Lee J. Cronbach (1916-2001): meten is begrijpen

Een interview met een bruggenbouwer tussen theorie en praktijk

■ **Henk van Berkel**
Dr. Mult. H. van Berkel is initiatiefnemer, oprichter en gedurende vele jaren hoofdredacteur van Examens, tijdschrift voor de toetspraktijk. Hij is auteur van meerdere boeken over toetskwaliteit. Zijn laatste boek is Mijlpalen in de test- en toetsleer; Van meten naar rechtvaardig beslissen.
E-mail: henkvanberkel22@gmail.com

Het onderstaande interview is fictief. Het is een bewerking van hoofdstuk 3 uit dit boek.

In de wereld van toetsing en psychometrie zijn er namen die niet alleen een tijdperk markeren, maar het vakgebied zelf opnieuw vormgeven. Lee J. Cronbach is zo'n naam. Zijn werk ligt aan de basis van hoe vandaag de dag wordt gedacht over betrouwbaarheid, validiteit en de rol van toetsing in onderwijs en beleid. Maar achter begrippen als Cronbachs alfa en generaliseerbaarheidstheorie schuilt een onderzoeker die voortdurend zocht naar verbinding: tussen statistiek en onderwijs, tussen theorie en praktijk, en tussen meten en begrijpen.
In dit fictieve interview blikt Cronbach terug op zijn leven, zijn werk en zijn ideeën, en wordt zichtbaar hoe zijn denken nog altijd doorwerkt in hedendaagse toetspraktijken.

U wordt vaak gezien als een van de grondleggers van de moderne testtheorie. Welke academische opleidingen heeft u gevolgd?

Ik begon eigenlijk vrij klassiek, met een studie psychologie aan het Fresno State College. Daar ontstond mijn eerste interesse in hoe mensen leren en hoe je dat kunt begrijpen. Vervolgens ging ik naar de University of California, Berkeley, waar ik me verder verdiepte in experimentele psychologie en statistiek. Die combinatie - gedrag en meten - is eigenlijk altijd centraal gebleven in mijn werk.
Mijn opleiding werd afgerond aan de University of Chicago, waar ik in 1940 mijn master behaalde en in 1944 promoveerde. Mijn proefschrift ging over leerprocessen en cog-

nitieve ontwikkeling. In die periode begon ik te beseffen dat psychologische metingen een krachtig instrument konden zijn om menselijk functioneren te doorgronden, mits ze zorgvuldig werden ontwikkeld en geïnterpreteerd.

Welke rol speelden leermeesters in die ontwikkeling?

Een cruciale rol. Aan Berkeley werkte ik met Jones, die mij leerde om altijd de mens achter de meting te blijven zien. Meten mag nooit een doel op zich worden. In Chicago kwam ik in aanraking met Tyler, die bekend stond om zijn visie dat toetsen altijd gekoppeld moeten zijn aan onderwijsdoelen. Dat idee, dat meten betekenis moet hebben binnen een bredere context, heeft mijn werk sterk beïnvloed. En dan was er Thurstone, een pionier op het gebied van factoranalyse. Van hem leerde ik hoe statistische technieken kunnen worden gebruikt om abstracte psychologische eigenschappen zichtbaar te maken. Als ik terugkijk, hebben deze drie leermeesters mij elk iets essentieels meegegeven: mensgerichtheid, onderwijskundige relevantie en statistische precisie. Mijn werk is eigenlijk een poging geweest om die drie samen te brengen.

Na uw promotie begon uw academische loopbaan. Hoe ontwikkelde die zich?

