwikkeling van generaliseerbaarheidstheorie, en de opkomst van computer-adaptieve en Bayesiaanse toetsmodellen. Elk van de twintig personen kan worden gezien als een cruciale schakel binnen deze intellectuele continuïteit.

Naast de historische spreiding is de selectie thematisch gemotiveerd. De bruggenbouwers zijn gekozen omdat zij een duidelijke en duurzame bijdrage hebben geleverd aan een specifiek deelgebied van de test- en toetsleer, zoals betrouwbaarheid, validiteit, normering, *equating*, *differential item functioning*, itembanking en toetsconstructie. Kenmerkend voor hun werk is dat deze bijdragen niet beperkt zijn gebleven tot abstracte theorievorming, maar expliciet zijn uitgewerkt in toepasbare modellen, procedures en standaarden voor testpraktijken in de context van onderwijs en beroep.

Een centraal selectiecriteria is het bruggenbouwende karakter van hun werk. De opgenomen auteurs hebben systematisch verbindingen gelegd tussen ogenschijnlijk tegengestelde domeinen: tussen statistische modellering en pedagogische interpretatie, tussen technische precisie en inhoudelijke betekenisgeving, en tussen methodologische kwaliteit en normatieve vragen over rechtvaardigheid, transparantie en verantwoord toetsgebruik. Bovenal hebben zij een brug geslagen tussen hun theoretische werk en de toetspraktijk. Hun werk laat zien dat testpraktijken slechts kunnen ontstaan wanneer deze verbindingen expliciet worden gemaakt en onderhouden.

Het aantal van twintig is het resultaat van bewuste afbakening. Het biedt ruimte om de diversiteit en complexiteit van het vakgebied recht te doen, terwijl tegelijkertijd de conceptuele samenhang behouden blijft. De selectie beoogt representativiteit in plaats van volledigheid: iedere opgenomen bruggenbouwer voegt een essentieel én origineel perspectief toe aan het geheel, en het weglaten van een van hen zou de inhoudelijke balans en samenhang verzwakken.

De centrale gedachte achter dit boek is dat vooruitgang in wetenschap, onderwijs en maatschappelijke besluitvorming niet ontstaat door afzonderlijke theorieën of technieken, maar door het bouwen van bruggen: tussen disciplines, tussen theorie en praktijk, tussen meten en betekenis geven, tussen testleer en toetsleer en tussen technische precisie en ethische verantwoordelijkheid. De gekozen auteurs en hun werk belichamen elk op eigen wijze deze brugfunctie.

Ten slotte is bij de selectie rekening gehouden met de blijvende invloed van de bruggenbouwers. Hun ideeën, modellen en concepten vormen nog steeds het fundament van hedendaagse test- en toetspraktijken, onderzoeksprogramma's en kwaliteitsstandaarden. In die zin worden zij niet uitsluitend gepresenteerd als

historische figuren, maar als blijvend relevante referentiepunten in het denken over meten, interpreteren en verantwoorden. De twintig bruggenbouwers vormen gezamenlijk een kader waarbinnen de ontwikkeling van de toets- en testleer kan worden begrepen als een continu proces van verbinden, verfijnen en legitimeren. Hun selectie is daarmee niet willekeurig, maar inhoudelijk en methodologisch verantwoord.

De selectie van de veertien verhalen

De selectie bestrijkt een periode waarin de test- en toetswetenschap zich ontwikkelde van een voornamelijk praktisch ambacht tot een theoretisch en methodologisch onderbouwd vakgebied. Elk van de veertien verhalen vertegenwoordigt een cruciale schakel in deze ontwikkeling. Sommige bijdragen leggen het fundament, bijvoorbeeld door het formuleren van kernbegrippen als betrouwbaarheid, validiteit en schaalconstructie. Andere verhalen tonen hoe deze concepten werden verfijnd, bekritiseerd of uitgebreid, bijvoorbeeld door aandacht te vragen voor context, beoordelaarsvariatie of maatschappelijke consequenties van toetsgebruik. Samen laten zij zien hoe het vakgebied zich niet lineair, maar in dialoog heeft ontwikkeld.

Het aantal van veertien verhalen is het resultaat van bewuste selectie en afbakening. Het boek beoogt geen volledigheid, maar representativiteit: elk opgenomen verhaal voegt een origineel perspectief toe aan het geheel. Het weglaten van één bijdrage zou de conceptuele samenhang verzwakken. Samen vormen deze veertien verhalen een samenhangend geheel dat inzicht biedt in de intellectuele bruggen die het vakgebied hebben gedragen en blijven dragen.

Werkwijze

Bij de beschrijving van de biografieën van de bruggenbouwers is primair gebruikt gemaakt van openbare bronnen die via internet raadpleegbaar zijn. De exacte bronnen staan in het betreffende hoofdstuk vermeld. Die werkwijze was niet mogelijk bij het schrijven van de verhalen omdat die verhalen bescheiden literatuurstudies zijn waar bronnen voor moeten worden geraadpleegd die voor een groot deel niet openbaar zijn maar via bibliotheekbezoek wel toegankelijk zijn. De verwijzingen staan in de tekst en de exacte referenties zijn gebundeld en in de referentielijst die achteraan in het boek is opgenomen.

Structuur van de hoofdstukken

Voor de hoofdstukken over de bruggenbouwers is steeds gezocht naar de volgende informatie:

- Opleiding: welke opleidingen hebben de bruggenbouwers genoten, inclusief een mogelijk doctoraat?
- Loopbaan: hoe verliepen de carrières van de bruggenbouwers?
- Professionele functies: welke functies hebben de bruggenbouwers vervuld op basis van hun inhoudelijke expertise, bijvoorbeeld redactielidmaatschappen?
- Onderscheidingen en prijzen: Welke onderscheidingen hebben de bruggenbouwers op grond van hun expertise ontvangen, waaronder eredoctoraten? Welke prijzen hebben zij in de loop van hun carrières ontvangen?
- Leermeesters: Welke invloedrijke begeleider of mentor die hun intellectuele en/of inhoudelijke vorming sterk heeft bepaald, hebben de bruggenbouwers gehad?
- Onderzoeksgebieden: op welke onderdelen van de test- of toetsleer zijn de bruggenbouwers invloedrijke wetenschappers (geweest)?
- Invloed: wat voor invloed hebben de bruggenbouwers (gehad) op de test- of toetsleer? Hebben zij leerlingen opgeleid?
- Belangrijkste publicaties: welke publicaties van de bruggenbouwers hebben de grootste impact gehad binnen het vakgebied?
- Bronnen: welke openbare bronnen zijn geraadpleegd bij het portretteren van de bruggenbouwers?



In de hoofdstukken over De Verhalen is de gezochte informatie als volgt gegroepeerd:

- Inleiding en definitie
- Wat was de noodzaak om het onderwerp ter hand te nemen?
- Hoe is de ontwikkeling geweest? Hoe zal de ontwikkeling in de toekomst gaan?
- Welke problemen lost het op? Welke problemen resteren?
- Wie waren primair de bedenkers?
- Wie zijn nu de deskundigen?

Leesadvies

Alle hoofdstukken zijn afzonderlijk en in willekeurige volgorde te lezen. Dat impliceert wel dat in de tekst verdubbelingen voorkomen.

Definities

- Een *toets* is elk instrument waarmee het leerresultaat van een student wordt gemeten (bijvoorbeeld: schriftelijk examen, presentatie, portfolio, opdracht).
- *Toetsleer* is het onderwijskundige vakgebied dat zich bezighoudt met het ontwerpen, uitvoeren, beoordelen en verbeteren van toetsen, met als doel op een verantwoorde en eerlijke manier uitspraken te doen over het leren en presteren van studenten.
- *Toetstheorie* is het theoretisch en wetenschappelijk kader dat beschrijft, modelleert en verklaart hoe metingen van kennis, vaardigheden en competenties tot stand komen, en onder welke voorwaarden deze metingen betrouwbaar, valide en rechtvaardig zijn.
- Een *test* is een gestandaardiseerd meetinstrument waarmee, onder gelijke en gecontroleerde omstandigheden, (psychologische) kenmerken van personen worden gemeten, met als doel vergelijkbare en interpreteerbare uitspraken te doen over individuele verschillen.
- *Testleer* is het vakgebied dat zich bezighoudt met de ontwikkeling, analyse, evaluatie en toepassing van (psychologische) tests, met als doel verantwoorde, betrouwbare en valide metingen van (psychologische) kenmerken mogelijk te maken.
- *Testtheorie* is het theoretisch en wetenschappelijk kader dat beschrijft, modelleert en verklaart hoe metingen op tests tot stand komen, en onder welke voorwaarde deze metingen betrouwbare, valide en betekenisvolle uitspraken toelaten over latente kenmerken van personen.
- *Psychometrie* is het deelgebied van de psychologie dat zich bezighoudt met het ontwikkelen, analyseren en toepassen van meetinstrumenten om psychologische kenmerken te meten, zoals intelligentie, persoonlijkheid, vaardigheden, attitudes en emoties. Psychometrie gaat dus over het meten van psychologische eigenschappen.

Verder

- Met de aanduidingen student, leerling, kandidaat en deelnemer, worden steeds personen benoemd die een test of toets afleggen;
- Met de aanduidingen docent, examinerator en instelling worden steeds de verantwoordelijken voor de test of toets bedoeld;
- De samenvattingen van de hoofdstukken zijn mede met behulp van ChatGPT geschreven.



