HET HANDHAVEN VAN EENMAAL AANGENOMEN NORMEN BIJ OPEENVOLGENDE OBJECTIEVE TOETSEN

R. F. VAN NAERSSEN

Enige tijd geleden werd een methode beschreven voor de bepaling van de caesuur voldoende/onvoldoende (de Groot, 1964). De hieronder beschreven methode (met nog twee varianten) komt voort uit een enigszins andere probleemstelling. Gaat het bij de Groot om het vaststellen van een norm aan de hand van items, die volgens de docent noodzakelijk te beheersen stof betreffen, hier gaat het vooral om het handhaven van een eenmaal – bijvoorbeeld door middel van de Groot's kernitemmethode – vastgestelde norm bij volgende tentamina (examens, proefwerken). Het belang van normeringsmethoden bij proefwerken, examens, etc. is evident. Zonder deze treedt immers de „Wet van Posthumus" in werking: de cijfers worden soms bewust doch vaker onbewust aangepast aan het toevallige niveau van de groep.

In de eerste twee paragrafen wordt de techniek van de „equating of norms", zoals deze o.a. beschreven is in Gulliksen (1950), enigszins uitgebreid. Daarna wordt een continu-procedure voorgesteld voor het constant houden van de normen.

1. DE HERHAALITEMMETHODE MET LINEAIRE TRANSFORMATIE

Bij academische – en vermoedelijk ook bij vele andere – leervakken treft men vaak de eigenaardige situatie aan, dat een bepaald tentamen twee of meerdere keren per jaar wordt afgenomen, bijvoorbeeld in juni en in september, en waarbij de examengroepen geheel verschillend zijn van niveau. De junigroep is meestal duidelijk intelligenter, ijveriger en beter in het betreffende vak dan de septembergroep. De „Wet van Postumus" – bij elk examen slaagt eenzelfde percentage – werkt hier bijzonder ongunstig. Een van de methoden, die in dergelijke situaties uitkomst kan brengen, is die met de herhaalitems:

Bij de nieuwe items van het tweede tentamen voegt men een aantal items toe, die reeds zijn gebruikt bij het eerste tentamen. Deze items, die samen de „deeltoets" vormen, dienen om de normen, die op een of andere wijze bij het eerste tentamen gelegd zijn, op objectieve wijze over te brengen op het tweede tentamen. Om deze taak goed te kunnen vervullen moet de deeltoets een tamelijk representatieve steekproef vormen van de totale toets. Dat wil zeggen, de items moeten gemiddeld ongeveer

dezelfde moeilijkheid en dezelfde intercorrelaties vertonen als de overige items van het tentamen. Om dit te bereiken kan men het kennisdomein, dat de toets bestrijkt, verdelen in een aantal gebieden. Uit elk gebied laat men door het lot items aanwijzen van elk van een aantal moeilijkheidscategorieën. Bij het tweede tentamen moet men er voor zorgen, dat de nieuwe items qua moeilijkheid en subgebied ongeveer overeenkomen met de herhaalitems. De wenselijkheid betreffende representativiteit van de herhaalitems betekent hier echter geen strenge eis: de representativiteit dient er alleen maar voor om de *vorm van de verdeling* van totale toets en deeltoets zoveel mogelijk identiek te maken. Gemiddelden en standaardafwijkingen van de vier scores mogen aanmerkelijk verschillen: de deeltoetsen verschillen van de totale toetsen omdat het aantal items minder is, eerste en tweede tentamenscores verschillen doordat het niveau van de personen hoger of lager ligt.

Een bepaalde score, gehaald bij het totale eerste tentamen, kan men op de gebruikelijke wijze omzetten in een standaardscore:

$$z_{t_1} = \frac{X_{t_1} - M_{t_1}}{s_{t_1}}, \tag{1}$$

waarin z standaardscore betekent, X ruwe score, M gemiddelde score, s standaardafwijking, de index t totale toets en de index 1 eerste tentamen.

De groep personen, die deze standaardscore gehaald heeft bij de totale toets, zal gemiddeld dezelfde standaardscore halen op de deeltoets, omdat we ervoor gezorgd hebben dat totale toets en deeltoets dezelfde factoren meten: de rangorde van de personen is bij beide toetsen dezelfde, afgezien van toevallige meetfouten. Deze standaardscore kan nu met behulp van gemiddelde en standaardafwijking van de deeltoets omgezet worden in de overeenkomstige ruwe score van de deeltoets.