Mijn eerste aanstelling was als docent aan de University of Illinois. Dat was een bijzonder vruchtbare periode. Daar werkte ik zowel aan theoretische modellen als aan praktische tests. Het was belangrijk voor mij dat theorie en toepassing elkaar versterkten. Later, in de jaren vijftig, werd ik hoogleraar aan Stanford University. Dat werd de plek waar ik het grootste deel van mijn loopbaan doorbracht. Aan Stanford bouwde ik een onderzoeksgroep op die internationaal bekend werd. Wat mij altijd heeft gedreven, is het begeleiden van jonge onderzoekers, hen leren dat nauwkeurigheid en maatschappelijke relevantie hand in hand moeten gaan. Daarnaast heb ik veel bestuurlijke rollen vervuld. Ik was president van de American Psychological Association (APA) en later van de American Educational Research Association (AERA). Ook werkte ik in commissies van de National Academy of Sciences en de National Research Council. In die rollen probeerde ik steeds de brug te slaan tussen wetenschap en beleid.

U stond ook bekend als een invloedrijk redacteur en bestuurder. Wat betekende dat voor uw werk?

Het gaf mij de mogelijkheid om richting te geven aan het vakgebied. Als redactielid van tijdschriften zoals *Psychometrika* en *Educational and Psychological Measurement* kon ik bijdragen aan de kwaliteit van onderzoek. Maar misschien nog belangrijker was dat ik zag hoe ideeën zich ontwikkelen binnen een gemeenschap van onderzoekers. Wetenschap is geen individuele onderneming; het is een collectief proces waarin ideeën worden getoetst, verfijnd en soms verworpen.

U heeft vele onderscheidingen ontvangen. Hoe kijkt u daar tegenaan?

Ze weerspiegelen verschillende aspecten van mijn werk. De E.L. Thorndike Award erkende mijn bijdrage aan de onderwijspsychologie, terwijl de Distinguished Contributions Award van de APA mijn bredere invloed op de psychologie onderstreepte. Eredoctoraten, zoals die van Chicago, Illinois, Göteborg en Madrid, waren voor mij vooral een erkenning van de internationale betekenis van het werk dat ik deed. Maar uiteindelijk zijn prijzen niet het belangrijkste. Wat telt, is of het werk daadwerkelijk bijdraagt aan beter onderwijs en aan betere beslissingen.

Laten we naar uw inhoudelijke werk gaan. Uw naam is onlosmakelijk verbonden met betrouwbaarheid en psychometrie. Wat was daar uw belangrijkste bijdrage?

De kernvraag was eigenlijk eenvoudig: hoe weten we of een test of toets betrouwbaar is? Psychometrie probeert abstracte eigenschap-

pen zoals intelligentie of motivatie meetbaar te maken. Maar metingen zijn nooit perfect. In 1951 introduceerde ik een maat voor interne consistentie, die later bekend werd als Cronbachs alfa. Het idee was om te bepalen in hoeverre items binnen een test hetzelfde construct meten. Als de samenhang tussen items hoog is, kunnen we meer vertrouwen hebben in de uitkomst.
Het belang daarvan is groot. Tests en toetsen worden gebruikt om beslissingen te nemen, over toelating, diagnose, selectie. Als die tests niet betrouwbaar zijn, kunnen de gevolgen ernstig zijn.
Wat mij altijd heeft gefascineerd, is dat betrouwbaarheid geen puur technisch probleem is. Het heeft directe maatschappelijke implicaties. Een slechte meting kan leiden tot onrechtvaardige beslissingen.

Tegelijkertijd is er tegenwoordig kritiek op alfa.

Dat is terecht. Alfa is gebaseerd op bepaalde aannames, zoals tau-equivalentie, en is geen universele maat voor betrouwbaarheid. Ik zou zeggen: gebruik alfa, maar begrijp wat het wel en niet meet. Nieuwe maatstaven zoals omega of de *greatest lower bound* bieden alternatieven. Transparantie is hier essentieel. Onderzoekers moeten expliciet maken welke aannames ze doen en waarom ze bepaalde maten gebruiken.

Wat betekenen die begrippen?

