Bert Meuffels

Cijfers en cijferschalen

Iedereen die heeft schoolgegaan, is vertrouwd met de cijferschaal van 1 tot 10. Ook buiten het onderwijs wordt deze cijferschaal trouwens veelvuldig gebruikt, ten behoeve van allerlei soorten evaluatieve doeleinden. Amper is een baby ter wereld gekomen, of hij of zij krijgt al een cijfer opgeplakt, de zogenaamde Apgar-score (genoemd naar de grondlegger van dit systeem, de Amerikaanse Virginia Apgar). Met dit cijfer wordt uitgedrukt, in welke conditie de pasgeborene verkeert. Voor elk van vijf functies (hartslag, ademhaling, spierspanning, reactie op externe prikkels en kleur van de huid) kan de baby maximaal 2 punten krijgen. Behaalt hij of zij minder dan 6 punten in totaal, dan is extra oplettendheid geboden.

Cijfers blijven de jonggeborene zijn verdere leven achtervolgen. Als zuigeling wordt zijn functioneren tijdens de visites op het consultatiebureau opnieuw in cijfers uitgedrukt, en vervolgens volgt een lange schoolloopbaan waarbij het cijfers regent. De school ontgroeid en opgenomen in een passende werkkring blijven cijfers in het leven van de volwassen geworden burger een rol spelen, nu bij de zogenaamde functionerings- en beoordelingsgesprekken die tegenwoordig in het bedrijfsleven en bij de overheid tot de standaardprocedures van het personeelsbeleid gerekend mogen worden. Eenmaal met pensioen is men nog niet van cijfers bevrijd: is de levensfase aangebroken waarin opname in een bejaardenhuis aan de orde is, dan kan men enkel worden toegelaten wanneer men eerst een voldoende aantal punten heeft behaald op een aantal verschillende controle-punten.

Ons hele leven lijkt doordrenkt te zijn van cijfers: van de wieg tot het graf wordt elke Nederlander geconfronteerd met de over-bekende cijferschaal van 1 tot 10. Waarom drukken wij een beoordeling in de vorm van een cijfer uit? Welke interpretaties kun je aan zo'n cijfer toekennen? En wat betekent het eigenlijk, wanneer je een zes voor je opstel krijgt?

Deze vragen betreffende de interpretatie van cijfers op een cijferschaal vormen de achtergrond van dit artikel, waarin de betrouwbaarheid van opstelbeoordeling centraal staat. Betoogd wordt, dat een interpretatie van cijfers in termen van het daaraan ten grondslag liggende type meetschaal (nominaal, ordinaal, interval en ratio) van doorslaggevende betekenis is voor de mate van betrouwbaarheid van opstelbeoordeling. Een nominale interpretatie van cijfers leidt tot een volstrekt ander beeld van de mate van betrouwbaarheid dan een ordinale of een interval-interpretatie. Voorts zal onder andere betoogd worden dat de in Nederland gebruikelijke manier om de betrouwbaarheid te berekenen (namelijk in termen van een correlatie) grotendeels bepaald is door traditie, opvoeding en onderwijs, en dat deze berekeningswijze tot grove vertekening kan leiden in het beeld wat we van 'de' betrouwbaarheid van opstelbeoordeling voorgeschoteld krijgen.

De interpretatie van opstelcijfers en de betrouwbaarheid van die opstelcijfers staan in dit artikel weliswaar centraal, maar dat betekent natuurlijk niet dat de bevindingen en conclusies zich zouden beperken tot de *schrijf*vaardigheid en de beoordeling van juist die vaardigheid - ze gelden onverkort voor (de beoordeling van) alle taalvaardigheden.

1. Het systeem van 'merckteekenen' en van 'notae'

Onze cijferschaal is van een historisch dateerbare oorsprong en kent een lange voorgeschiedenis. Tegen het eind van de 16de eeuw werd op de Latijnse scholen (in de vroege middeleeuwen gesticht door de kloosters of bisschop en later ook wel door de steden) het systeem van 'merckteekenen' ingevoerd, de eerste rudimentaire vorm van een (verbaal) beoordelingssysteem. Om leerlingen tot gehoorzaamheid te dwingen en in het gareel te houden, brachten de leermeesters vrijwel dagelijks de stelling "Een goed leermeester spare de roede niet" in praktijk. Andere middelen stonden niet of nauwelijks tot hun beschikking. Om uitwassen te voorkomen en om een vermindering van die lijfstraffen te bewerkstelligen werden de 'merckteekenen' ingevoerd, aantekeningen over negatief en ongewenst gedrag van de leerling. Liepen die aantekeningen voor een bepaalde leerling de spuigaten uit, dan volgde er straf.

Op de Latijnse school in Zwolle hanteerde men in de 16de eeuw (1563) de volgende drie 'merckteekenen': (1) "Van alle soorten baldadigheid, en van ongeschikte manieren, spreeckingen, en daden" (2) Van gebreck en versuim aangaande boecken, inckt, schrijfpennen en andere schoolgeredschappen" (3) Van Nederduitsch spreecken op de straten en in de School".

De 'merckteekenen' werden in de loop van de 17de en 18de eeuw op de Schola Latina uitgewerkt tot het systeem van notae, dat de basis zou gaan vormen voor de dagelijkse beoordeling van de leerlingen. In dit systeem werd een onderscheid gemaakt tussen 'notae bonae' (goede aantekeningen) en 'notae malae' (slechte aantekeningen). De specifieke notae die konden worden toegekend, en de wijze van verrekening varieerde van school tot school, maar de opzet van het systeem was overal dezelfde.

In Haarlem hanteerde men in 1802 drie verschillende soorten notae bonae: 1. notae diligentiae (aantekeningen van attentheid) 2. notae industriae (aantekeningen van ijver) 3. notae modestiae (aantekeningen van goed gedrag). Daar tegenover stonden drie verschillende notae malae: 1. notae negligentiae (aantekeningen van onattentheid) 2. notae pigritiae (aantekeningen van luiheid) 3. notae petulantiae (aantekeningen van brutaliteit). Aan de verschillende notae werd een verschillend gewicht toegekend: één nota petulentiae bij voorbeeld stond gelijk met twee notae pigritiae. Met deze laatste nota werd elke leerling gestraft die weerspannig of ongehoorzaam gedrag vertoonde, of die zich te buiten ging aan onbetamelijke tegenspraak of wanordelijk gedrag binnen of buiten de school. Een nota diligentiae verdiende een leerling wanneer hij gedurende 14 dagen geen boete had

opgelopen wegens het niet-spreken van Latijn, of wanneer hij gedurende één hele week geen notae negligentiae had opgelopen. Met een nota diligentiae kon een leerling zijn foutentotaal met 1/12 terugbrengen.

Van al die verschillende notae met hun verschillend gewicht werd door de docenten nauwkeurig boek gehouden. Niet alleen bepaalden de notae de rangorde, de plaats in de klas (de 'primus' kwam de eer toe helemaal vooraan in de klas te mogen zitten, de 'secundus' daar vlak achter, enzovoort, terwijl voor de mindere goden de achterste rijen waren gereserveerd), maar ook vormden ze de basis voor de bevordering van een leerling. Bevordering was immers alleen mogelijk bij een zeker overwicht van notae bonae boven notae malae. Om voor bevordering in aanmerking te komen, moest een leerling in Gouda (1816) minstens 2/3 meer goede dan slechte notae hebben. Met dit stelsel van notae was, aldus Fortgens, de beoordeling van leerlingen zeer vergemakkelijkt, "daar alle prestaties, themafouten en gedragingen op één noemer (...) (waren) gebracht" (Fortgens 1958: 179). Van een examen in de eigenlijke zin des woords was dus geen sprake, aangezien de uitslag reeds van tevoren vaststond.

Als we Fortgens mogen geloven, is onze gewoonte om (school)prestaties in *cijfers* uit te drukken iets meer dan honderd jaar oud. "Een beoordeling in cijfers trof ik voor het eerst in Doesburg aan (cursus 1874/1875). Deze verving nu de opgave van het aantal fouten gemaakt in de Latijnse en Griekse themata. Men beperkte zich tot de volgende predikaten: 4 zeer goed, 3 goed, 2 voldoende, 1 gebrekkig" (Fortgens 1958: 182).