Een persoon, die bij het tweede tentamen een bepaalde ruwe score haalt op de deeltoets, staat op hetzelfde (kennis-)niveau als de persoon, die bij de eerste deeltoets diezelfde ruwe score haalde, want beide deeltoetsen zijn identiek. De deeltoetsen meten op dezelfde maatstok en kunnen daardoor de schakel vormen tussen de beide totale toetsen.

De ruwe score van de tweede deeltoets kan nu weer omgezet worden in standaardscore. Deze komt overeen met dezelfde standaardscore van de tweede totale toets, en deze kan ten slotte omgezet worden in de ruwe score van het tweede tentamen.

Aldus kan bij elke ruwe score van de eerste totale toets de overeenkomstige ruwe score gevonden worden van de tweede totale toets, met andere woorden: de tweede toets wordt genormeerd op de eerste. Voor

de normering moeten acht basisgegevens berekend worden: de gemiddelden en de standaardafwijkingen van de totale toetsen en van de deeltoetsen. Voor het overige kan het proces echter aanmerkelijk vereenvoudigd worden. We kunnen namelijk een formule afleiden, die de ruwe score van de tweede toets uitdrukt als een lineaire functie van die van de eerste toets:

$$X_{t_2} = AX_{t_1} + B. \tag{2}$$

Heeft men A en B met behulp van de acht basisgegevens berekend, dan kan men bijvoorbeeld voor X_{t_1} de ruwe-scorepunten invullen, die de eindscores (de tentamencijfers) van elkaar afgrenzen, waaronder ook de caesuur voldoende/onvoldoende. Formule 2 geeft dan de overeenkomstige grenspunten op de ruwe-scoreschaal van het tweede tentamen. Hiermee is de normering voltooid en is de Wet van Posthumus doorbroken, evenals bij de kernitemmethode.

De formules, die A en B uitdrukken als functies van de acht basisgegevens, luiden:

$$A = \frac{s_{t_2} s_{d_1}}{s_{d_2} s_{t_1}} \text{ en} \tag{3}$$

$$B = M_{t_2} - \frac{s_{t_2} s_{d_1}}{s_{d_2} s_{t_1}} . M_{t_1} + \frac{s_{t_2}}{s_{d_2}} (M_{d_1} - M_{d_2}), \tag{4}$$

waarin de index *d* slaat op „deel" en 2 op „tweede tentamen". Een afleiding van deze formules, die bovenstaande explicatie practisch op de voet volgt, wordt gegeven in de appendix.

2. DE HERHAALITEMMETHODE MET NIET-LINEAIRE TRANSFORMATIE

Indien de items van een deeltoets gemiddeld veel moeilijker of gemakkelijker zijn dan die van de totale toets, dan wel duidelijk meer of minder met elkaar samenhangen, dan zijn beide verdelingen niet meer gelijkvormig. Het gevolg is dat niet meer op deel- en geheel-toets eenzelfde percentage personen een hogere score gehaald hebben dan een bepaalde standaardscore; met andere woorden, de op bovenstaande wijze gepaarde punten van deeltoets en totale toets komen in werkelijkheid niet meer met elkaar overeen. In dit geval moet men de herhaalitemmethode op iets andere wijze toepassen. Het principe is precies hetzelfde, maar in plaats van via omzetting in standaardscores moet men nu werken via omzetting in percentielscores. Hierbij is het echter niet meer mogelijk gebruik te

maken van een formule. Men zal zijn toevlucht moeten nemen tot een grafische transformatie.

Men zet bijvoorbeeld op millimeterpapier af: op de X-as de ruwe scores van de vier tests, en op de Y-as de percentielscores. De percentielscore, behorende bij een bepaalde ruwe score is, zoals bekend, het percentage personen, dat een lagere score gehaald heeft dan die ruwe score; waarbij men rekent, dat ook nog de helft van de personen, die precies de betreffende score gehaald hebben, onder dit scorepunt thuishoren. Het resultaat is, dat men vier ogiefvormige (S-vormige) krommen heeft getekend, die men zoveel mogelijk vloeiend laat lopen. De krommen van de deeltoetsen D1 en D2 liggen door het geringere aantal items meer links dan die van de totale tentamina T1 en T2. Het probleem is nu van een bepaald punt X1 op de X-as, dat een ruwe score van de eerste totale toets markeert, het overeenkomstige scorepunt X2, ook op de X-as, van de tweede toets te vinden.