Sorry, ik zal ze even uitleggen. Bij tau-equivalentie zijn er geen systematische verschillen tussen de vragen wat betreft hun moeilijkheid. Hun meetfout kan echter wel verschillend zijn. Omega geeft de betrouwbaarheid van een test aan zonder dat tau-equivalentie noodzakelijk is. De *greatest lower bound* is ook een schatter van de betrouwbaarheid en heeft een waarde die de betekenis heeft dat de betrouwbaarheid niet lager kan zijn.

Een andere belangrijke bijdrage is de generaliseerbaarheidstheorie, de G-theorie. Wat maakte die zo vernieuwend?

De klassieke testtheorie gaat uit van één bron van meetfout. Maar in werkelijkheid zijn er veel meer. Denk aan verschillen tussen items, beoordelaars en situaties. Met de generaliseerbaarheidstheorie wilden we een kader bieden om al die bronnen van variatie tegelijkertijd te analyseren. Testscores zijn nooit puur; ze zijn altijd het resultaat van meerdere invloeden. Dat inzicht heeft grote gevolgen. In het onderwijs kan een leerling anders scoren afhankelijk van de toets of de beoordelaar. Met G-theorie kun je berekenen hoe groot die effecten zijn en hoe je ze kunt beperken.

Een mooi voorbeeld is de beoordeling van schrijfvaardigheid. Daar spelen beoordelaars een grote rol. Met G-theorie kun je bepalen hoeveel beoordelaars nodig zijn om tot een betrouwbaar oordeel te komen. Het doel was altijd praktisch: betere beslissingen mogelijk maken door beter inzicht in meetfouten.

Uw werk reikt verder dan meten alleen. U heeft ook veel geschreven over evaluatie en onderwijsbeleid.

Ja, dat klopt. Ik vond dat evaluatie te vaak werd gereduceerd tot cijfers en gemiddelden. In mijn werk, zoals *Toward Reform of Program Evaluation*, heb ik gepleit voor een bredere benadering. Onderwijs vindt plaats in complexe contexten. Leerlingen verschillen, scholen verschillen, omstandigheden verschillen. Evaluaties moeten die complexiteit erkennen. Ik heb ook benadrukt dat toetsen niet alleen selectiemiddelen zijn. Ze moeten bijdragen aan leren. Toetsen die feedback geven, passen veel beter bij dat idee dan alleen summatieve toetsen. En misschien nog belangrijker: beleidsmakers moeten voorzichtig zijn met het trekken van algemene conclusies. Een toetsresultaat is nooit zonder context.

Hoe kijkt u naar uw invloed op de toetspraktijk?

Het is bijzonder om te zien hoe ideeën blijven doorwerken. Betrouwbaarheid en validiteit zijn nu standaardonderdelen van toetsontwikkeling. Formatieve toetsen wordt steeds belangrijker. Digitale systemen geven leerlingen of studenten direct feedback. Adaptieve toetsen passen zich aan aan het niveau van de leerling of student. Methoden uit de generaliseerbaarheidstheorie worden gebruikt om beoordelaarsvariatie te analyseren. Wat mij vooral verheugt, is dat er meer aandacht is gekomen voor rechtvaardigheid. Meten is nooit neutraal; het heeft gevolgen voor mensen. Dat besef lijkt sterker aanwezig dan vroeger.

Welke van uw onderzoeksterreinen en publicaties beschouwt u als het meest invloedrijk? Kunt u beginnen met uw bijdrage aan de psychometrie en betrouwbaarheid?