Dit systeem van notae dat in de loop der tijden uitgroeide tot een volwaardig cijfersysteem heeft het, zij het in sterk gemitigeerde vorm, zelfs tot in de jaren vijftig en zestig van deze eeuw uitgehouden. Op veel lagere scholen was het de gewoonte om op overgangsrapporten van de leerlingen niet alleen cijfers voor de reguliere vakken te vermelden, maar ook cijfers of verbale mededelingen voor de categorieën 'gedrag' (nota modestiae), 'vlijt' (nota industriae) en 'orde en netheid' (nota diligentiae).

2. De geboorte van onze cijferschaal

Vanaf 1815 vindt er via de reeks zogeheten onderwijswetten (Hubrecht, 1881) een steeds verdergaande standaardisering en uniformering van het Nederlandse schoolwezen plaats, inclusief de te hanteren beoordelingsschaal: leerplannen, aantal klassen, bevoegdheden van de leerkrachten en aard en inrichting van de toelatingsexamens worden aan een uniforme regeling onderworpen (Idenburg, 1964: 28). Niettemin behielden de scholen in het Organiek Besluit van 2 augustus 1815, no. 14, een zekere vrijheid in het toelatingsbeleid en in de samenstelling van hun leerplan, maar met die vrijheid was het, althans wat de door de landelijke overheid gecontroleerde Hogere Burgerscholen betreft, vanaf 2 mei 1863 definitief gedaan. De op die dag gedateerde *Wet op het middelbaar onderwijs* van de hand

van de liberaal Thorbecke bepaalde onverkort het aantal te onderwijzen vakken op 18.

De verplichting om die 18 vakken te onderwijzen gold slechts voor de rijksscholen, want de provinciale, gemeentelijke en bijzondere scholen mochten die vakkenlijst naar omstandigheden uitbreiden of inkrimpen. De leerlingen op de rijksscholen waren overigens niet verplicht alle lessen te volgen, alleen die welke ze nuttig vonden. En ook mochten ze plaatsnemen in de klas van hun eigen keuze. In dit relatief 'liberale' systeem pasten uiteraard geen verplichte overgangs- en eindexamens, maar "Thorbecke besefte dat de maatschappij waarborgen nodig had dat wie de middelbare school verliet tot min of meer zelfstandige denkarbeid in staat zou zijn. Er moest dus een eindexamen komen" (o.c.:41), dat zou worden afgenomen door externe examencommissies, benoemd door de "Commissarissen des Konings". "Toen begon de druk van het eindexamen op het leerprogramma der school, welke nimmer meer zou aflaten. De leerplan- en toelatingsvrijheden verdwenen als sneeuw voor de zon. Reeds na enkele jaren kwamen er nauwkeurige regels voor de inrichting van het eindexamen, een omschrijving van de kennis, welke per vak zou worden verlangd en voorschriften omtrent de beoordeling" (o.c. 41).

In mei 1868 vaardigde de toenmalige minister van Onderwijs Heemskerk een voorschrift uit betreffende de wijze van beoordeling van de eindexamens. In 1869 volgde hierop een "Ontwerp van Algemeen Reglement", dat bij Koninklijk Besluit van 10 maart 1870 tot wet werd verheven. Op deze datum werd de cijferschaal van 1 tot en met 10 officieel ingevoerd, zij het - alweer - uitsluitend voor de eindexamens van de Hogere Burgerscholen: "Het eindoordeel over de kennis der kandidaten (...) wordt (...) uitgedrukt door een der cijfers van 1 tot 10, aan welke de volgende beteekenis is te hechten:

10, uitmuntend
9, zeer goed
8, goed
7, ruim voldoende
6, voldoende
5, even voldoende
4, onvoldoende
3, gering
2, slecht
1, zeer slecht

(...) Is aan eenen kandidaat (...) het cijfer 5 of hooger toegekend, dan wordt hem het getuigschrift wegens voldoend afgelegd examen uitgereikt" (Staatsblad 49, art. 22 & 23).

Dat Fortgens pas in 1874 voor het eerst een beoordeling in cijfers aantreft - dus vier jaar na de officiële invoering van de cijferschaal op de Hogere Burgerschool - heeft waarschijnlijk te maken met het feit dat hij zich in zijn geschiedschrijving van het Nederlandse schoolwezen beperkt tot de Schola Latina. Zoals gezegd, voor andere dan de Burgerscholen gold de wet uit 1870 immers (nog) niet. Ook

toen de tiendelige cijferschaal al was doorgedrongen tot vrijwel alle lagen van ons onderwijsbestel, hanteerden de Gymnasia, de opvolgers van de vroegere Scholae Latinae, een afwijkende vorm van beoordeling, tot aan de jaren dertig van deze eeuw toe (beoordelingen vonden plaats op een 5-punt schaal, met de 1 als hoogste en de 5 als laagste kwalificatie).

De tiendelige cijferschaal die op 10 maart 1870 officieel in Nederland werd ingevoerd, heeft sinds haar invoering maar één belangrijke wijziging ondergaan. Bij Koninklijk Besluit van 8 juni 1929 werd bepaald dat de 5, die voor een 'even voldoende' prestatie stond (i.c. net voldoende), de betekenis kreeg van 'even onvoldoende', dus net niét voldoende. Een beslissing van hogerhand over de betekenis van de 5 was noodzakelijk, omdat er een vrij grote onzekerheid heerste aan welke kant de 5 nu eigenlijk stond (cf. De Groot 1968: 45). "Een stroom van artikelen werd in de vakpers en ook daarbuiten aan dit feit gewijd, zelfs in de Tweede Kamer werd het ter sprake gebracht, zodat het waarlijk leek, alsof de zegsman van Minister Waszink gelijk had, die van mening was, dat deze wijziging de belangrijkste was in de wetgeving op het middelbaar onderwijs sinds 1863" (Bartels, 1947: 126). De reglementen en de programma's voor de eindexamens van de Hogere Burgerscholen werden in 1943, als gevolg van de conjunctureel penibele situatie, en in 1962 als gevolg van de sterke stijging van het aantal eindexamenkandidaten nogal ingrijpend gewijzigd. De cijferschaal werd daarbij ongemoeid gelaten.

3. Cijferschalen over de grens

Het is onduidelijk, waarom minister Heemskerk voor een cijferschaal koos die uit 10 categorieën bestond, met klassen oplopend van 1 tot 10. Waarom niet een schaal van bij voorbeeld 1 tot 100, of van 1 tot 5? In andere landen binnen en buiten Europa figureren heel andere typen schalen (zie tabel 1).

Ter toelichting op deze tabel waarin de landen alfabetisch gerangschikt zijn, het volgende. Van de in de verschillende landen gehanteerde schaal worden alleen de extremen weergegeven tenzij de schaal uit letters bestaat), en wel zo dat het eerste getal verwijst naar het laagste en het tweede getal naar het hoogste cijfer. Concreet: voor de in Duitsland gehanteerde schaal van '6-1' geldt dus, dat 6 het laagste en 1 het hoogste cijfer vormt. Verder staat in de tabel de bij elke schaal behorende caesuur, de grens tussen voldoende en onvoldoende. Voor de Nederlandse cijferschaal ligt die caesuur bij de 6 (6 tot 10 voldoende, 1 tot en met 5 onvoldoende), voor Duitsland bij voorbeeld bij de 5 (alleen de 6 fungeert daar als onvoldoende; verdere differentiaties in 'onvoldoende' kent de Duitse schaal niet)[1].

In alle landen lopen de cijfers op de schaal netjes op (of af, zoals in Duitsland), maar zo niet in Denemarken (althans op het gymnasium aldaar). Daar bevat de schaal van 0 tot 13 niet veertien cijfers, zoals je zou verwachten, maar slechts tien: 0 (volledig onacceptabel), 3 (erg zwak), 5 (zwak), 6 (enigszins zwak) en vervolgens 7, 8, 9, 10 en 11 (excellent) - de 12 ontbreekt weer, net zoals de 1, 2,

en 4 en tot slot de 13 (briljant, excellent werk). Waarom juist de cijfers 1, 2, 4 en 12 ontbreken, is niet geheel duidelijk (cf. Schultze 1969: 408-409).