INLEIDING

De inleiding schetst de historische en conceptuele ontwikkeling van de psychometrie, beginnend bij de vroege intelligentietest van Binet en Simon en de introductie van het IQ. Het laat zien hoe het meten van abstracte eigenschappen als intelligentie leidde tot de opkomst van de psychometrie als kwantitatieve wetenschap, met Spearman als vroege grondlegger. Centraal staat de menselijke behoefte om prestaties niet willekeurig, maar systematisch en rechtvaardig te meten. Het boek positioneert de te bespreken psychometristen als bruggenbouwers die testleer, toetsleer en toetspraktijk met elkaar verbinden. Aan de hand van Angoff, Cronbach, Brennan en Ebel wordt in deze inleiding duidelijk hoe toetsing zich ontwikkelde tot een discipline waarin betrouwbaarheid, validiteit en maatschappelijke verantwoordelijkheid onlosmakelijk met elkaar zijn verbonden.

In 1905 publiceerden Binet samen met zijn medewerker Simon de, naar wordt aangenomen, eerste intelligentietest. De test werd later bekend als de Binet-Simon test. Zij ontwikkelden deze test als hulpmiddel bij het maken van onderscheid tussen kinderen die wel of geen speciaal onderwijs nodig hadden. Daartoe hadden ze een soort vragenlijst ontwikkeld die tot doel had te discrimineren tussen leerlingen. Zij wilden dus met de ontwikkelde test toekomstige leerproblemen voorspellen. Pas later nadat Stern in 1912 de mentale leeftijd (= de uitkomst van de test) relateerde aan de werkelijke leeftijd van de leerling, kreeg de test de toevoeging intelligentie. Stern vermenigvuldigde dit quotiënt met 100 en zo is de term IQ ontstaan. De IQ-test werd een methode om individuele verschillen in kaart te brengen. Daardoor werd de intelligentietest gehanteerd in het kader van de persoonlijkheidsleer, al is het beter te spreken van een intelligentietest omdat al snel meerdere testen werden ontwikkeld, alle met verschillende namen.

Het intelligentiequotiënt is in feite een kwantitatieve operationalisering van het abstracte begrip intelligentie. Later werd dit een latente variabele genoemd. De kwantitatieve operationalisering van het begrip intelligentie bood onderzoekers allerlei mogelijkheden om ermee te gaan rekenen en er een theorie omheen te bouwen. Overigens, Galton was, rond de eeuwwisseling tussen de 18^{de} en 19^{de} eeuw, de eerste die het idee opperde dat intelligentie meetbaar is, maar hij slaagde er niet in om een betrouwbare methode te ontwikkelen.

Spearman stelde in 1904, dus voordat de IQ-test haar intrede deed, dat de menselijke eigenschap intelligentie, een algemene factor heeft (de *g-factor*) en enkele specifieke (de *s-factoren*). Door de komst van de IQ-test werd het gemakkelijker te rekenen met het begrip intelligentie. Spearman introduceerde ook de factoranalyse om onderzoek te verrichten naar de *g-factor*. Zo was hij de eerste die van testonderzoek een kwantitatieve wetenschap maakte. Na hem kwamen andere onderzoekers die met de intelligentietest aan de slag gingen en er telkens een theorie aanbouwden. Zij werden zo de eerste *psychometristen*.

Er bestaat een oude menselijke wens om grip te krijgen op datgene wat zich niet gemakkelijk laat vangen, bijvoorbeeld talent, aanleg, kennis, redeneren en persoonlijkheid. In allerlei culturen en tijden hebben mensen geprobeerd elkaar te beoordelen, te selecteren, te onderwijzen en te rangschikken. Maar telkens opnieuw doemde dezelfde vraag op: hoe meten we op een manier die niet willekeurig is? Hoe scheppen we orde in de veelvormigheid van menselijke prestaties? Hoe rechtvaardigen we beslissingen die voortkomen uit één enkel getal op een schaal?

Het boek vertelt de geschiedenis van de personen, en in bredere zin van de denkers, die deze vragen niet alleen stelden, maar ze trachtten te beantwoorden door nieuwe theorieën, nieuwe methoden en



een nieuw ethos van zorgvuldigheid te ontwikkelen. Zij leefden in een tijd waarin toetsing nog verre van gestandaardiseerd was, in een wereld waarin docenten vaak individueel bepaalden wat een toets precies inhield, hoe streng of mild ze beoordeelden, en welke score voldoende was. Langzaam maar zeker begon het besef te groeien dat een toets geen persoonlijk oordeel kon blijven, maar een systematisch ontworpen instrument moest zijn. Een toets moet recht doen aan de persoon die ze beoordeelt, ongeacht wie de toets afneemt en wanneer dat gebeurt.

Ter illustratie, vier bruggenbouwers die in dit boek worden besproken, Angoff, Brennan, Cronbach en Ebel, hebben op hun eigen manier die zoektocht vormgegeven. Zij bouwden geen bruggen van steen of staal, maar intellectuele bruggen tussen disciplines die elkaar nodig hebben: de psychologie die wil begrijpen wát een prestatie betekent, de psychometrie die wil vastleggen hóé je die prestatie kunt meten, en de onderwijskunde die wil bepalen hoe en of studenten van het onderwijs wat hebben opgestoken. Hun denken is niet los te zien van de tijd waarin zij leefden. Ze werkten in een eeuw waarin massaal onderwijs groeit, waarin het onderwijs, en in het bijzonder het hoger onderwijs, toegankelijker werd voor bredere bevolkingsgroepen, waarin toetsing een instrument werd voor selectie en doorstroom, maar ook voor emancipatie en gelijke kansen.

Het verhaal van Angoff begint bij ETS, de Educational Testing Service, in Princeton. Daar ging hij zich bezighouden met de vraag hoe scores van verschillende toetsversies vergelijkbaar zijn te maken zijn, hoe vragen eerlijk kunnen worden geanalyseerd, en hoe experts op een systematische manier kunnen bepalen welke score als cesuur moet gelden. Zijn naam is onlosmakelijk verbonden met de methode die later beroemd zou worden als de Angoff-methode: een procedure waarin deskundigen inschatten welk percentage minimaal competente kandidaten een item correct zal beantwoorden. Het klinkt eenvoudig, maar het idee was vernieuwend. Niet langer werd een cesuur bepaald op basis van traditie, politieke afspraken of groepsgegevens; het werd een inhoudelijk beredeneerde schatting, gebaseerd op expertise en onderbouwd met itemanalyse. Angoff wilde een toetswereld waarin elke beslissing uitlegbaar is, waarin “voldoende” niet toevallig maar principieel is.

Wat Angoff onderscheidde, was zijn vermogen om de praktische problemen van grote toetsprogramma's te begrijpen en ze te vertalen naar oplossingen die theoretisch én uitvoerbaar zijn. De SAT en TOEFL staan symbool voor een nieuw tijdperk van massatoetsing, maar juist daardoor wordt het gebrek aan vergelijkbaarheid van scores een urgent probleem. Angoff bestudeerde schaalconstructie, *equating* en normering, onderwerpen die tot dan toe vooral statistische specialismen waren, maar die hij begrijpelijk maakte voor toetsontwikkelaars. Zijn publicaties, zoals *Scales, Norms, and Equivalent Scores*, zijn handboeken die zowel de wiskundige precisie als de morele verplichting van toetsing laten zien. Hij vond dat validiteit niet alleen een technische eigenschap is, maar ook een maatschappelijke verantwoordelijkheid: een test moet eerlijk zijn voor iedereen die hem maakt!

Chronologisch gezien was Cronbach een generatiegenoot van Angoff, maar in intellectueel opzicht vormden zij twee tegenpolen. Waar Angoff zich richtte op de operationalisering van toetsprocessen, boog Cronbach zich over de fundamenteën van het meten zelf. In 1951 publiceerde hij het artikel dat een van de meest geciteerde in de sociale wetenschappen zou gaan worden: zijn beschrijving van alfa, een maat voor de interne consistentie van tests en toetsen. Met alfa gaf Cronbach onderzoekers een manier om te bepalen in welke mate items binnen een toets hetzelfde construct meten. Hoewel de maat later kritiek zal krijgen omdat hij soms te ondoordacht wordt gebruikt, is de impact enorm. Cronbachs alfa wordt een taal, een gemeenschappelijk begrip dat psychologen en onderwijskundigen in staat stelt om op dezelfde manier te praten over de kwaliteit van hun instrumenten.



Maar Cronbach bleef niet bij alfa staan. Hij was een denker die altijd verder wilde kijken dan het getal. Zijn latere werk, geschreven samen met Meehl, richtte zich op constructvaliditeit, het idee dat validiteit niet een eigenschap van de toets is, maar van de interpretatie van toetsresultaten. Hiermee belichtte hij een nieuw perspectief: waar alfa iets zegt over de inhoud van items, zegt constructvaliditeit iets over de betekenis van scores.

Voor Cronbach was meten nooit los te zien van de context waarin het gebeurt. Hij vond dat toetsresultaten niet zonder meer moeten worden gezien als objectieve weergaven van bekwaamheid. Interpretatie vraagt om theorie, nuance en begrip van de situatie waarin iemand presteert. In zijn publicaties over evaluatieonderzoek verzette hij zich tegen simplistische aggregaties van resultaten. Onderwijscontexten zijn rommelig, divers, complex en volgens Cronbach moest evaluatie die complexiteit erkennen in plaats van reduceren.

Juist in deze kritische reflectie herkent men een gedachte die ook vandaag nog uiterst actueel is: dat cijfers schijnzekerheid suggereren en dat betekenis altijd meer vraagt dan statistiek. Cronbach is misschien wel de meest filosofische van de bruggenbouwers: hij stelt vragen die het vakgebied dwingen tot zelfreflectie.