Als je psychometrie in de kern bekijkt, gaat het om een fundamenteel probleem: hoe maak je het onzichtbare zichtbaar? We proberen eigenschappen als intelligentie, motivatie of attitude te meten, maar dat zijn geen direct observeerbare grootheden. We construeren instrumenten, tests, toetsen, vragenlijsten, en hopen dat die een betrouwbare afspiegeling geven van wat we willen meten. Mijn bijdrage lag vooral in het expliciet maken van de vraag naar betrouwbaarheid. Wat betekent het eigenlijk als een test betrouwbaar is? In praktische zin betekent het dat je, als je dezelfde meting herhaalt onder vergelijkbare omstandigheden, tot vergelijkbare resultaten komt. Maar dat klinkt eenvoudiger dan het is. Binnen een test kunnen allerlei dingen misgaan: items kunnen verschillende aspecten meten, sommige vragen kunnen slecht geformuleerd zijn, of respondenten kunnen inconsistent reageren. In mijn artikel uit 1951 heb ik geprobeerd een hanteerbare maat te bieden voor wat we interne consistentie noemen. Wat ik wilde weten, was: in hoeverre ‘hangen’ de items binnen een test samen? Met andere woorden,

Als je psychometrie in de kern bekijkt, gaat het om een fundamenteel probleem: hoe maak je het onzichtbare zichtbaar?

meten ze daadwerkelijk hetzelfde onderliggende construct? De coëfficiënt die later mijn naam kreeg, biedt een samenvattend getal dat deze samenhang uitdrukt.

Het belang daarvan bleek al snel groot. Onderzoekers en docenten kregen een praktisch instrument om de kwaliteit van hun meetinstrumenten te beoordelen. Denk bijvoorbeeld aan een vragenlijst voor depressie. Als de alfa laag is, wijst dat erop dat sommige vragen niet goed aansluiten bij het onderliggende construct. Dat betekent dat de test moet worden herzien voordat je er vergaande conclusies aan verbindt. Tegelijkertijd moet men voorzichtig zijn. Alfa is geen universele maat voor betrouwbaarheid. Het veronderstelt bijvoorbeeld dat items in zekere zin uitwisselbaar zijn

– wat we dus tau-equivalentie noemen – en dat is lang niet altijd het geval. In moderne psychometrie zie je dan ook dat men pleit voor aanvullende maatstaven, zoals omega of de greatest lower bound. Wat mij betreft is dat een gezonde ontwikkeling. Het gaat er niet om één getal te hebben, maar om een goed begrip van wat je meet en onder welke voorwaarden.

U heeft dat denken verder uitgebreid met de generaliseerbaarheidstheorie. Wat was daar de kern van?

De aanleiding voor de generaliseerbaarheidstheorie was eigenlijk onvrede met de beperkingen van de klassieke testtheorie. Die theorie behandelt meetfouten vaak als één homogene bron van ruis. Maar in werkelijkheid is de situatie veel complexer. Neem een toets in het onderwijs. De score van studenten wordt niet alleen bepaald door hun vaardigheid, maar ook door de specifieke vragen die worden gesteld, door de beoordelaar die het werk nakijkt, en zelfs door de omstandigheden van het moment. In de klassieke benadering verdwijnen al die invloeden in één restterm. Dat vond ik onbevredigend.

Met de generaliseerbaarheidstheorie hebben we geprobeerd die complexiteit expliciet te maken. In plaats van één foutbron onderscheiden we meerdere componenten van variantie: variantie tussen personen, tussen items, tussen beoordelaars, en interacties daartussen. Door die componenten afzonderlijk te schatten, kun je precies zien waar de onzekerheid in een meting vandaan komt.

Wat deze benadering bijzonder maakt, is dat zij recht doet aan de complexiteit van menselijk gedrag. Meten is geen zuiver proces; het is altijd een samenspel van factoren. Door die factoren zichtbaar te maken, kun je betere, eerlijkere beslissingen nemen. Dat geldt niet alleen voor onderwijs, maar ook voor selectieprocedures, certificering en evaluatieonderzoek.

Een derde belangrijk terrein in uw werk is evaluatieonderzoek en onderwijsbeleid. Hoe verhoudt zich dat tot uw psychometrische werk?