Tabel 1: Cijferschaal (extremen met de caesuur) in verschillende landen

Land	Schaal	Caesuur
België	1-10	6
Denemarken	0-13	6
Duitsland	6-1	5
Frankrijk	1/20-20/20	12/20
Groot-Brittannië	1-100	41
	F E D C B A	E
Ierland	1-100	41
Italië	0-10	6
Luxemburg	0-60	30
Nieuw-Zeeland	F D C B A	D
Noorwegen	0-6	2
Oostenrijk	6-1	5
Portugal	0-20	5
	M m S B MB	m
Sovjet-Unie	1-5	3
Spanje	1-10	5
Verenigde Staten	0-100	65
	F D C B A	D
Zweden	1-5	-
	C Bc B Ba AB a A	Bc
Zwitserland	6-1	5

In de meeste landen bestaat er, anders dan de suggestie die in tabel wellicht wordt gewekt, geen uniforme cijferschaal voor zowel het primair, secundair als tertiair onderwijs. In Noorwegen bij voorbeeld worden op de basisschool geen cijfers gegeven, maar alleen verbale beoordelingen als 'goed', 'kan beter', enzovoort. Op de Noorse middelbare school wordt daarentegen wel gebruik gemaakt van een cijferschaal (van 0 t/m 6), maar deze schaal verschilt dan weer van die in het hoger onderwijs (van 1 t/m 4, waarbij de 4 voor 'onvoldoende' staat, en de overige drie cijfers voor graden van voldoende) (Hove, 1958).

Zelfs binnen één type onderwijs bestaan soms, zoals in de Verenigde Staten op de High School, verschillende cijferschalen (of beter uitgedrukt: beoordelingssystemen) naast elkaar: de letteraanduiding (A t/m F) en de cijferschaal (0 t/m 100). Die letteraanduiding heeft in landen als de Verenigde Staten, het Verenigd Koninkrijk en Nieuw-Zeeland uitsluitend een symbolische betekenis, waarbij de A staat voor 'outstanding', de B voor 'very good', de C voor 'satisfactory', de D voor 'generally unsatisfactory' en de F voor 'failed'; de letter E komt op de betreffende schaal niet voor, vanwege de verwarring met de letter A die bij het uitspreken ervan zou kunnen ontstaan (Crombag & De Gruyter, 1974). In Portugal daarentegen staan de letters M, m, S, B en MB voor afkortingen van respectieve-

lijk mau (=slecht), mediocre (=middelmatig), sufficiente (=bevredigend), bom (=goed) en muito bom (=zeer goed). "I can think of less confusing (systems)", aldus Newcombe (1977: 171) over dit Portugese afkortingen-stelsel.

Tussen 1930 en 1940 veranderden de meeste onderwijsinstellingen in de V.S. het toen vigerende cijfersysteem 'percentage grading' in het beoordelingssysteem met letteraanduiding (Geisinger, 1982: 1142). In het eerstgenoemde cijfersysteem (van 0 tot 100) werd elk aan een leerling toegekend getal/cijfer geacht te corresponderen met het percentage stof (materiaal) dat de betreffende leerling zou beheersen. Getallen lager dan 50 werden zelden gegeven. Tegenwoordig wordt nog maar op 16 percent van de Amerikaanse High Schools het 'percentage grading' gebruikt, de overgrote meerderheid hanteert het 'letter-grade' systeem (Terwilliger, 1966). In vrijwel alle landen worden er overigens voortdurend meer of minder grote veranderingen in de bestaande cijferschalen en beoordelingssystemen aangebracht. Het is dan ook nog maar de vraag, of de in tabel getabelleerde cijferschalen waarvan het bestaan vaak achterhaald is uit literatuur van enkele decennia oud, nog wel als zodanig dienst doen en niet door modernere varianten zijn achterhaald.

Op het eerste gezicht lijkt het er veel op, dat de Amerikaanse High Schools met de introductie van het 'letter-grade' systeem heel wat differentiatiemogelijkheden overboord hebben gegooid. Bestond het oude systeem nog uit 100 klassen, het nieuwe bevat er slechts vijf. Dit bezwaar wordt echter sterk afgezwakt als men beseft, dat uit onderzoek is gebleken dat leerkrachten niet of nauwelijks in staat zijn om betrouwbare differentiaties aan te brengen die kleiner zijn dan 3 tot 7 percentage-punten (Starch, 1913). De effectieve differentiatiemogelijkheden reduceren daarmee tot 20. Voeg daar nog bij dat het nieuwe beoordelingssysteem met letteraanduiding in de praktijk vaak gemodificeerd wordt door aan de letters de symbolen + en - toe te voegen, zodat het aantal differentiatiemogelijkheden niet uit vijf, maar uit 15 klassen bestaat, en het wordt duidelijk dat beide beoordelingssystemen feitelijk nog maar weinig van elkaar verschillen.

Het vergroten van het aantal differentiatiemogelijkheden van een bestaande beoordelingsschaal met een aantal extra klassen, door gebruik te maken van symbolen als de 'plus', de 'min', de 'schuine streep' (A/B), het 'vraagteken' of simpelweg decimalen is karakteristiek voor de manier waarop de schalen in vrijwel alle landen uit tabel 1 in de praktijk gebruikt worden (het 'halfje' schijnt een typisch Nederlandse verworvenheid te zijn; nergens anders ben ik dit tegengekomen). Dat impliceert dat het aantal klassen per schaal zoals weergegeven in tabel 1, in feite als een absolute ondergrens moet worden gezien.

Weer een andere, relevante opmerking naar aanleiding van tabel 1 betreft het volgende. Ook de schalen die gekenmerkt worden door letters, zoals in de V.S., Portugal en Zweden, worden vaak in een numerieke vorm gegoten. Zo worden in de V.S. de afzonderlijke letters gemiddeld tot een totaalindex, de zogeheten GPA (Grade Point Average), door respectievelijk de getallen 4, 3, 2, 1, en 0 toe te kennen aan A, B, C, D en F.

Eén van de conclusies die men naar aanleiding van het ontstaan van de cijferschaal in Nederland en naar aanleiding van al deze opmerkingen over de beoordelingsschalen in tabel 1 kan trekken is, dat de in verschillende landen gehanteerde systemen min of meer arbitrair zijn, op *conventies* berusten en dat de betekenis van de in die schalen figurerende cijfers (of letters) een strikt geconventionaliseerd karakter hebben die, in beginsel althans, niets van doen hebben met ons numerieke stelsel en de daaruit voortvloeiende interpretatie van *getallen*. Een ruim voldoende prestatie uitdrukken in de vorm van het cijfer 7 is niets anders dan een 'short cut', symbolischgetalsmatige weergave van de evaluerende kwalificatie "ruim voldoende". Dit heeft, zoals in de volgende paragraaf nader zal worden toegelicht, vergaande consequenties voor de interpretatie van cijfers en voor de analyse van de betrouwbaarheid van opsteloordelen.

Een andere conclusie die naar aanleiding van tabel 1 kan worden getrokken is, dat ondanks hun grote verscheidenheid de gehanteerde beoordelingsschalen niettemin drie fundamentele kenmerken gemeen hebben:

(1) alle schalen maken een differentiatie mogelijk in het prestatieniveau
(2) alle schalen leggen een rangorde in de prestaties vast
(3) alle schalen bevatten een caesuur tussen voldoende en onvoldoende prestaties.

Ook aan deze kenmerken zijn consequenties verbonden voor de betrouwbaarheidsproblematiek.

4. Schaaltypen

Stel dat een leerling voor zijn opstel een 8 heeft gekregen. Is deze leerling nu twee keer zo goed als iemand met een 4? Kan men met andere woorden staande houden, dat de schrijfvaardigheid van deze persoon met een 8 twee keer zo groot is als die met een 4? Het hoeft weinig betoog, gelet op het arbitraire karakter van cijferschalen, dat een dergelijke interpretatie niet adequaat is. Het zou bij voorbeeld impliceren, dat in theorie - in Amerika de ene persoon 100 maal zo schrijfvaardig kan zijn dan de ander, in Luxemburg daarentegen 'slechts' 60 maal, terwijl men in Nederland niet verder zou komen dan 'de een is tien maal zo schrijfvaardig dan de ander'. En hoe zou je, wanneer je uitgaat van dit type interpretaties, dan moeten omgaan met de Noorse cijferschaal (gebruikt op de toelatingsexamens voor de universiteit (Hylla & Wrinkle, 1953: 92)) die de cijfers 4, 3, 2, 1, -2 en -3 bevat? Hoeveel 'beter' is nu de persoon met het cijfer 4 dan degene met bij voorbeeld het cijfer -3? De hier gewraakte interpretatie gaat er ten onrechte van uit, dat cijfers absolute grootheden zijn, en miskent hun relatieve karakter.