Brennan, een leerling van deze traditie maar ook een vernieuwer op eigen kracht, ging een ander onderzoeksgebied van Cronbach, de generaliseerbaarheidstheorie (G-theorie), verder uitwerken tot een volwaardig raamwerk. Waar klassieke testtheorie uitgaat van één ongedifferentieerde foutterm, maakt de G-theorie het mogelijk om meerdere bronnen van variatie te onderscheiden: items, beoordelaars, taken, afnamesituaties. Brennan beseftte dat, in een wereld waarin toetsen steeds vaker gebruikmaken van open vragen, schrijfoopdrachten en *authentic tasks*, beoordelaarsvariatie een probleem wordt. Zijn werk laat zien dat de betrouwbaarheid van een score niet alleen afhangt van de kandidaat, maar ook van de nauwkeurigheid van de beoordelaar, de aard van de taak en de interactie tussen die factoren.

In zekere zin stond Brennan op de schouders van Cronbach, maar hij gaf diens ideeën een praktisch gevolg. Hij ontwikkelde een theorie die eerst vooral theoretisch was en die later meer praktisch werd voor toetsontwikkelaars die moeten beslissen hoeveel taken of beoordelaars nodig zijn. Zo werd hij een van de belangrijkste architecten van betrouwbaarheid in complexere evaluatievormen, een rol die alleen maar relevanter werd naarmate toetsen zich ontwikkelen in de richting van *authentic assessment* en competentiemetingen.

Hoewel Angoff, Cronbach en Brennan veel aandacht kregen voor hun theoretische bijdragen, mag Ebel niet ontbreken in deze inleiding. Hij bracht toetsing terug naar haar didactische kern. Zijn werk richtte zich niet primair op statistiek of filosofie, maar op de vraag hoe toetsing onderwijs kan ondersteunen. Als docent en later als hoogleraar zag hij dagelijks hoe een toets het leren van studenten beïnvloedt: welke vragen worden gesteld ("Meneer, komt dit onderwerp op het tentamen?"), welke kennis wordt belangrijk gemaakt ("Meneer, krijgen we hier vragen over?"), welke vaardigheden worden zichtbaar en welke verdwijnen onder de radar.

De bekende Ebel-matrix, waarin items worden geordend op basis van hun moeilijkheid en relevantie, is veel meer dan een classificatieschema. Het was Ebels manier om docenten te dwingen na te denken over de inhoud van hun vak. Wat wil je echt meten? Wat moeten studenten echt weten of kunnen? Wat is essentieel en wat is bijzaak?

Ebel bracht een pedagogische gevoeligheid in de psychometrie die soms dreigde te verdwijnen in de wiskundige elegantie van theoretici. Hij herinnerde het veld eraan dat een toets begint bij onderwijs:



bij leerdoelen en bij betekenisvolle inhoud. Zijn boeken vormden generaties docenten en onderzoekers in hun denken over toetskwaliteit, niet als abstracte eigenschap, maar als integraal onderdeel van onderwijsontwerp.

De samenhang tussen deze vier psychometrici, Angoff, Cronbach, Brennan en Ebel, wordt pas zichtbaar wanneer men hen beschouwt als personages in een groter verhaal. Zij werken niet als één school of gezamenlijk project, maar hun ideeën vormen een doorlopende dialoog. Angoff zorgt dat de infrastructuur van grootschalige toetsing eerlijk functioneert; Cronbach bekritiseert de fundamenteën en geeft ze nuance; Brennan verfijnt de modellering van foutbronnen; Ebel zorgt dat toetsing geworteld blijft in betekenisvol leren.

Gezamenlijk leggen zij de basis voor wat men vandaag moderne psychometrie noemt, een discipline waarin statistiek, psychologie, filosofie en onderwijskunde elkaar voortdurend beïnvloeden. De beoefenaars van deze disciplines leven in een tijd waarin toetsing een steeds grotere rol krijgt in de samenleving, in onderwijs, maar ook daarbuiten: in beroepen, in immigratieprocedures, in selectieprocessen, in beleid. Daarom hebben hun theorieën en methoden ook een maatschappelijke betekenis.

Toetsen impliceren kansen. Zij openen deuren, sluiten ze soms, voorspellen succes, bepalen toegang tot beroepen, instituties en sociale mobiliteit. Dat maakt toetsing tot een van de krachtigste mechanismen in moderne samenlevingen. En juist daarom staat eerlijkheid bij al deze denkers centraal, of ze nu statistici, onderwijskundigen of theoriecritici zijn.

Wanneer men hun werk bestudeert, wordt duidelijk dat bijna alle hedendaagse discussies over toetsing al eens door deze wegbereiders zijn voorzien. Digitale adaptieve toetsen? De noodzaak van equating en normering is alleen maar groter geworden. Brennan en Cronbach hebben al laten zien waarom betrouwbaarheid daar complexer en tegelijk belangrijker is. Selectiekritiek en toetsdruk? Cronbachs waarschuwingen voor contextloze interpretaties blijven relevant. Programmatisch toetsen en toetsmatrijzen? Ebel heeft er de pedagogische grondslagen voor gelegd.

De bruggen die de twintig denkers in dit boek bouwen, zijn niet af en zullen misschien nooit voltooid zijn. Dat is niet erg; het is zelfs passend. Want toetsing is geen gesloten systeem maar een voortdurend proces van herzien, verfijnen en opnieuw nadenken. Elke generatie krijgt nieuwe onderwijsdoelen, nieuwe technologieën, nieuwe maatschappelijke verwachtingen. En elke generatie zal dus opnieuw moeten nadenken over validiteit, betrouwbaarheid en eerlijkheid.

Soms werken de bruggenbouwers samen met anderen, soms individueel. Wat hen onderscheidt van anderen in het vakgebied is dat ze ieder de expert zijn van een deeldiscipline in de toetsleer. Bijvoorbeeld, denk je aan generaliseerbaarheidstheorie, dan denk je aan Cronbach, denk je aan de itemresponstheorie, dan denk je aan Lord en Novick, denk je aan computer-adaptieve toetsen, dan denk je aan Van der Linden en Mislevy. Het is echter opvallend dat geen van de genoemden kan worden gezien als dé vader van een bepaald onderwerp. Het is typisch aan wetenschappelijke onderzoekers dat ze op een bestaand concept voortbouwen en dat vervolgens uitdiepen zodat ze een expert op dat gebied worden. Zo is de gedachte dat een score op een toetsvraag de optelsom is van de ware score en een foutscore al gesuggereerd door Spearman, maar verder uitgewerkt door anderen, op de eerste plaats door Lord.

De experts hebben het vermogen, en misschien wel het lef, om een bepaald onderwerp tot het hunne te maken en dat consequent uit te werken tot een hanteerbaar concept. Niet alleen hebben zij het onderwerp theoretisch verdiept, zonder uitzondering hebben zij aandacht gegeven aan de consequenties voor de



toetspraktijk. Ze zijn weliswaar begonnen als theoretici, maar vervolgens kwam de praktijk. Dat is ook de correcte volgorde. Ook zij kennen de slogan: *niets is zo praktisch als een goede theorie*. Dus op de eerste plaats is er een doortimmerde en robuuste theorie en vandaaruit is het pas mogelijk deze in de praktijk vorm te geven. Helaas zijn er talloze voorbeelden te geven van toetsinnovaties die op grond van een leuk idee of een goed verhaal, breeduit zijn geïmplementeerd in een onderwijssetting, en vervolgens faliekant zijn mislukt. De toetsvernieuwers hadden verzuimd te werken vanuit een theorie. Een voorbeeld hiervan is de invoering van competentietoetsing, meestal op hogescholen. In de praktijk bleek dat de invoering te veel haken en ogen met zich meebracht die onoplosbaar bleken. Bij competentietoetsing gaat het om op een holistische wijze het toetsen van drie onderliggende gebieden: kennis, vaardigheden en attitudes. Maar in praktijk ontstond de wens om toch inzicht te willen hebben op het niveau van ieder van de drie onderliggende gebieden. Dus kwamen er drie toetsen die werden gemiddeld. Het hele idee van een holistische wijze van toetsing werd zo verlaten.

De onderzoekers die in dit boek worden beschreven, kunnen met recht worden gekarakteriseerd als expert van onderdelen van de toetsleer en de toetspraktijk. Het is daarom beter te spreken van bruggenbouwer. Want wat ze hebben gedaan, is de concepten, afkomstig uit de testtheorie, toe te passen op de praktijk van het toetsen. Zo ontstond, naast een testtheorie, ook een toetstheorie. Zonder ook maar een van hen, zou die er anders hebben uitgezien en inhoudelijk armer zijn. Deze wetenschappers hebben vanuit de testtheorie de toetstheorie doen ontstaan en vervolgens vanuit de toetstheorie de toetspraktijk. Zij zijn ware bruggenbouwers.

In dit boek staan de portretten van de bruggenbouwers. Het is goed wanneer toetsdeskundigen uit latere generaties kennismaken met degenen die de toetsleer en toetstheorie hebben ontwikkeld. Zo zien ze dat succes niet komt aanwaaien, maar een gevolg is van doorzettingsvermogen, samenwerking en verantwoording afleggen. Met het laatste wordt bedoeld dat wanneer briljante geesten verzuimen hun zienswijze aan het wetenschappelijke forum voor te leggen, zij gedoemd zullen zijn niet te worden opgemerkt.

De verhalende geschiedenis van de psychometrie is niet alleen een geschiedenis van tabellen, formules en methoden. Het is een geschiedenis van mensen die proberen recht te doen aan andere mensen. Het is een geschiedenis van wetenschappers die begrijpen dat een toets nooit louter een technisch product is, maar een morele uitspraak over wie iemand is, wat hij of zij kan, en welke mogelijkheden een samenleving bereid is toe te kennen.