Voor mij zijn die terreinen nooit gescheiden geweest. Meten heeft alleen betekenis binnen een context, en evaluatieonderzoek gaat precies over die context. In de jaren zeventig en tachtig zag ik dat evaluaties vaak werden gereduceerd tot cijfers en gemiddelden. Programma's werden beoordeeld op basis van gemiddelde effecten, zonder aandacht voor verschillen tussen leerlingen of scholen. Dat vond ik problematisch. In *Toward Reform of Program Evaluation* heb ik betoogd dat evaluatie veel breder moet worden opgevat. Onderwijs is geen homogeen systeem. Wat werkt in de ene context, werkt niet noodzakelijk in een andere. Daarom moeten evaluaties niet alleen de gemiddelde effecten onderzoeken, maar ook variatie en de omstandigheden waaronder effecten optreden.

Daarnaast heb ik altijd benadrukt dat toetsen niet alleen selectiemiddelen zijn. Ze moeten ook dienen als feedbackinstrumenten. Formatieve toetsen, toetsen die bedoeld zijn om het leerproces te ondersteunen, sluiten veel beter aan bij wat onderwijs in wezen is: een proces van ontwikkeling.

Er zit ook een belangrijke waarschuwing in dit alles. Wanneer beleidsmakers toetsresultaten gebruiken voor beslissingen over financiering of verantwoording, bestaat het risico dat onderwijs zich gaat aanpassen aan wat wordt gemeten. Dan vernauwt het curriculum tot datgene wat toetsbaar is.
Mijn pleidooi was daarom altijd tweeledig: gebruik toetsen, maar gebruik ze verstandig. Begrijp hun beperkingen, interpreteer resultaten in context, en wees terughoudend met generalisaties. Meten kan veel inzicht geven, maar alleen als het wordt ingebed in een breder begrip van onderwijs en samenleving. ■

Slotbeschouwing
Lee J. Cronbach was meer dan de bedenker van een statistische maat. Hij was een denker die het meten van menselijk gedrag steeds opnieuw in verband bracht met de context waarin dat gedrag plaatsvindt. Zijn werk laat zien dat toetsing nooit slechts een technische aangelegenheid is, maar altijd ook een normatieve en maatschappelijke dimensie heeft. In een tijd waarin data en metingen een steeds grotere rol spelen in onderwijs en beleid, blijft zijn boodschap opvallend actueel: meet zorgvuldig, interpreteer voorzichtig en verlies nooit de mens achter de cijfers uit het oog. Cronbachs nalatenschap leeft voort in elke toets die wordt geanalyseerd, in elke evaluatie die rekening houdt met context, en in elke onderzoeker die probeert bruggen te slaan tussen theorie en praktijk.

Zijn belangrijkste publicaties:

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671–684.
- Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). Harper & Row.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Wiley.

Gesignaleerd

NOS Nieuws
vrijdag 8 mei, 05:59

Eindexamens beginnen met voor het eerst ook een hulplijn tegen stress
Deel dit artikel. Bijna 145.000 leerlingen beginnen vandaag aan hun centrale examens. Voor vmbo gl en tl staat wiskunde op het programma, voor havo filosofie en Nederlands en voor vwo kunst en bedrijfseconomie.In totaal doen ruim 185.000 leerlingen eindexamen. Leerlingen die examen vmbo-basis en -kader doen zijn al eerder begonnen. Zij maken de algemene vakken zoals Nederlands digitaal. Die konden vanaf 1 april worden afgenomen en plant de school zelf in.

Mondeling examen
Voor wie het niet mogelijk is om mee te doen aan het reguliere examen is er het staatsexamen. Dit jaar hebben ruim 9500 leerlingen zich daarvoor aangemeld. Dat zijn er zo'n 1200 meer dan drie jaar geleden. Het College van Toetsen en Examens heeft geen verklaring voor de toename.
Lees verder: https://nos.nl/artikel/2613524-eindexamens-beginnen-met-voor-het-eerst-ook-een-hulplijn-tegen-stress