De interpretatie van cijfers op een cijferschaal (of, om het even, letters op een letterschaal waaraan men een numerieke interpretatie heeft toegekend) is afhankelijk van het type *meetschaal*. In de regel worden er vier verschillende typen meetschalen onderscheiden: (1) een nominale schaal, (2) een ordinale schaal, (3)

een intervalschaal en (4) een ratioschaal. Op elk van deze vier meetschalen hebben getallen c.q. cijfers telkens een heel andere interpretatie.

Getallen op een nominale schaal hebben als functie objecten of individuen te identificeren. Een getal op een nominale schaal kent aan een individu of object een naam (nomen) toe, een label, en in deze zin fungeert een dergelijk getal als kengetal. Voorbeelden van getallen op een nominale schaal zijn rugnummers van voetballers, kentekennummers van auto's, telefoonnummers.

Getallen die worden gebruikt om individuen of objecten te identificeren en daarnaast relaties tussen de betreffende individuen of objecten te symboliseren, constitueren een ordinale schaal. Kenmerkende relaties voor een ordinale schaal zijn relaties als 'groter dan', 'beter dan', 'meer dan', enzovoort. Zulk soort relaties leggen een volledige *rangorde* van de individuen of objecten vast, en de getallen op een ordinale schaal zijn derhalve *rangnummers*. Illustratieve voorbeelden van getallen op een ordinale schaal vormen huisnummers, of de nummers 1, 2 en 3 in de sportwereld.

Een intervalschaal is een schaal waarvoor geldt dat de getallen nominale en ordinale eigenschappen bezitten, en waarvoor bovendien geldt dat de *afstand* tussen twee getallen een reële betekenis heeft. Voorbeelden van een intervalschaal zijn de temperatuurschaal en de jaartelling. Een gebeurtenis die in het jaar 2000 plaatsvindt, treedt niet alleen in een ánder jaar op dan bij voorbeeld in het jaar 1000 (nominale interpretatie), maar ook later (ordinale interpretatie). Bovendien is de afstand, het verschil tussen het jaar 2000 en het jaar 1000 even groot als het verschil tussen het jaar 1000 en het jaar 0 (interval-interpretatie). Men kan echter niet beweren dat het jaar 2000 twee maal 'zo laat' plaatsvindt als het jaar 1000, net zo min als men kan volhouden dat een temperatuur van 20 graden twee keer zo warm is als die van tien graden. Dat dergelijke beweringen onzinnig zijn is een gevolg van het feit dat het jaar 0 in onze jaartelling of een temperatuur van nul graden geen 'echt', absoluut nulpunt constitueert (immers, ook het jaar 25 voor Christus of een temperatuur van min drie graden behoort tot de mogelijkheden). Op een intervalschaal is het nulpunt met andere woorden arbitrair.

Bevat een schaal alle eigenschappen van zowel een nominale, een ordinale als een intervalschaal, en bevat die schaal daarnaast een absoluut nulpunt dat reële betekenis heeft, dan wordt zo'n schaal een ratioschaal genoemd. Het is deze schaal waarmee elke Nederlander in het lager onderwijs in de rekenles kennis heeft gemaakt, en die hem zó vertrouwd voorkomt dat hij bij getallen niet of nauwelijks nog kan denken in termen van andere schalen en daarmee van andere interpretaties dan die, welke op een ratioschaal van toepassing zijn. Voorbeelden van een ratioschaal zijn de gewichtschaal en de lengteschaal. Omdat de gewichtschaal een absoluut nulpunt kent, kan men, anders dan bij getallen op een intervalschaal, beweren dat een gewicht van 10 kg twee maal zo zwaar is als een gewicht van 5 kg.

Het onderscheid tussen de hier genoemde schaaltypen is uiteraard van belang voor een adequate interpretatie van getallen in het algemeen, en van schoolcijfers op een cijferschaal in het bijzonder. De nummer 3 in een hardloopwedstrijd is niet

drie maal zo langzaam als de nummer 1, maar van een gewicht van 3 kg kan wel volgehouden worden dat het drie maal zo zwaar is als een gewicht van 1 kg. Evenmin kan men staande houden dat het verschil tussen de nummers 1 en 2 in de sportwereld even groot is als dat tussen de nummers 2 en 3, of dat een voetbalspeler met rugnummer 14 twee keer zo goed is als die met rugnummer 7, maar het is wel zinvol om bij voorbeeld te beweren dat het verschil tussen 80 cm en 70 cm even groot is als dat tussen 50 en 40 cm.

Op welk type schaal moeten onze schoolcijfers nu gesitueerd worden? Schoolcijfers laten in ieder geval een nominale interpretatie toe: de persoon met een 5 heeft een andere prestatie geleverd dan die met een 8. Bovendien impliceert een nominale interpretatie dat de leerlingen die eenzelfde cijfer hebben gekregen, een identieke prestatie hebben geleverd. Ook een ordinale interpretatie is zinvol: een 8 staat voor een betere prestatie dan een 7, een 7 voor een betere dan een 6, enzovoort. Het zijn deze *twee eigenschappen* van het ons bekende tientallig numerieke stelsel (i.c. identiteit en ordening) die maken dat cijfers bij uitstek geschikt zijn om er, kort en krachtig en zonder veel omhaal van woorden, prestaties mee aan te duiden. Het feit, dat in verschillende landen over de hele wereld zonder uitzondering *cijfer*schalen worden gebruikt om er (graden van) prestaties mee uit te drukken, moet dan ook verklaard worden uit deze nominale en ordinale eigenschappen van ons getallenstelsel.

Verder kan gesteld worden, dat een ratio-interpretatie van schoolcijfers ontoelaatbaar is. De vraag echter of schoolcijfers getallen op een intervalschaal constitueren, is veel lastiger te beantwoorden. Waarom zou het verschil tussen de kwalificatie 'ruim voldoende' en 'voldoende' even groot moeten zijn als het verschil tussen 'ruim voldoende' en 'goed'? Aan de oplopende reeks verbale aanduidingen uit het "Besluit van den l0den Maart 1870" is geen enkel argument te ontlenen waarom dat noodzakelijk zo zou moeten zijn. Dat impliceert dat - in beginsel althans - onze cijferschaal geen interpretaties op intervalniveau toelaat.

In de *praktijk* echter zullen de meeste leerkrachten, juist krachtens het feit dat ze bij het beoordelen van prestaties gebruik maken van het numerieke stelsel en weten dat de afstanden tussen opeenvolgende getallen even groot zijn, er voor zorgen dat die afstanden bij benadering corresponderen met min of meer identieke verschillen in beoordeelde prestaties. Anders uitgedrukt: bij het toekennen van een 6 en een 7 waken ze ervoor, dat het prestatieverschil tussen de leerlingen met een 6 en met een 7 min of meer even groot is als het prestatieverschil tussen leerlingen met bij voorbeeld een 8 en een 9. Maar nogmaals, uit de gegeven verbale omschrijvingen van de klassen van onze cijferschaal (10= uitmuntend; 9= zeer goed; enzovoort) valt geenszins dwingend af te leiden, dat onze cijferschaal interval-eigenschappen bezit, het is de psychologie van de beoordelaar/cijferaar die een dergelijke interval-interpretatie mogelijk maakt.

Maar zelfs de leerkracht die bewust bij zijn cijfergeving geen interval-eigenschappen nastreeft, gaat er bij de berekening van de jaarlijkse eindcijfers ten behoeve van de overgang nolens volens vanuit dat de getalsmatige prestaties van

de leerlingen de facto op een intervalschaal zijn uitgedrukt. Immers, het optellen van de cijfers voor de prestaties die de leerlingen in de loop van het jaar op verschillende gelegenheden geleverd hebben en het middelen daarvan vooronderstelt dat de afzonderlijke prestaties op een intervalschaal gesitueerd kunnen worden (middelen op een ordinale of een nominale schaal is zinloos).