In de hoofdstukken die volgen, worden deze bruggenbouwers in detail beschreven. Hun (werk)levens lopen uiteen. Maar samen vertellen zij één verhaal: het verhaal van een discipline die volwassen wordt, die leert te kijken naar zichzelf, die fouten durft te erkennen en oplossingen durft te zoeken.

Wie dit boek leest, betreedt niet alleen een galerij van historische figuren, maar wandelt door de fundamenteën van een vakgebied dat nog dagelijks vormgeeft aan de levens van miljoenen mensen. Het is een herinnering dat meten nooit simpel is, dat interpretatie altijd zorg vraagt, en dat elke test of toets, hoe klein ook, een daad is met ethische gevolgen. En misschien is dat wel de belangrijkste les van deze bruggenbouwers: dat de wetenschap van het meten pas echt menselijk wordt wanneer zij begrijpt dat achter elk getal een verhaal schuilgaat.





DEEL 1

DE BRUGGENBOUWERS

- W. H. Angoff
- R.L. Brennan
- L.J. Cronbach
- R.L. Ebel
- L. Guttman
- R.K. Hambleton
- M.T. Kane
- F.J. Kelly
- F.M. Lord
- P.E. Meehl
- W.A. Mehrens
- S.J. Messick
- R.J. Mislevy
- G.W. Rasch
- C.E. Spearman
- R.L. Thorndike
- W.J. van der Linden
- A.A. von Davier
- H. Wainer
- D.J. Weiss

1. WILLIAM H. ANGOFF

1919-1993

Samenvatting

William H. Angoff was een psycholoog en psychometricus, vooral bekend door zijn werk bij de Educational Testing Service (ETS) in Princeton. Hij studeerde psychologie en statistiek aan de City College of New York en Harvard University, en promoveerde met een specialisatie in psychometrie aan Purdue University. In 1950 startte hij zijn carrière bij ETS, waar hij tot zijn overlijden bleef werken. Daar droeg hij bij aan de ontwikkeling van gestandaardiseerde toetsen, zoals de SAT en TOEFL. Angoff legde de nadruk op zorgvuldigheid, eerlijkheid en precisie in toetsontwikkeling. Zijn belangrijkste onderzoeksgebieden waren toetsconstructie, score-equivalering, normering en standaardsetting. Hij introduceerde de invloedrijke Angoff-methode, een procedure waarmee deskundigen bepalen waar de cesuur van een examen moet liggen. Deze methode wordt wereldwijd toegepast, onder andere in medische en juridische examens. Verder schreef hij standaardwerken zoals Scales, Norms, and Equivalent Scores (1971), waarin hij liet zien hoe toetsscores betekenisvol en eerlijk vergeleken kunnen worden. Zijn bijdragen versterkten de wetenschappelijke basis van psychometrie en toetsleer. Hij benadrukte steeds dat validiteit en betrouwbaarheid niet alleen technische, maar ook ethische kwesties zijn. Angoff wordt herinnerd als een integere, analytische denker die zowel methodologische innovaties als morele zorgvuldigheid bracht in de toetspraktijk.

Opleiding

Het studiegebied van Angoff was psychologie. Hij begon zijn opleiding aan het City College of New York, maar switchte naar de Harvard University waar hij zijn bachelor en master behaalde. Daarna vervolgde hij zijn studie aan de Purdue University, waar hij promoveerde in de psychologie, met een specialisatie in statistiek en psychometrie.

Loopbaan

Na het afronden van zijn studie startte Angoff in 1950 zijn loopbaan bij ETS in Princeton, New Jersey. Hij begon daar als onderzoeker en hield zich bezig met de ontwikkeling en analyse van gestandaardiseerde toetsen, waaronder de inmiddels wereldberoemde SAT, een gestandaardiseerde toelatingstest voor Amerikaanse universiteiten. Zijn werk richtte zich op psychometrie, toetsbetrouwbaarheid en de vraag hoe scores eerlijk konden worden geïnterpreteerd. In de loop van de jaren vijftig en zestig groeide hij uit tot een senior onderzoeker. Een belangrijk keerpunt volgde in 1976, toen Angoff werd benoemd tot Director of Developmental Research. In die functie kreeg hij de leiding over projecten die gericht waren op de innovatie van toetsmethodieken. Hij combineerde daarbij zijn theoretische kennis met praktische toepassingen die ETS internationaal toonaangevend maakten. Angoff bleef tot aan zijn overlijden in 1993 verbonden aan ETS en publiceerde in die periode meerdere invloedrijke werken.

Onderscheidingen en prijzen

In openbare bronnen worden geen onderscheidingen of prijzen genoemd die aan Angoff zijn uitgereikt. Wel heeft het ETS na zijn overlijden de William H. Angoff Memorial Lecture Series in het leven geroepen, een jaarlijks evenement waar vooraanstaande onderzoekers worden uitgenodigd om over toetsontwikkeling en psychometrie te spreken. Dit geldt als een belangrijke postume eer en onderstreept de impact van zijn werk.



Leermeesters

Angoff werd gevormd door een aantal leermeesters die zijn academische en professionele ontwikkeling hebben beïnvloed. Tijdens zijn studie in New York kwam hij in aanraking met vooraanstaande psychologen zoals Allport, bekend om zijn werk op het gebied van persoonlijkheidspsychologie, en Boring, die de geschiedenis van de psychologie doceerde. Deze omgeving legde bij Angoff de basis voor een brede psychologische blik, maar ook voor respect voor de wetenschappelijke traditie van zijn discipline. Toen hij later aan de Purdue University promoveerde, werd hij begeleid door Odell, een psycholoog die gespecialiseerd was in psychometrie en statistiek. Onder diens begeleiding verdiepte Angoff zich in de kwantitatieve kant van de psychologie en raakte hij overtuigd van de waarde van nauwkeurige metingen in onderwijs en selectie. Zijn promotieonderzoek bracht hem niet alleen methodologische expertise, maar ook een kritische houding tegenover de praktische toepassing van toetsinstrumenten.

Toen hij in 1950 zijn loopbaan begon bij ETS, kreeg Angoff nieuwe mentoren. Lord, een van de grondleggers van de itemresponstheorie (IRT), als een van zijn meest directe partners. Lord en Angoff hielden zich bezig met de vraag hoe toetsen niet alleen betrouwbaarder maar ook eerlijker konden worden gemaakt. Waar Lord zich vaak richtte op de theoretische verfijning van meetmodellen, was Angoff degene die deze inzichten hielp vertalen naar de praktijk van de SAT en andere examens. Een andere belangrijke leermeester, hoewel jonger, was Messick die beroemd zou worden om zijn werk over validiteit. Messick en Angoff deelden de overtuiging dat toetsen nooit alleen technische instrumenten konden zijn, maar altijd in relatie moesten staan tot maatschappelijke rechtvaardigheid. Hun discussies maakten Angoff gevoeliger voor de ethische dimensies van psychometrie: meten moest niet alleen correct gebeuren, maar ook eerlijk en betekenisvol.

Door deze opeenvolgende leermeesters ontwikkelde Angoff zich tot een veelzijdige toetsdeskundige. Van Allport en Boring nam hij de brede psychologische basis mee, van Odell de statistische precisie, en van Lord en Messick de overtuiging dat toetsing niet alleen een wiskundige, maar ook een menselijke aangelegenheid is. Deze invloeden samen maakten hem tot een bruggenbouwer tussen theorie en praktijk, iemand die de fundamenteën legde voor de huidige toetstheorie én tegelijkertijd het belang van rechtvaardigheid en zorgvuldigheid nooit uit het oog verloor.

Onderzoeksgebieden

Toetsontwikkeling en toetsanalyse

Toetsontwikkeling en toetsanalyse vormen de kern van het werk van Angoff. Toetsontwikkeling is het proces waarbij toetsen zorgvuldig worden ontworpen om eigenschappen, kennis of vaardigheden te meten. Toetsanalyse houdt zich bezig met de evaluatie van die toetsen: of de vragen goed functioneren, of ze de juiste moeilijkheidsgraad hebben en of ze daadwerkelijk meten wat ze beogen te meten. Het gaat hierbij niet alleen om het bedenken van vragen, maar om een systematische benadering waarbij elke vraag wordt geanalyseerd op betrouwbaarheid en validiteit. Dit onderzoeksgebied is van groot belang omdat gestandaardiseerde toetsen vaak beslissend zijn voor de toegang tot onderwijs of beroep. Een examen dat niet goed is ontwikkeld of geanalyseerd, kan leiden tot oneerlijke uitkomsten, discriminatie of verkeerde beslissingen over iemands geschiktheid. Angoff zag dat en werkte aan methoden om tests zo objectief en betrouwbaar mogelijk te maken. Zijn invloed is bijvoorbeeld zichtbaar in de ontwikkeling van de reeds genoemde SAT, waarbij hij ervoor zorgde dat de vragen niet alleen de juiste moeilijkheid hadden, maar ook eerlijk en vergelijkbaar waren tussen kandidaten.

Het belang van Angoffs werk lag in de combinatie van theorie en praktijk. Hij ontwikkelde methoden om systematisch te analyseren welke vragen te makkelijk of te moeilijk waren, welke onvoldoende onderscheid maakten tussen sterke en zwakke kandidaten, en hoe een toets kon worden verbeterd zonder aan



betrouwbaarheid te verliezen. Daarmee loste hij een praktisch probleem op dat tot op heden relevant is: hoe maak je een toets die niet alleen efficiënt is af te nemen, maar ook rechtvaardig beslissingen ondersteunt? Voorbeelden daarvan zijn de SAT in de Verenigde Staten en vergelijkbare examens in Europa die volgens dezelfde principes worden geanalyseerd.