Twijfel over de rechtmatigheid van het toekennen van een interval-interpretatie aan schoolcijfers is onder meer gebaseerd op het cijfergedrag van leerkrachten rond de fatale caesuur, de 6. Is de afstand tussen het cijfer 5 en 6 wel even groot als die tussen de 6 en de 7, gelet op de onmiskenbare en in pedagogisch opzicht begrijpelijke neiging van leerkrachten om in twijfelgevallen niet voor een onvoldoende, maar voor een mager zesje te kiezen? Empirisch onderzoek naar het cijfergedrag van leerkrachten (Van den Ende, 1954a; Van den Ende, 1954b; De Groot, 1968: 92-93) maakt duidelijk, dat de bedoelde afstanden beslist niet even groot zijn en dat dus de aanname van een intervalschaal (i.c. gelijke afstanden) niet terecht is.

Of het nu wel of niet gerechtvaardigd is om aan opstelcijfers interval-eigenschappen toe te kennen, een empirisch te constateren feit is dat in het gros van het empirisch onderzoek naar de betrouwbaarheid van opstelbeoordeling er zonder meer vanuit wordt gegaan, dat opstelcijfers op een intervalschaal gesitueerd kunnen worden.

5. Drie vormen van betrouwbaarheid

Volgens De Groot heeft de hier behandelde materie,"de keuze van een meetschaal, niets te maken (...) met het probleem van de objectiviteit en van de betrouwbaarheid van schoolcijfers" (De Groot, 1968: 34). In abstracto mag deze bewering wel waar zijn, maar als we de betrouwbaarheid concretiseren op het niveau van schoolcijfers en daaraan een *kwantitatieve* interpretatie moeten geven, dan is die bewering, hoe apodictisch ook geformuleerd, apert onjuist.

Wat moeten we onder betrouwbaarheid verstaan? Wanneer kunnen we opstelcijfers betrouwbaar noemen? Betrouwbaarheid is een vrij omvattend en gecompliceerd begrip. In veel boeken wordt uitgebreid en diepgaand op verschillende aspecten van dit begrip ingegaan (Davis, 1964; Horst, 1966; Lord & Novick, 1968; Nunnaly, 1967; Rozeboom, 1966) en in maar liefst drie tijdschriften (*Psychometrika; Educational and Psychological Measurement; Journal of Educational Statistics*) houdt men zich intensief bezig met de betrouwbaarheidsproblematiek. Het spreekt voor zich dat de hier te presenteren behandeling van betrouwbaarheid, toegespitst op opstelbeoordeling, niet anders dan onvolledig kan zijn. Er komt slechts een beperkt aantal aspecten van de betrouwbaarheidsproblematiek aan de orde. Daar komt nog bij dat betrouwbaarheid een sterk geformaliseerd begrip is, dat definitorisch is vastgelegd in axioma's en daarvan afgeleide mathematische formuleringen. De hier te presenteren behandeling van het begrip

betrouwbaarheid is echter niet-formeel, zodat onvermijdelijk enige precisie en exactheid verloren gaat.

Wat houdt betrouwbaarheid nu in? Wat betekent het eigenlijk wanneer we een persoon, een politicus bijvoorbeeld, betrouwbaar noemen? We zijn geneigd een politicus het predicaat 'betrouwbaar' te verlenen als *bij herhaling* gebleken is dat hij zich aan zijn woord houdt. Betrouwbare politici komen hun verkiezingsbeloften na, verdedigen tegenover een hen welgezind publiek precies hetzelfde standpunt als tegenover een vijandig auditorium, liegen niet, zijn geloofwaardig, kortom men kan staat op hen maken. Hoe globaal deze karakterisering van het begrip betrouwbaarheid ook moge zijn, één aspect daarvan komt toch duidelijk naar voren: *konsistentie*. Betrouwbare personen handelen zowel in verbaal als niet-verbaal opzicht konsistent en daarom zij hun gedragingen en handelingen - gegeven zekere restricties - *voorspelbaar*. Een onbetrouwbaar iemand daarentegen zegt A, maar doet B of C, of misschien ook wel A; zijn gedrag is niet of nauwelijks voorspelbaar.

De overwegingen die bij het toekennen van het predicaat 'betrouwbaar' aan personen een rol spelen, namelijk konsistentie en voorspelbaarheid, zijn evenzeer van belang wanneer het gaat om de betrouwbaarheid van meetinstrumenten. We noemen een weegschaal betrouwbaar indien dit instrument op konsistente wijze iemands gewicht bepaalt. Heeft de weegschaal uitgewezen dat iemands gewicht 65 kg bedraagt, dan zal die weegschaal (aangenomen dat hij betrouwbaar is) bij een herhaalde weging die onmiddellijk na de eerste plaatsvindt, datzelfde gewicht aangeven. De eerste en de tweede weging zijn onderling konsistent, en de tweede weging is perfect voorspelbaar op basis van de eerste.

Conform de voorgaande beschouwing noemen we de beoordeling van opstellen betrouwbaar wanneer de resultaten van een eerste meting konsistent zijn met die van een tweede. Maar anders dan bij de konsistentie van de weegschaal, kan de konsistentie bij de beoordeling van opstellen verschillende dingen betekenen, die ook verschillende vormen van betrouwbaarheid inhouden. Wanneer een beoordelaar met een tussenperiode van twee of drie weken dezelfde opstellen beoordeelt en hij bij die twee gelegenheden hetzelfde oordeel over de kwaliteit van die opstellen velt, dan is hij een betrouwbaar beoordelaar. In dat geval is hij immers 'konsistent' met zichzelf. Deze specifieke vorm van betrouwbaarheid wordt *stabiliteit* genoemd.

Een tweede vorm van betrouwbaarheid die bij de beoordeling van opstellen onderscheiden kan worden, betreft de *interbeoordelaarsovereenstemming* (ook wel intersubjectieve overeenstemming genoemd). Hierbij is de vraag niet zozeer of een beoordelaar het met zichzelf eens is, maar of hij het met anderen eens is. Komen zijn beoordelingen overeen met die van anderen, die onafhankelijk van hem en van elkaar dezelfde opstellen hebben beoordeeld?

Zelfs als het mogelijk zou zijn om een perfecte stabiliteit en interbeoordelaarsovereenstemming te bereiken, dan nog hoeft de beoordeling daarmee niet per se in alle opzichten betrouwbaar te zijn. Dit hangt samen met de variabiliteit in prestatie. Als men iemands rijvaardigheid wil beoordelen, krijgt men natuurlijk

geen betrouwbaar beeld wanneer die persoon alleen in een relatief kort tijdsbestek en in een specifieke situatie geobserveerd wordt. Gedurende die korte observatieperiode kan de persoon in kwestie net even wat minder geconcentreerd zijn, of door de toevallige verkeersdrukte in de stad kan hij net even enigszins verkrampt rijden. Het beeld dat wij op basis van de observaties in dat korte tijdsbestek en in die specifieke situatie van zijn rijvaardigheid krijgen, hoeft niet representatief te zijn voor zijn rijvaardigheid op andere momenten en in andere situaties. Afhankelijk van het tijdstip waarop en de situatie waarin we iemand aantreffen, fluctueert het te beoordelen gedragsaspect in zekere mate. Kortom, er is variabiliteit in de prestatie en dat maakt dat beoordelingen die op verschillende momenten en in verschillende situaties plaatsvinden, niet volledig konsistent hoeven zijn. Dat geldt ook bij de beoordeling van schrijfvaardigheid. Iemands schrijfprestatie kan, afhankelijk van zijn kennis over het specifieke onderwerp waarover hij schrijft, van de tijd die hij ervoor tot zijn beschikking heeft, van zijn interesse in het te behandelen onderwerp enzovoort, variëren. Hier gaat het bij de betrouwbaarheid dus om de vraag, in hoeverre iemands prestaties bij het verrichten van een specifieke taak konsistent zijn met zijn prestaties in een veel groter assortiment gelijksoortige taken. Deze vorm van betrouwbaarheid wordt *teststabiliteit* genoemd (soms ook wel testvariabiliteit).

De drie hier onderscheiden vormen van betrouwbaarheid stabiliteit, interbeoordelaarsovereenstemming en test-stabiliteit - zijn met het oog op de betrouwbaarheid van opstelbeoordeling weliswaar de belangrijkste, maar zeker niet de enige vormen die onderscheiden zouden kunnen worden (cf. Meuffels, 1983). Hoe het ook zij, betrouwbaarheid is in elk geval geen ééndimensioneel begrip: "There is no single, universal and absolute reliability coefficient" (Stanley, 1971: 363).

We beperken onze beschouwingen over betrouwbaarheid en meetschalen verder tot de eerste twee vormen van betrouwbaarheid (i.c. stabiliteit en intersubjectieve overeenstemming), namelijk die vormen van (in)konsistentie die het gevolg zijn van het feit, dat menselijke beoordelaars imperfecte meetinstrumenten zijn.

6. Betrouwbaarheid op niveaus

Anders dan De Groot betoogt, heeft het bepalen van de betrouwbaarheid aan de hand van gegeven cijfers juist alles te maken met de keuze van een meetschaal, i.c. met de vraag welke schaaleigenschappen cijfers kunnen worden toegekend: nominale, ordinale, of ook interval-eigenschappen? Stel dat twee leraren (A en B) elk, onafhankelijk van elkaar, dezelfde drie opstellen nakijken. Oordelen zij betrouwbaar, dat wil zeggen is er sprake van intersubjectieve overeenstemming?

Aan geen enkel opstel wordt door beoordelaar A en B hetzelfde cijfer (i.c. label) toegekend, zodat een onderzoeker die opstelcijfers op nominaal niveau analyseert tot de conclusie moet komen dat de beoordelingen volledig inkonsistent met elkaar zijn, dus absoluut onbetrouwbaar. Een onderzoeker echter die ordinale eigenschappen aan opstelcijfers toekent, trekt de diametraal tegenovergestelde

conclusie, namelijk dat de beoordeling perfect betrouwbaar is. Immers, de rangorde die elk van de twee beoordelaars heeft toegekend, de relatieve positie van de drie opstellen ten opzichte van elkaar, is bij elke beoordelaar exact dezelfde. Weer een andere onderzoeker die interval-eigenschappen aan opstelcijfers toekent, komt weer tot iets andere conclusies. Hij zal betogen dat de twee beoordelaars weliswaar betrouwbaar zijn voorzover het gaat om de rangorde en om het algemene niveau (zowel bij A als bij B is het gemiddelde van de drie opstellen een 6), maar dat ze onbetrouwbaar beoordelen wat betreft de range, i.c. het verschil tussen het beste en het slechtste opstel (voor A bedraagt die range 5, voor B 4). Verder zal hij misschien betogen dat A en B onbetrouwbaar beoordelen omdat ze andere minimum-eisen hanteren. A kent slechts één onvoldoende toe, B daarentegen twee. Kortom, de mate van betrouwbaarheid van (school)cijfers is, anders dan De Groot ons wil doen geloven, juist volledig afhankelijk van de keuze van een meetschaal.

Tabel 2: Cijfers voor drie opstellen, toegekend door twee beoordelaars A en B

Opstel	A	B
1	8	8,5
2	7	5
3	3	4,5

Uiteraard blijven de hier gemaakte opmerkingen onverkort van kracht wanneer men niet de intersubjectieve overeenstemming, maar de stabiliteit van een beoordelaar onder de loep zou nemen, i.c. de konsistentie van de door een beoordelaar gegeven cijfers voor dezelfde drie opstellen op twee verschillende tijdstippen. In tabel 2 leze men dan in plaats van beoordelaar A en B het tijdstip A en B waarop de drie opstellen door dezelfde beoordelaar beoordeeld zijn.

Wat is nu *de* betrouwbaarheid van de beoordelingen in bovenstaande tabel? Het antwoord is simpel: *de* betrouwbaarheid bestaat niet. Eén van de bedoelingen van de bovenstaande exercitie op het vlak van betrouwbaarheidsanalyse is te laten zien, dat de oordelen van beoordelaars op heel veel verschillende manieren al of niet konsistent met elkaar kunnen zijn. Beoordelaars kunnen van mening verschillen over de kwaliteit van één opstel, of over de relatieve positie van de opstellen ten opzichte van elkaar, of over het maximale verschil in kwaliteit, of over het niveau van de groep als geheel, of over de te hanteren minimum-eisen, enzovoort. Het punt waar het hier nu om gaat, is dat het merendeel van al deze onderscheiden aspecten van de betrouwbaarheid *onafhankelijk* van elkaar is (een uitzondering op deze claim vormt de afhankelijkheid tussen het niveau en het aantal onvoldoendes; iemand die het niveau erg laag inschat, zal door de bank genomen veel onvoldoendes geven). Onafhankelijk betekent hier, dat uit het feit dat twee beoordelaars bij voorbeeld niet unaniem oordelen, dus aan geen enkel opstel

precies hetzelfde cijfer toekennen (nominale interpretatie), men geen enkele conclusie kan trekken over de mate van betrouwbaarheid op ordinaal niveau. Neem, ter illustratie van deze onafhankelijkheid, de volgende drie situaties waarin twee beoordelaars elk dezelfde vijf opstellen beoordelen.

Tabel 3: Drie verschillende vormen van konsistentie

Opstel	situatie I		situatie II		situatie III	
	A	B	A	B	A	B
1	1	6	4	1	6	10
2	2	7	4.5	3	7	9
3	3	8	5	5	8	8
4	4	9	5.5	7	9	7
5	5	10	6	9	10	6

A en B zijn het in situatie I niet eens over het niveau van de groep als geheel, maar ze zijn daarentegen wel volledig konsistent in hun oordeel over de rangorde van de vijf opstellen en ook in hun oordeel over de spreiding, de range. Geheel anders dan in situatie I, blijkt er in situatie II volledige overeenstemming te bestaan over de rangorde en het niveau, maar niet over de spreiding. In situatie III ten slotte bestaat er wel overeenstemming over het niveau en de spreiding, maar niet over de rangorde.

Deze voorbeelden maken opnieuw duidelijk, dat verschillende aspecten van betrouwbaarheid onderling onafhankelijk zijn, dat het ene betrouwbaarheidsaspect niet voorspeld kan worden op grond van kennis van het andere betrouwbaarheidsaspect, en dat uitspraken in de trant van "die beoordeling is onbetrouwbaar" of "die beoordelaars zijn het niet met elkaar eens" zonder verdergaande specificatie misleidend onvolledig zijn. "There is no single, universal and absolute reliability coefficient" (Stanley o.c.). Op abstract niveau is betrouwbaarheid al geen ééndimensioneel begrip, gelet op de differentiatie van dit begrip in stabiliteit, interbeoordelaarsovereenstemming en test-stabiliteit, laat staan wanneer dit begrip in kwantitatieve zin geconcretiseerd wordt aan de hand van gegeven cijfers.

7. Analyse-culturen

Een andere implicatie van de in tabel 2 en 3 gegeven voorbeelden is, dat een cijferschaal op zichzelf niets dwingends voorschrijft over hoe we de daarop gegeven cijfers moeten interpreteren: een nominale en ordinale interpretatie is toegestaan, en soms ook een interval-interpretatie. Afhankelijk van die interpretatie varieert het beeld wat we van 'de' betrouwbaarheid van opstelbeoordelaars voorgeschoteld krijgen. Een onderzoeker die opstelcijfers uitsluitend op nominaal niveau analyseert, presenteert, bewust dan wel onbewust, een veel negatiever beeld (en in ieder geval een onvolledig beeld) van die betrouwbaarheid dan een

onderzoeker die bij zijn analyse ook het ordinale karakter van opstelcijfers betrekt. Het is deze 'analyse-vrijheid', inherent aan elke cijferschaal, die door sommige onderzoekers van opstelbeoordeling willens en wetens is misbruikt ten behoeve van 'propagandistische' doeleinden.