Later is dit werk voortgezet door psychometrici als Embretson en Mislevy, die geavanceerde methoden ontwikkelen voor toetsconstructie. Ook organisaties als de ETS in de VS en het Cito in Nederland bouwen voort op Angoffs nalatenschap door toetsontwikkeling te combineren met digitale en adaptieve technologieën. Wat Angoff begon, wordt nog altijd doorgezet: toetsen worden niet langer intuïtief ontworpen, maar zijn de uitkomst van wetenschappelijk onderzoek en continue evaluatie.

In recente literatuur wordt Angoffs aanpak opnieuw relevant door digitale en adaptieve toetsing. Onderzoeken tonen aan dat standaardsetting-methoden zoals *Modified Angoff* goed toepasbaar zijn in digitale *high-stakes* examens, inclusief medische licentiëring. De focus ligt op betrouwbaarheid van expertschattingen, consensusvorming en hybride methoden waarin Angoff-schattingen worden gecombineerd met empirische data.

Test equating(score-equivalering)

Test equating, ook wel score-equivalering genoemd, was ook een van de gebieden waaraan Angoff een grote bijdrage heeft geleverd. Het verwijst naar het proces waarbij scores van verschillende versies van een examen gelijkwaardig worden gemaakt. Omdat grote examens zoals de SAT of de TOEFL jaarlijks of zelfs vaker in nieuwe versies verschijnen, is het onvermijdelijk dat de moeilijkheidsgraad per versie enigszins verschilt. Test equating zorgt ervoor dat aan een score op versie A dezelfde betekenis kan worden toegekend als aan een gelijke score op versie B. Zonder deze procedure zouden kandidaten ongelijke kansen hebben, simpelweg omdat zij een moeilijker of makkelijker examen kregen. De essentie van equating is dat de prestatie van kandidaten altijd eerlijk moet worden geïnterpreteerd, ongeacht welke toetsvorm zij maken. Angoff benadrukte dat zonder een betrouwbaar systeem van score-equivalering het vertrouwen in examens snel verloren zou gaan. Dit onderwerp is dus van groot maatschappelijk belang, omdat toelating tot universiteiten, toegang tot beroepen en soms zelfs immigratiebeslissingen afhangen van eerlijke toetsresultaten. Zijn onderzoek gaf inzicht in hoe statistische methoden konden worden ingezet om verschillen in moeilijkheidsgraad te neutraliseren. Zijn activiteiten rond dit onderwerp staan beschreven in het hoofdstuk *Scales, Norms, and Equivalent Scores* (1971), dat tot op heden als een standaardwerk wordt gezien. Daarin gaf Angoff richtlijnen om ruwe scores om te zetten in schaal- en normtabellen, waardoor resultaten over de tijd en over verschillende populaties konden worden vergeleken. Een concreet voorbeeld is de SAT, waar een score van 600 in 1975 dezelfde betekenis moest hebben als een score van 600 in 1985. Zonder Angoffs methoden zou het onmogelijk zijn geweest die consistentie te garanderen.

Het belang van dit onderzoeksgebied strekt zich tot ver buiten de VS. Ook in Nederland en andere landen worden staatsexamens en beroepskwalificaties jaarlijks opnieuw afgenomen en moeten scores gelijkwaardig zijn. Organisaties zoals Cito gebruiken vergelijkbare methodieken om prestaties over verschillende jaren met elkaar te vergelijken.

Later is dit werk voortgezet door onderzoekers als Kolen en Brennan, die het standaardwerk *Test Equating, Scaling, and Linking* schreven (Kolen & Brennan, 2024). Zij bouwden direct voort op Angoffs ideeën en pasten deze toe in steeds complexere toetsomgevingen, waaronder digitale en adaptieve toetsen. Zijn pionierswerk blijft dus een fundament waarop latere generaties hun onderzoek en praktijk baseren.



Normering en standaardsetting

Normering en standaardsetting zijn nauw verwante concepten die centraal stonden in het werk van Angoff. Normering betekent dat individuele scores worden geplaatst binnen een referentiekader, bijvoorbeeld door ze te vergelijken met een normgroep of om te zetten in percentielen. Standaardsetting verwijst naar het vaststellen van een drempel of cesuur die aangeeft wie geslaagd of niet geslaagd is. Samen geven deze processen betekenis aan toetsresultaten: ze maken duidelijk of een bepaalde score voldoende is en hoe die zich verhoudt tot de scores van andere kandidaten.

Het belang van dit onderzoeksgebied kan moeilijk worden overschat. Zonder normering en standaardsetting zouden toetsscores losstaande cijfers blijven zonder context. Een score van 75 betekent weinig als niet duidelijk is of dit boven of onder het (groeps)gemiddelde ligt, en of dit voldoende is om door te stromen naar een opleiding of beroep. Angoff zag dat een betrouwbaar en transparant systeem nodig was om beslissingen te kunnen rechtvaardigen. Vooral in onderwijs en beroepscertificering, waar de uitkomst van een examen directe invloed heeft op iemands toekomst, is dit een cruciaal vraagstuk.

De betekenis van Angoffs onderzoek blijkt vooral uit de door hem ontwikkelde en naar hem genoemde methode: de Angoff-methode. Deze procedure vraagt experts om per toetsvraag in te schatten hoeveel kandidaten met minimale competentie de vraag correct zouden kunnen beantwoorden. De som van die verwachtingen vormt de cesuur. Het grote voordeel van deze methode is dat niet willekeurig of achteraf wordt bepaald wie slaagt of faalt, maar dat er een systematisch, reproduceerbaar en wetenschappelijk onderbouwd proces aan ten grondslag ligt. Een concreet voorbeeld van de toepassing is het Amerikaanse *bar exam*, een examen voor juristen om advocaat te mogen worden, dat in veel staten gebruikmaakt van (varianten van) de Angoff-methode. Ook medische licentietoetsen wereldwijd baseren hun cesuur op dit principe.

Angoffs bijdrage loste daarmee een fundamenteel probleem op: hoe bepaal je objectief wat voldoende bekwaam is? Voorheen waren cesuren vaak gebaseerd op historische gemiddelden of op arbitraire beslissingen van examinatoren. Met de Angoff-methode werd het proces transparanter en beter verdedigbaar. Dat zorgde voor meer vertrouwen in examens en voor eerlijker beslissingen.

In het heden wordt dit werk voortgezet en verfijnd. Onderzoekers als Kane en Glass hebben de methode uitgebreid en aangepast aan nieuwe omstandigheden, zoals digitale en adaptieve toetsen. Ook internationale organisaties, waaronder de World Health Organization, maken gebruik van Angoff-varianten om standaarden te zetten in certificeringssystemen. Daarmee blijft zijn naam direct verbonden met een van de meest invloedrijke en toegepaste methoden in de toetsleer.

Validiteit en betrouwbaarheid

Validiteit en betrouwbaarheid vormen de pijlers van elke toets en waren een van de centrale thema's in het werk van Angoff. Validiteit verwijst naar de vraag of een toets daadwerkelijk meet wat zij geacht wordt te meten. Betrouwbaarheid heeft betrekking op de consistentie van de resultaten: bij herhaalde afname moet een toets hetzelfde resultaat opleveren. Een toets kan betrouwbaar zijn zonder valide te zijn, bijvoorbeeld een rekentoets die vooral taalvaardigheid meet omdat de opgaven te talig zijn, maar een valide toets moet altijd ook betrouwbaar zijn.

Het belang van dit onderzoeksgebied ligt voor de hand. Examens zijn vaak bepalend voor iemands toegang tot een opleiding, beroep of maatschappelijke positie. Wanneer een toets onbetrouwbaar of niet valide is, kunnen studenten onterecht slagen of zakken, met verstreckende gevolgen. Angoff zag dat validiteit en betrouwbaarheid daarom niet alleen technische kwesties zijn, maar ook ethische: ze gaan over rechtvaardigheid en gelijke kansen. Zijn werk legde nadruk op de noodzaak om deze begrippen niet



te behandelen als abstracte statistische formules, maar als concrete kwaliteitscriteria die richting geven aan toetsontwikkeling en -analyse. Angoffs bijdrage was dat hij validiteit en betrouwbaarheid niet los van elkaar behandelde, maar integreerde in de bredere context van toetsconstructie. Hij werkte aan methoden om de betrouwbaarheid van toetsen te vergroten, bijvoorbeeld door te onderzoeken hoe de lengte van een toets de betrouwbaarheid beïnvloedt. In zijn artikel uit 1953, waarin hij de relatie tussen testlengte en betrouwbaarheid analyseerde, staat dat beschreven. Daarmee gaf hij toetsontwikkelaars praktische handvatten om te bepalen hoeveel vragen een toets minimaal moet bevatten. Tegelijkertijd benadrukte hij dat validiteit het uiteindelijke doel was: een toets moest niet alleen consistent meten, maar ook precies datgene meten wat relevant is voor de beslissingen die erop volgen. Een concreet voorbeeld is zijn werk aan de SAT. Hij leverde een bijdrage aan de validering van de toets door te onderzoeken of de toets daadwerkelijk academisch succes voorspelde. Hij leverde ook bijdragen aan betrouwbaarheidsanalyses van andere grootschalige toetsen, zoals de TOEFL. Door deze studies werd duidelijk dat toetsen alleen maatschappelijk betekenisvol kunnen zijn als zij voldoen aan deze strenge criteria.

Later is dit onderzoeksgebied voortgezet door psychometrici als Kane en Mislevy die werken aan moderne validiteitsmodellen en aan de argument-gebaseerde benadering. Ook Messick, een collega van Angoff bij ETS, bouwde direct voort op deze basis en ontwikkelde een invloedrijk raamwerk voor validiteit dat breed is overgenomen. Hun werk bevestigt dat de onderwerpen die Angoff aansneed, nog steeds actueel zijn. Validiteit en betrouwbaarheid zijn niet langer alleen onderzoeksbegrippen, maar bepalende voorwaarden voor elk examen dat wordt ingezet.