Een voorbeeld. In de jaren zestig deed Schröter onderzoek naar de betrouwbaarheid van opstelbeoordeling. In totaal 1113 leraren Duits beoordeelden 617 opstellen, met dien verstande dat elk opstel door gemiddeld 18 docenten werd nagekeken. Per opstel ging Schröter na, wat het verschil was tussen het hoogste en het laagste daaraan toegekende cijfer. Door de opstellen op deze manier te analyseren wordt niet alleen een *extreem* beeld van de mate van overeenkomst tussen beoordelaars geschetst (immers, uitsluitend de twee meest extreme cijfers worden in de beschouwing betrokken, terwijl alle daartussen liggende cijfers worden veronachtzaamd), maar bovendien een beeld waarin het accent op het gebrek aan overeenstemming ligt. En, last but not least, het geschetste beeld van de betrouwbaarheid van opstelbeoordeling is ook nog eens verre van volledig.

Het is eenvoudigweg niet mogelijk, zoals eerder betoogd, om alle denkbare vormen van konsistentie of inkonsistentie bij gegeven beoordelingen in één enkel getal te gieten, één getal dat indicatief zou zijn voor 'de' betrouwbaarheid. Toch wordt in de onderzoekspraktijk van opstelbeoordeling net gedaan alsof dat wel zou kunnen, althans die suggestie wordt sterk gewekt. Onderzoekers in het Angelsaksische taalgebied, en in navolging daarvan de onderzoekers in Nederland, analyseren 'de' betrouwbaarheid vrijwel zonder uitzondering in termen van een correlatiecoëfficiënt, een getal dat varieert tussen de 0 en 1. "Wanneer wij nu van 'betrouwbaarheid' spreken", aldus De Groot (1968: 119), "dan bedoelen wij (...) alleen die verschillen die het gevolg zijn van een niet perfecte correlatie. Men kan ook zeggen: met 'betrouwbaarheid' bedoelen wij die correlatie(coëfficiënt) zelf".

Gegeven onze eerdere conceptuele analyse van betrouwbaarheid (i.c. konsistentie) en gegeven de bovenstaande cijfervoorbeelden van verschillende vormen van konsistentie komt deze opvatting neer op een ontoelaatbare inperking van het begrip 'betrouwbaarheid'. Immers, de door De Groot bedoelde coëfficiënt, voor de berekening waarvan moet worden uitgegaan van de veronderstelling dat de opstelcijfers op *intervalniveau* liggen, houdt alleen rekening met de (in)konsistentie wat betreft de *rangorde* en de *spreiding* van de cijfers - verschillen in bij voorbeeld het niveau of in het percentage onvoldoendes komen daarin helemaal niet tot uitdrukking. Betrouwbaarheid behelst beslist meer dan deze twee typen (in)konsistentie.

Tot welke warwinkel van misverstanden over 'de' betrouwbaarheid van beoordelingen deze visie leidt, mogen de onderstaande twee voorbeelden verduidelijken.

Berekenen we De Groots correlatie voor het linker voorbeeld, dan vinden we als uitkomst 1 (conclusie: A en B oordelen perfect betrouwbaar), terwijl de correlatie in het rechter voorbeeld 0 bedraagt (conclusie: absoluut onbetrouwbaar) (de absolute grootte van de correlatie varieert tussen 0 en 1. Bij 1 is er sprake van een perfecte samenhang, bij 0 van geen enkele). Maar iedereen die beide voor-

beelden aan een nadere beschouwing onderwerpt, moet toch tot de conclusie komen dat hiermee de 'werkelijkheid' ernstig geweld wordt aangedaan.

Tabel 4: Betrouwbaarheid van twee beoordelaars, uitgedrukt in termen van een correlatie

Opstel	A	B	A	B
1	8	5	8	7
2	7	4	7	8
3	6	3	6	7
	r = 1.0		r = .00	

Wordt in Nederland en in de Angelsaksische landen 'de' betrouwbaarheid in het gros van de gevallen als een correlatie opgevat, in het Duitse taalgebied bestaat een heel andere analyse-cultuur. Daar gaan onderzoekers in de regel uit van een nominale interpretatie door te berekenen, wat het percentage opstellen is wat van verschillende beoordelaars exact hetzelfde cijfer heeft gekregen, of door te berekenen hoe groot het percentage van door leraren gegeven cijfers is, wat een 1, een 2, enzovoort heeft gekregen (zie bij voorbeeld Ulshofer, 1963; Weiss, 1965). Het spreekt voor zich dat het beeld van 'de' betrouwbaarheid van opstelbeoordeling wat uit deze berekening van de *unanimiteit* resulteert, heel wat negatiever is dan dat uit de Angelsaskische landen. Sterk uitgedrukt: de analysepraktijk in de Angelsaksische landen geeft door de bank genomen een te optimistisch beeld van 'de' betrouwbaarheid van opstelbeoordeling, die in het Duitstalige landen een te pessimistisch beeld.

Hoe men opstelcijfers analyseert dient idealiter af te hangen van de specifieke functie die een betrouwbaarheidsberekening geacht wordt te vervullen. Voor bepaalde, specifieke situaties is er niets op tegen wanneer onderzoekers opstelcijfers enkel op bij voorbeeld intervalniveau analyseren, en dus 'de' betrouwbaarheid uitdrukken in één enkel getal, de correlatiecoëfficiënt. Koppelt men echter waarde-oordelen en kwalificaties over de ter zake kundigheid van opstelbeoordelaars aan de resultaten van een dergelijke analyse, dan gaat het niet aan een door traditie en cultuur bepaalde eenzijdigheid te betrachten die tot ernstige vertekening kan leiden. De cijfers dienen in een dergelijk geval zowel op nominaal, ordinaal als intervalniveau geanalyseerd te worden.

8. Psychometrie versus pedagogiek

De zo van elkaar verschillende analyse-culturen in de Verenigde Staten en in Duitsland staan overigens niet op zichzelf, maar vormen een integraal onderdeel van de bestaande onderwijskundige oriëntaties in de betreffende landen - die op hun beurt weer het resultaat zijn van historische omstandigheden, traditie,

opvoeding en onderwijs. In Amerika ontstond tussen 1910 en 1920 de zogeheten 'educational measurement movement' (Smith & Dobbin, 1960: 784), een onderwijskundige beweging waarin sterk de nadruk werd gelegd op het gebruik van scorings-objectieve toetsen en de daarmee onlosmakelijk verbonden begrippen als validiteit en betrouwbaarheid, het afwijzen van schoolcijfers als basis voor bij voorbeeld selectie en het benadrukken van het idee, dat zeker niet elk kind even goed presteert als een ander wanneer het maar genoeg zijn best doet: er zijn verschillen in aanleg, en deze kunnen/moeten wetenschappelijk gemeten worden, via scorings-objectieve tests en toetsen, zodat elk kind dat onderwijs ontvangt dat het beste aansluit bij zijn capaciteiten. In dit alles speelt de correlatie als betrouwbaarheidsmaat een centrale rol: met deze maat wordt immers aangegeven, hoe goed een test/toets kan discrimineren tussen leerlingen met een verschillende aanleg.

De nadruk in de Duitse onderwijskunde, althans die in de jaren zestig en zeventig, ligt daarentegen juist niet op psychometrische kwesties als het gebruik van scorings-objectieve toetsen (integendeel zelfs, deze worden openlijk afgewezen - en tot nu toe met succes), maar op het individu, met zijn specifieke mogelijkheden en beperkingen die een eigen, unieke pedagogische aanpak vereisen. Die beperkingen zijn niet zozeer het resultaat van aanleg, maar vloeien veeleer voort uit het maatschappelijk systeem. Bij de analyse van dit systeem wordt sterk de nadruk gelegd op de onrechtvaardigheid van bestaande instituties en praktijken voor het individuele kind (bij voorbeeld het zitten-blijven). Vanuit deze optiek wordt het wellicht begrijpelijk dat in Duitsland cijfers op nominaal niveau geanalyseerd worden: uitgangspunt vormt het opstel van het individuele kind, waarvan bekeken wordt hoe verschillend (met als implicatie: onrechtvaardig) dit door verschillende beoordelaars nagekeken wordt.

De hier geschetste analyse-culturen in het Angelsaksische en het Duitse taalgebied en de daaraan ten grondslag liggende psychometrische en pedagogische oriëntaties zijn uiteraard abstracties, waarop vele uitzonderingen en nuanceringen mogelijk zijn. Maar dat neemt niet weg dat, algemeen gesproken, uitgangspunten, perspectieven, beschouwingswijzen, aard van de onderzochte problemen en de gehanteerde data-analysetechnieken in de onderwijsresearch van beide landen sterk van elkaar verschillen, en dat de communicatie tussen beide onderwijskundige 'polen' zo niet afwezig, dan toch wel uiterst problematisch is. Vele onderzoekers die, zoals gebruikelijk, zich slechts op één van beide onderwijskundige polen baseren, beseffen niet of onvoldoende hoezeer het beeld dat zij van de betrouwbaarheid van cijfers geven, door traditie bepaald is. En evenmin beseffen ze, hoezeer dat beeld vertekend kan zijn.

Noot

1. Het is geen eenvoudige opgave om de in verschillende landen gehanteerde cijferschalen te achterhalen - een standaardwerk waarin al die schalen netjes

worden opgesomd, ontbreekt. Een deel van de informatie in tabel 1 putte ik uit Lauwerys & Scanlon (1969), voor de landen Rusland, Frankrijk, Oostenrijk en Duitsland. Voor Portugal uit Newcombe (1977) en Schultze (1970), voor Zweden uit Orring (1967), Marklund e.a. (1967), Velema (1959) en Hettema (1965), voor Noorwegen Hove (1958) en Hylla & Wrinkle (1953), voor Italie Newcombe (1977), voor Ierland en Groot-Brittannië Heywood (1977), voor Denemarken Schultze(1969), voor Spanje Newcombe (1977), voor de VS onder andere Geisinger (1982), Thorndike (1972), en voor Nieuw-Zeeland uit Clift & Imrie (1981). Benadrukt moet worden, dat de hierboven aangehaalde auteurs in lang niet alle gevallen precies duidelijk maken, op welk onderwijsniveau (primair, secundair of tertiair) de betreffende schaal gehanteerd wordt. Ook is het vaak niet te achterhalen, waar precies de caesuur ligt. In evidente twijfelgevallen raadpleegde ik de Culturele Dienst van de ambassade, bij voorbeeld die van Frankrijk, Noorwegen, Ierland en Luxemburg. Zo ligt volgens de Culturele Dienst van de Franse ambassade de caesuur op de Franse middelbare scholen (waar men de typische gewoonte heeft een cijfer uit te drukken als een deel van het maximum te behalen punten (i.c. 20), zoals "dix sur vingt" (10/20)), strikt genomen niet vast: op de ene school vinden ze 11/20 onvoldoende (en hogere cijfers voldoende), op de andere 9/20 onvoldoende (en 10/20 tot 20/20 voldoende). Op de lagere school in Frankrijk gebruiken ze doorgaans de schaal 1-10, maar dit systeem is aan het veranderen. Steeds meer scholen gaan gebruik maken van letters.

Bibliografie

Bartels, A. (1947), *75 jaar Middelbaar Onderwijs 1863-1938*. Groningen: Wolters

Clift, J.C. & B.W. Imrie (1981), *Assessing Students, Appraising Teaching*. New York: Halsted Press.

Combag, H.F. & D.N. de Gruyter (1974)(eds.), *Contemporary Issues in Educational Testing*. Den Haag: Mouton.

Davis, F.B. (1964), *Educational measurements and their interpretation*. Bellmont Calif.: Wadsworth.

Ende, J.N. van den (1954a), Cijfers op de middelbare school, in: *Pedagogische Studiën, 31*, p. 69-86.

Ende, J.N. van den (1954b), Cijfers op de middelbare school, in: *Pedagogische Studiën, 31*, p. 112-129.

Fortgens, H.W. (1958), *Schola Latina*. Uit het verleden van ons voorbereidend hoger onderwijs. Zwolle: Tjeenk Willink.

Geisinger, K.F. (1982), Marking systems, in: H.E. Mitzel (ed.) *Encyclopedia of Educational Research*. Fifth ed., Vol. 3, 1139-1149. New York: The Free Press.

Groot, A.D. de (1968), *Vijven en zessen*; Cijfers en beslissingen: het selectieproces in ons onderwijs. Groningen: Wolters-Noordhoff.

Hettema, H. (1965), Het beoordelingssysteem in het Zweedse lager onderwijs, in: *Pedagogische Studiën, 42*, p. 393-399.

Heywood, J. (1977), *Assessment in Higher Education*. London etc.: Wiley & Sons.

Horst, P. (1966), *Psychological measurement and prediction*. Bellmont Calif.: Wadsworth.

Hove, 0. (1958), *An Outline of Norwegian Education*. 2nd. rev. ed. Oslo: The Royal Norwegian Ministry of Foreign Affairs.

Hubrecht, P.F. (1881), *De onderwijswetten en hare uitvoering*. C. Derde afdeling. Lager onderwijs, tweede deel. Den Haag: Stemberg.

Hylla, E. & W.L. Wrinkle (1953), *Die Schulen in Westeuropa*. Bad Neuheim: Im Christian Verlag.

Idenburg, Ph.J. (1964), *Schets van het Nederlandse Schoolwezen*. Groningen: Wolters.

Lauwerys, J.A. & D.G. Scanlon (1969)(eds.), *The World Year Book of Education 1969*. Examinations. London: Evans.

Lord, F.M. & M.R. Novick (1968), *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

Marklund, S. & P. Söderberg (1967), *The Swedisch Comprehensive School*. London: Longmans, Green & Co.

Meuffels, B. (1983), Hoe meer, hoe beter?, in: *Tijdschrift voor Taalbeheersing, 5*, p. 243-256.

Newcombe, N. (1977), Europe at School. *A study of primary and secundary schools in France, West Germany, Italy, Portugal and Spain*. London; Methuen.

Nunnaly, J.C. (1967), *Psychometric Theory*. New York: McGraw-Hill.

Orring, J. (1967), *Die Schule in Schweden*. Eine Übersicht über das Unterrichtswesen bis zur gymnasialen Stufe. Skolöverstyrelsen: Sö-förlaget.

Roozeboom, W.W. (1966), *Foundations of the theory of prediction*. Homewood Ill.: Dorsey.

Schröter, G. (1971), *Die ungerechte Aufsatzzensur*. Bochum: Verlag Kamp.

Schultze, W., (1969)(ed.) *Schools in Europe*. Vol. I: Part A. Weinheim/Berlin: Verlag Julius Beltz.

Schultze, W. (1970)(ed.), *Schools in Europe*. Vol. II: Part A. Weinheim/Berlin: Verlag Julius Beltz.

Smith, A.Z. & J.E Dobbin (1960), Marks and Marking Systems, in: C.W. Harris (ed.) *Encyclopedia of Educational Research*. Third Edition, 783-791. New York: Macmillan.

Stanley, J.C. (1971), Reliability, in: R.L. Thorndike (ed.) *Educational Measurement*. 2nd ed. p. 356-443, Washington D.C.: One Dupont Circle.

Starch, D. (1913), Reliability and distribution of grades, in: *Science, 38*, p. 630-636.

Terwilliger, J.S. (1966), Self-reported marking practices and policies in public secundary schools, in: *National Association of Secundary School Principals Bulletin, 50*, p. 5-37.

Thorndike, R.L. (1972), Marks and Marking Systems, in: G. Bracht e.a. (eds.) *Perspectives in Educational and Psychological Measurement*, p. 164-180. Englewood Cliffs, New Jersey: Prentice-Hall.

Ulshöfer, R. (1963), Welcher Grad von Objectivitat lasst sich bei der Beurteilung Deutscher Aufsatze erreichen?, in: *Der Deutschunterricht, 15*, p. 104-108.

Velema, E. (1959), *De comprehensive school in Zweden en Noorwegen*. Groningen: Wolters.

Weiss, R. (1965), Über die Zuverlassigkeit von Ziffernbenotung bei Aufsätzen, in: *Schule und Psychologie, 9*, p. 257-269.

(manuscript binnengekomen 14 juni 1993)
(manuscript aanvaard 12 juli 1993)