Statistische en psychometrische methodologie

Een van de meest fundamentele onderzoeksgebieden waarin Angoff actief was, betrof de ontwikkeling en toepassing van statistische en psychometrische methoden. Psychometrische methodologie kan worden omschreven als het geheel van wiskundige en statistische technieken die worden gebruikt om eigenschappen en prestaties van personen op een betrouwbare en valide manier te meten. Het omvat onderwerpen als schaalconstructie, itemanalyse, factoranalyse en later ook de meer geavanceerde modellen zoals de itemresponstheorie. Angoff werkte in een periode waarin psychometrie zich sterk ontwikkelde, en hij stond aan de basis van het omzetten van abstracte theorieën naar praktische instrumenten. Het belang van dit onderzoeksgebied is groot, omdat zonder statistische methoden toetsresultaten betekenisloos zouden blijven. Een ruwe score zegt op zichzelf weinig; pas door middel van psychometrische berekeningen kan men bepalen of die score eerlijk, vergelijkbaar en interpreteerbaar is. Bovendien maken deze methoden het mogelijk om toetsen jaar na jaar te verbeteren, bias te signaleren en scores van verschillende versies gelijk te stellen. Angoff zag dit en pleitte ervoor om toetsontwikkeling altijd te verankeren in solide wiskundige en methodologische principes. Daarmee professionaliseerde hij de toetsleer en verhef hij het vak tot een academische discipline.

Zijn bijdrage lag vooral in het toegankelijk maken van deze complexe methoden voor de praktijk. Hij schreef overzichtswerken die duidelijk maakten hoe schaalmethoden en normtabellen konden worden toegepast. Zijn reeds genoemde hoofdstuk *Scales, Norms, and Equivalent Scores* uit 1971 was daarin bijzonder invloedrijk omdat het een brug sloeg tussen theorie en praktijk. Hij maakte inzichtelijk hoe percentielen moesten worden berekend en geïnterpreteerd, hoe je equivalente scores tussen verschillende testvormen kon vaststellen, en hoe schaalconstructies konden bijdragen aan eerlijke beslissingen. Dit loste het praktische probleem op dat veel onderwijsinstellingen en beroepsorganisaties hadden: hoe vergelijk je scores die op het eerste gezicht niet gelijkwaardig lijken? Een concreet voorbeeld van de toepassing is de manier waarop de SAT en TOEFL zijn opgebouwd. Dankzij psychometrische methoden zoals die van Angoff, kunnen scores van verschillende jaargangen of versies eerlijk worden vergeleken. Ook bij licentietoetsen voor artsen, juristen of leraren zijn deze methoden onmisbaar gebleken.



Later zetten onderzoekers zoals Reckase, Wainer en Van der Linden het werk voort. Zij hebben de methoden verfijnd met behulp van moderne statistiek, zoals Bayesian models en computerized adaptive testing (hoofdstuk 24). Hoewel de technieken complexer zijn geworden, blijft de basis dezelfde: toetsen moeten methodologisch onderbouwd zijn om betrouwbaar en eerlijk te zijn. Angoffs pionierswerk leverde het fundament waarop deze latere generaties konden bouwen, en zijn invloed blijft zichtbaar in vrijwel elk grootschalig examen dat wereldwijd wordt afgenomen.

Invloed

De invloed van Angoff op de toetsleer is groot. Hij maakte van toetsontwikkeling een vakgebied dat niet langer alleen draaide om het construeren van vragen, maar dat stevig werd verankerd in psychometrische theorie. Vóór zijn tijd werd toetsing vaak gezien als een praktisch ambacht: docenten en instellingen stelden examens samen en interpreteerden resultaten op basis van ervaring of intuïtie. Angoff bracht daar verandering in door methoden te ontwikkelen waarmee toetsen betrouwbaarder, eerlijker en transparanter konden worden geconstrueerd en afgenomen.

Zijn werk op het gebied van equating en schaalconstructie zorgde ervoor dat toetsscores door de tijd heen vergelijkbaar werden. Daarmee werd het mogelijk om prestaties van verschillende cohorten of versies van een toets eerlijk naast elkaar te leggen, iets wat essentieel is in zowel nationaal als internationaal onderwijs. Daarnaast introduceerde hij met de Angoff-methode een manier om cesuren te bepalen die niet willekeurig of politiek bepaald waren, maar gebaseerd op wetenschappelijke inschattingen van deskundigen. Deze procedure wordt tot op de dag van vandaag wereldwijd gebruikt in onderwijs en beroepscertificering.

Een ander aspect van zijn invloed was de nadruk op validiteit en betrouwbaarheid als fundamenteën van toetsing. Hij maakte duidelijk dat toetsen alleen waardevol zijn als ze meten wat ze beogen te meten en dat de resultaten consistent moeten zijn. Daarmee legde hij de basis voor kwaliteitsstandaarden die tegenwoordig vanzelfsprekend zijn. Zijn publicaties en zijn rol bij ETS droegen ertoe bij dat psychometrie en toetsconstructie zich ontwikkelden tot een academische discipline. Organisaties als Cito in Nederland en vergelijkbare instituten wereldwijd werken nog altijd volgens principes die teruggaan op zijn werk. Daarmee kan worden gesteld dat Angoff de toetsleer niet alleen heeft beïnvloed, maar mede heeft vormgegeven tot het wetenschappelijk verantwoorde vakgebied dat het nu is.

Belangrijkste publicaties

- Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika*, 18(1), 1–14.
In dit artikel onderzoekt Angoff de relatie tussen de lengte van een toets en de betrouwbaarheid van de resultaten. Hij bouwt voort op de Spearman-Brown-formule en toont aan dat langere toetsen doorgaans betrouwbaarder zijn, maar dat er een punt van afnemende meeropbrengst optreedt. Het artikel levert toetsontwikkelaars een praktische basis voor het bepalen van de optimale testlengte.
- Angoff, W. H. (1971b). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.
Dit hoofdstuk biedt een systematisch overzicht van schaalconstructie, normering en score-equivalering. Angoff legt uit hoe ruwe scores betekenis krijgen door omzetting naar normtabellen en equivalente scores, en hoe deze vergelijkingen noodzakelijk zijn om prestaties over verschillende populaties en jaargangen eerlijk te interpreteren. Het hoofdstuk geldt als een hoeksteen in de toetsleer.



- Angoff, W. H. (1972). *Patterns of test and item difficulty for six foreign language groups on the Test of English as a Foreign Language*. Educational Testing Service.
In deze studie analyseert Angoff de prestaties van zes verschillende taalgroepen op de TOEFL. Hij laat zien dat de moeilijkheidsgraad van items kan variëren per taalachtergrond, wat inzicht gaf in differential item functioning (DIF). Dit onderzoek was een belangrijke stap richting eerlijkere toetsen en beïnvloedde de ontwikkeling van internationale taalexamens.
- Angoff, W. H., & Schrader, W. B. (1981). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educational Measurement*, 18(4), 229–244.
In dit artikel onderzoekt Angoff samen met Schrader het verschil tussen goed-scoring (alleen goede antwoorden tellen) en formule-scoring (puntenaftrek bij foute antwoorden). Ze analyseren de effecten van gokgedrag en rechtvaardigheid. De studie levert belangrijke inzichten voor de scoring van meerkeuze-examens, waaronder de SAT.
- Assess.com. (2021). *The modified Angoff method: A practical guide*. Assess.
Deze uitgave beschrijft hoe de Angoff-methode kan worden toegepast binnen digitale en moderne toetsplatforms. De publicatie geeft richtlijnen voor panelectie, training van experts, en het bereiken van consensus tussen beoordelaars. Nieuw is dat het document laat zien hoe statistische feedback (bijvoorbeeld item-p-waarden na pilotafnames) kan worden gebruikt om schattingen te verfijnen zonder dat de methode haar inhoudelijke basis verliest. De auteurs bespreken ook varianten zoals Borderline Regression en Modified Angoff, waarin experts hun schattingen iteratief bijstellen.

Bronnen

- Assess.com. (2021). *The modified Angoff method: A practical guide*. Assess. <https://assess.com/modified-angoff-method/>
- Cirrus Assessment. (2022). *The pros and cons of the Angoff method for setting standards*. Cirrus Assessment. <https://cirrusassessment.com/pros-and-cons-of-the-angoff-method-for-setting-standards/>
- Educational Testing Service. (z.d.). *William H. Angoff (1919–1993): A tribute*. ETS. <https://www.ets.org/Media/Research/pdf/PICANG13.pdf>
- Sage Publishing. (2018). The Angoff method. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vol. 1, pp. 108–110). Sage. <https://doi.org/10.4135/9781506326139.n47>
- Wikipedia contributors. (2023, December). *William H. Angoff*. In Wikipedia, https://en.wikipedia.org/wiki/William_H._Angoff



2. ROBERT L. BRENNAN

1944-2018

Samenvatting

Robert L. Brennan bouwde een indrukwekkende loopbaan op binnen de psychometrie en toetsing. Na zijn promotie aan de Harvard University werkte hij eerst als docent en onderzoeker, waarna hij achttien jaar bij ACT werkte. In 1994 werd hij hoogleraar in Iowa en leidde hij de Iowa Testing Programs. Hij richtte later CASMA op en bleef tot na zijn emeritaat invloedrijk in het vakgebied. Brennan stond bekend om zijn werk aan de generaliseerbaarheidstheorie, waarmee hij de betrouwbaarheid van toetsresultaten juist wist te analyseren, en om zijn bijdragen aan het gelijkstellen en schalen van toetsen. Ook onderzocht hij de betrouwbaarheid van beoordelaars bij performance assessments. Zijn werk bracht de toetsleer dichter bij de onderwijspraktijk en maakte toetsen eerlijker en consistent. Naast onderzoek bekleedde hij belangrijke bestuursfuncties en redactionele rollen. Zijn bijdragen leverden hem prestigieuze prijzen op en zijn invloed blijft wereldwijd zichtbaar in toetsontwikkeling en -onderzoek.

Loopbaan

Na het behalen van zijn PhD aan de Harvard Graduate School of Education in 1970, begon Brennan zijn academische loopbaan aan de SUNY Stony Brook, waar hij les gaf en onderzoek deed in psychometrie en toetsing. In 1976 maakte hij de overstap naar de praktijk: hij trad in dienst bij ACT (American College Testing). Daar bekleedde hij diverse functies gedurende achttien jaar, onder andere als Assistant Vice-President for Measurement Research en als Distinguished Research Scientist. In 1994 verhuisde Brennan naar de University of Iowa, waar hij de prestigieuze E. F. Lindquist Chair in Measurement and Testing aanvaardde en werd aangesteld als directeur van de Iowa Testing Programs. In 2002 trad hij terug uit de leiding van Iowa Testing Programs om zich volledig te richten op onderzoek en onderwijs, waarbij hij het Center for Advanced Studies in Measurement and Assessment (CASMA) oprichtte. Uiteindelijk ging Brennan met emeritaat, maar hij is niet volledig uit het vakgebied verdwenen: hij blijft betrokken als adviseur bij toetsorganisaties en onderzoekscentra, en zijn standaardwerken worden wereldwijd aangehaald in handleidingen voor toetsontwikkeling en test equating.

Nadat Brennan zijn academische loopbaan had verankerd, ging hij zich ook steeds nadrukkelijker inzetten voor de bredere toetsgemeenschap. Hij nam niet alleen plaats in redacties, maar werd ook hoofdredacteur van invloedrijke publicaties, waaronder de vierde editie van *Educational Measurement*, een standaardwerk binnen het vakgebied uit 2006. Zijn vermogen om complexe vraagstukken helder te structureren maakten hem tot een veelgevraagde redacteur en adviseur bij toonaangevende publicaties op het gebied van psychometrie en toetsing.

Naast zijn werk in de wetenschap vervulde Brennan ook tal van bestuursfuncties. Zo werd hij gekozen tot president van de National Council on Measurement in Education (NCME), waar hij richting gaf aan de ontwikkeling van toetsmethodologie in de Verenigde Staten. Binnen de American Educational Research Association (AERA) bekleedde hij de rol van vicepresident van Division D, de divisie die zich richt op methoden en meting. Ook dichterbij huis droeg hij verantwoordelijkheid: hij was president van de Mid-Western Educational Research Association en later ook van de Iowa Academy of Education. In al deze rollen wist Brennan zijn expertise te verbinden met leiderschap, waardoor hij niet alleen zijn eigen onderzoek aanscherpte, maar ook de koers van het bredere vakgebied medebepaalde.



Onderscheidingen en prijzen

Brennan ontving in de loop der jaren meerdere onderscheidingen. In 1997 kreeg hij de NCME Award for Outstanding Technical Contribution to the Field of Educational Measurement. Daarna, in 2000, ontving hij een andere award van de NCME: de Award for Career Contributions to Educational Measurement als erkenning voor zijn impact op het vakgebied. In 2004 ontving hij de E. F. Lindquist Award (gegeven door AERA en ACT) voor zijn onderzoek op het gebied van toetsing en meting. In 2011 volgde een andere award, namelijk de Career Achievement Award van de Association of Test Publishers, waarmee zijn gehele loopbaan in assessment werd geëerd. In 2019 werd hij bovendien vereerd met de Albert Nelson Marquis Lifetime Achievement Award, een erkenning vanuit het Who's Who-netwerk voor personen met langdurige en indrukwekkende bijdragen aan hun vakgebied.

Onderzoeksgebieden

Generaliseerbaarheidstheorie

Een belangrijk onderzoeksgebied van Brennan is de zogenoemde generaliseerbaarheidstheorie (G-theorie). Hij was niet de onderzoeker die met het idee kwam van de G-theorie. Dat was Cronbach met zijn collega's (Cronbach et al., 1963 en 1972). De publicaties van Cronbach et al. waren echter voor de docenten die werkten in de praktijk zeer moeilijk te doorgronden en het is de verdienste van Brennan geweest dat hij de theorie in een begrijpelijke taal wist uit te leggen. Deze theorie kan het best worden omschreven als een uitbreiding van de klassieke testtheorie, waarbij de betrouwbaarheid van toetsresultaten niet wordt samengevat in één enkele coëfficiënt, maar wordt opgesplitst in verschillende bronnen van de meetfout. De G-theorie gaat uit van het idee dat elke meting wordt beïnvloed door meerdere facetten, zoals de items in een toets, de beoordelaars die scores toekennen of de situatie waarbinnen een test of toets wordt afgenomen. In plaats van slechts de verhouding tussen ware score en meetfout te onderzoeken, analyseert G-theorie de variantie die door deze facetten wordt veroorzaakt. Daarmee ontstaat een veel genuanceerder beeld van de betrouwbaarheid van toetsresultaten. Het onderwerp is belangrijk omdat beslissingen in onderwijs en selectie vaak worden gebaseerd op toetsuitslagen. Als men alleen een globaal betrouwbaarheidsgetal gebruikt, worden mogelijke foutbronnen over het hoofd gezien. Door de benadering van Brennan en zijn collega's werd het mogelijk om beslissingen over toetsconstructie en scoringsprocedures beter te onderbouwen. Zijn onderzoek liet zien hoe de G-theorie bijvoorbeeld kan analyseren of het toevoegen van meer items, het inzetten van meerdere beoordelaars of het variëren van opdrachten de betrouwbaarheid verhoogt. Een concreet voorbeeld is het gebruik van schrijfp opdrachten in taaltoetsen: de variantie kan dan worden uitgesplitst in effecten van schrijftaak, beoordelaar en student, waardoor zichtbaar wordt welke component het meest bijdraagt aan (on)betrouwbaarheid. Het werk van Brennan leverde praktische modellen en software waarmee toetsontwikkelaars analyses kunnen uitvoeren en scenario's kunnen doorrekenen. Daarmee werden belangrijke problemen opgelost, zoals de vraag hoeveel taken of beoordelaars nodig zijn voor betrouwbare metingen. Na hem zetten onderzoekers als Kane, Clauser en collega's binnen onderwijs- en licentiëringsorganisaties de traditie van de leer van de generaliseerbaarheidstheorie voort. Ook in Europa, bijvoorbeeld bij het Cito in Nederland, of bij de ETS in de Verenigde Staten, zijn er groepen die zijn werk verder hebben uitgebouwd, zodat G-theorie inmiddels een standaardinstrument is in toetsontwikkeling en onderzoek.

Equating and Scaling of Educational Tests

Een tweede belangrijk onderzoeksgebied van Brennan betreft het gelijkstellen en schalen van toetsen, beter bekend als *test equating* en *scaling*. *Equating* wordt gedefinieerd als de procedure om scores van verschillende versies van een toets zodanig te transformeren dat ze onderling vergelijkbaar worden. *Scaling*, schalen, gaat een stap verder en plaatst toetsresultaten op een gemeenschappelijke meetlat, vaak om de voortgang over jaren heen te kunnen volgen. Deze technieken houden in dat er statistische



methoden worden toegepast om te corrigeren voor verschillen in moeilijkheidsgraad tussen toetsvormen, zodat kandidaten die verschillende varianten maken toch op eerlijke wijze met elkaar kunnen worden vergeleken. Het onderwerp is cruciaal, omdat grootschalige toetsprogramma's, zoals toelatingstoetsen of eindtoetsen, nooit voor alle kandidaten exact dezelfde opgaven kunnen gebruiken. Zonder equating zouden kleine verschillen in toetsconstructie kunnen leiden tot ongelijkheid in kansen. Brennans onderzoek was belangrijk omdat hij methoden ontwikkelde en systematiseerde waarmee equating en scaling betrouwbaar en transparant konden worden uitgevoerd. Zijn werk maakte duidelijk welke assumpties bij verschillende methoden gelden en hoe men keuzes kan maken tussen bijvoorbeeld equating op basis van gemeenschappelijke items of op basis van parallelle groepen. Een praktisch voorbeeld is de ACT-toets: elke afname bevatte nieuwe items, maar door equating bleven scores door de tijd heen vergelijkbaar. Ook in onderwijsonderzoek in Iowa werd scaling toegepast om prestaties van leerlingen over cohorten te volgen. Brennans bijdragen losten hiermee problemen op met betrekking tot eerlijkheid en vergelijkbaarheid, waardoor toetsprogramma's betrouwbaar konden functioneren. Heden ten dage houden onderzoekers zoals Dorans, Hong Jiao en Kolen (die nauw met Brennan samenwerkte) zich intensief bezig met equating en scaling. Organisaties als ETS, ACT en nationale toetscentra wereldwijd bouwen nog steeds voort op zijn modellen en toepassingen, waarmee zijn invloed in dit domein hoog is gebleven.

Performance Assessment en de betrouwbaarheid van beoordelaars

Een derde onderzoeksgebied waarin Brennan veel betekende, is dat van *performance assessment* en de rol van beoordelaars bij het vaststellen van prestaties. Performance assessment verwijst naar toetsvormen waarin studenten of kandidaten complexe taken uitvoeren die dichterbij de werkelijkheid liggen dan meerkeuzevragen, zoals het schrijven van essays, het uitvoeren van wetenschappelijke experimenten of het geven van presentaties. De aantrekkingskracht van deze toetsvormen is dat zij beter in staat zijn hogere-ordevaardigheden te meten, zoals kritisch denken, probleemoplossend vermogen of communicatieve vaardigheden. Het nadeel is echter dat de beoordeling van zulke taken vaak subjectief is en daardoor gevoelig voor inconsistenties en fouten. Brennan onderzocht dit probleem met zijn expertise in psychometrie en betrouwbaarheidstheorie. Hij liet zien dat beoordelaars zelf een belangrijke bron van meetfouten zijn en dat de variantie in scores vaak niet alleen voortkomt uit de kandidaat, maar ook uit verschillen in beoordelingsstijl of interpretatie van rubrics. Met behulp van de generaliseerbaarheidstheorie maakte hij deze bronnen van variantie zichtbaar en kwantificeerbaar. Daardoor konden toetsontwikkelaars beter inschatten hoeveel beoordelaars per taak nodig waren, hoe rubrics moesten worden verfijnd of welke trainingen noodzakelijk waren om de interbeoordelaarsbetrouwbaarheid te verhogen. Een concreet voorbeeld hiervan is te vinden in schrijfvaardigheidstoetsen in het hoger onderwijs. Brennans analyses toonden aan dat een enkele beoordelaar vaak onvoldoende is om tot een stabiele score te komen, waardoor systemen met meerdere beoordelaars of herbeoordelingen noodzakelijk werden.

Het belang van dit onderzoeksgebied is groot. Zonder betrouwbare beoordelingen verliezen performance assessments hun waarde, juist omdat ze zijn bedoeld om genuanceerdere vaardigheden te meten. Brennans werk leverde praktische richtlijnen waarmee instellingen konden zorgen dat zulke toetsen eerlijker en consistentier werden. Daarmee werd een probleem opgelost dat zowel technisch als maatschappelijk urgent was: hoe beoordeel je complexe vaardigheden op een manier die geloofwaardig is voor studenten, docenten en beleidsmakers?

Vandaag de dag zetten onderzoekers zoals Bennett (ETS), Darling-Hammond (Stanford, Learning Policy Institute) en meer toegepaste toetscentra in Europa en Azië dit werk voort. Ook de opkomst van digitale performance assessments met geautomatiseerde scoring bouwt verder op de fundamenteen die Brennan legde in het zichtbaar maken en corrigeren van beoordelaarsfouten. Zijn onderzoek blijft dus richtinggevend in complexe situaties op het grensgebied van het onderwijs en de beroepspraktijk.



Invloed

De invloed van Brennan op de toetsleer is zowel theoretisch als praktisch geweest. Theoretisch heeft hij de klassieke test- en toetstheorie verbreed met zijn werk aan de G-theorie, waarmee het mogelijk werd om toetsresultaten te analyseren op meerdere bronnen van meetfout. Dit bracht een verschuiving teweeg: van een enkel betrouwbaarheidsgetal naar een rijker inzicht in hoe taken, beoordelaars en omstandigheden samen de betrouwbaarheid beïnvloeden. Daarmee gaf hij de toetsleer een fundament dat recht deed aan de complexiteit van onderwijscontexten.

Ook voor de toetspraktijk was zijn invloed groot. Door zijn bijdragen aan equating en scaling werden grootschalige toetsen eerlijker en consistentier gemaakt. Hij legde uit hoe men scores van verschillende toetsvormen vergelijkbaar kan maken, waardoor beleidsmakers en onderwijsinstellingen betrouwbare beslissingen konden nemen over toelating, plaatsing en verantwoording. Zijn werk bood oplossingen voor een van de kernproblemen van de toetsleer: hoe zorg je ervoor dat een score vandaag en een score morgen dezelfde betekenis hebben, ook als de toets niet identiek is?

Daarnaast bracht Brennan met zijn analyses van performance assessments en beoordelaarsbetrouwbaarheid de toetstheorie dichterbij de realiteit van onderwijspraktijken. Hij maakte zichtbaar dat subjectieve beoordelingen niet genegeerd, maar juist moeten worden geanalyseerd en beheerst. Dit versterkte de geloofwaardigheid van toetsen die meer complexe vaardigheden meten, iets wat steeds belangrijker is geworden in moderne curricula.

Zijn invloed strekte zich bovendien uit via leiderschap en redactie: door het samenbrengen en ordenen van kennis in standaardwerken zoals *Educational Measurement* (2006) gaf hij generaties onderzoekers en toetsontwikkelaars een gemeenschappelijk referentiekader. In die zin was zijn betekenis niet alleen wetenschappelijk, maar ook institutioneel en didactisch: hij hielp de toetsleer bij het zich ontwikkelen tot een volwaardig en veelzijdig vakgebied dat methodologische strengheid combineert met onderwijsrelevantie.

Belangrijkste publicaties

- Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3), 687–699.
Brennan bespreekt wanneer Cohens kappa passend is voor beoordelaarsovereenstemming met categorische data, waar het mis kan gaan (bijvoorbeeld bij scheve verdelingen), en welke kappa-achtige alternatieven of correcties zinvoller zijn. Dit artikel is invloedrijk omdat het misinterpretaties corrigeert en praktische houvast biedt voor onderzoek met beoordelaars, bijvoorbeeld bij rubric-scores voor essays of klinische beoordelingen.
- Brennan, R. L. (1992). *Elements of generalizability theory* (Rev. ed.). ACT, Inc.
Deze monografie, een herziene versie van de ACT-uitgave uit 1983, biedt een toegankelijke, stapsgewijze behandeling van G-theorie, inclusief notatie, designs, variantiecomponenten en decision-studies. Het diende jarenlang als praktische gids vóór het latere Springer-boek, en wordt nog steeds veel aangehaald in handleidingen, rapporten en onderzoek waarin G-analyses worden toegepast op 'echte' toetsprogramma's.
- Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, 24(4), 339–353.
Brennan past G-theorie toe op performance assessments (zoals essays of praktische taken) en laat zien hoe beoordelaars, taken en situaties aan de meetfout bijdragen. Hij bespreekt keuzes als aantal taken en aantal beoordelaars en geeft rekenvoorbeelden die helpen robuustere beoordelingen te ontwerpen. Deze bijdrage biedt een methodische basis voor betrouwbaardere performance-toetsen.



- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.
Dit boek biedt het meest complete overzicht van de generaliseerbaarheidstheorie: een raamwerk om meerdere bronnen van meetfout (items, beoordelaars, situaties) te scheiden en ontwerpkeuzes te optimaliseren. Brennan behandelt zowel univariate als multivariate G-analyses, decision-studies en praktische richtlijnen voor toetsontwerp. Het werk wordt breed gebruikt door onderzoekers en toetsontwikkelaars als standaardreferentie en heeft het veld geholpen om betrouwbaarheid genuanceerder en ontwerpgericht te benaderen.
- Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). American Council on Education & Praeger Publishers.
Van dit omvangrijke boek was Brennan redacteur. Het bevat hoofdstukken over validering, betrouwbaarheid, itemresponstheorie, testontwikkeling, toetsingstechnologie en toepassingen. Brennan zelf levert hierin de inleidende hoofdstukken over de evolutie en toekomst van toetsing.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21.
Dit overzichtsartikel zet de G-theorie af tegen de klassieke testtheorie. Brennan laat zien hoe de G-theorie verschillende foutbronnen kwantificeert en waarom dat cruciaal is voor beslissingen in onderwijs en certificering. Het artikel biedt definities, vergelijkingen en voorbeeldtoepassingen en is daardoor een veelgeciteerde instap voor onderzoekers en personen uit de toetspraktijk die betrouwbaarheid willen analyseren met meer dan één enkele coëfficiënt.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
Kolen en Brennan presenteren een uitgebreid en gezaghebbend overzicht van theorieën en methoden voor test equating, scaling en linking. Het boek combineert klassieke en itemresponstheoretische benaderingen met praktische toepassingen in grootschalige toetsprogramma's. Door aandacht voor aannames, ontwerpkeuzes en validiteit laat het zien hoe toetsscores over tijd, populaties en toetsvormen eerlijk en verantwoord vergelijkbaar kunnen worden gemaakt. Het werk geldt internationaal als een standaardreferentie binnen de psychometrie en toetsleer.

Bronnen

- 24-7 Press Release. (2019, juli 19). *Robert L. Brennan presented with the Albert Nelson Marquis Lifetime Achievement Award by Marquis Who's Who*. 24-7 Press Release. <https://www.24-7pressrelease.com/press-release/465944/robert-l-brennan-presented-with-the-albert-nelson-marquis-lifetime-achievement-award-by-marquis-whos-who>
- Education UIowa. (2017, april). *Retirements*. College of Education, University of Iowa. <https://education.uiowa.edu/news/2017/04/retirements>
- New Meridian Corporation. (z.d.). *Dr. Robert L. Brennan*. New Meridian. <https://newmeridiancorp.org/dr-robert-l-brennan>
- ResearchGate. (z.d.). *Robert L. Brennan, Scientific contributions*. ResearchGate. <https://www.researchgate.net/scientific-contributions/Robert-L-Brennan-80498163>
- Tests.com. (z.d.). *Robert L. Brennan*. Tests.com. <https://www.tests.com/robert-brennan>
- University of Iowa, College of Education. (z.d.). *Robert L. Brennan biography* (IAE pages). University of Iowa. <https://www2.education.uiowa.edu/html/iae/Pages/bio-brennan.html>
- Wiley Online Library. (1992). *Generalizability theory. Educational and Psychological Measurement*. Wiley. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.1992.tb00260.x>