De constructie van dit punt zal na het bovenstaande duidelijk zijn:

a. Men trekt door X een vertikale lijn tot deze de ogief T1 snijdt, in „punt-T1".
b. Men trekt een horizontale lijn door punt-T1 tot punt-D1.
c. Men trekt een vertikale lijn door punt-D1 tot punt-D2.
d. Men trekt een horizontale lijn door punt-D2 tot punt-T2.
e. Men trekt een vertikale lijn door punt-T2 tot X2 op de X-as.

Deze vijfstaps-constructie kan men herhalen voor elk punt, waarvan men het overeenkomstige punt van de andere tests wil weten.

Een vereenvoudiging kan worden aangebracht wanneer de scores van de beide totale toetsen ongeveer een gelijkvormige verdeling hebben (de deeltoetsen kunnen dan nog een geheel andere verdeling hebben). Het is hier voldoende slechts twee punten op bovengenoemde wijze te transformeren, waarna men de andere punten kan aanbrengen met een evenredige verdeling, bijvoorbeeld met behulp van evenwijdige lijnen. Voor één van beide punten kan men desgewenst het punt tussen voldoende en onvoldoende kiezen.

De methode met niet-lineaire transformatie heeft het voordeel dat geen enkele eis wordt gesteld aan de representativiteit van de herhaalitems wat betreft hun moeilijkheid.

Voorbeeld: Beide normeringsmethoden werden o.a. toegepast bij een eerstejaarstentamen „Psychologie" aan het Nutsseminarium voor Pedagogiek. In juni 1965 participeerden 108 en in september 38 personen aan het tentamen, dat beide malen uit 48 items bestond, waarvan 25 herhaal-

items. De aan te brengen caesuur voldoende/onvoldoende lag in september volgens de lineaire methode bij een ruwe score van 31,6 en volgens de niet-lineaire (grafische) methode bij 31,0. De grens lag in juni bij een score van 29,5, waarbij 64 % slaagde, terwijl in september slechts 42 % slaagde, namelijk bij gebruik van de niet-lineaire methode (37 % bij de lineaire methode).

3. DE ITEMSTEEKPROEFMETHODE

Bij de herhaalitemmethode mag de deeltoets niet uit te weinig items bestaan, want dan zouden standaardafwijking en gemiddelde van de deeltoetsen te onbetrouwbaar worden; en deze gegevens zitten in de constanten A en B van de normeringsformule. Maar aan de andere kant mag de deeltoets ook niet een te grote portie van de totale toets vormen, want als een dergelijke strategie bekend raakt, dan zullen de studenten zich voortaan concentreren op de toch wel enigszins uitgelekte vragen van het eerste tentamen, in plaats van de gehele stof te bestuderen. Dit moeten laveren tussen Skylla en Charybdis kan men vermijden door de beide deeltoetsen niet volledig identiek te maken maar daarentegen tot werkelijk representatieve steekproeven uit eenzelfde verzameling van items. Aan het trekken van de steekproef, zoals deze in de eerste paragraaf beschreven is, moeten dan *wel* strenge eisen gesteld worden.

De normeringsmethode, die hier voorgesteld wordt – maar die in tegenstelling tot de herhaalitemmethoden nog niet in de practijk beproefd is – komt neer op het volgende:

a. Men deelt de voorraad items in in een aantal groepen, die wat betreft inhoud (kennisgebied) en moeilijkheid dicht bijeenliggen. Voor dit laatste is het noodzakelijk dat de items al eerder zijn afgenomen.

b. Men stelt de aantallen vast van de items van de deeltest, die uit elk van de itemgroepen gekozen moeten worden. Wat de inhoud betreft hangt dit af van wat de docent belangrijk vindt. Bij de verdeling van de moeilijkheidsgraden moet men met twee feiten rekening houden: enerzijds is er een bepaalde optimale moeilijkheidsgraad voor alle items, die afhangt van het aantal alternatieven en van het percentage personen, dat verwacht wordt te slagen, anderzijds is het meestal niet mogelijk voldoende goede items te ontwerpen met een optimale moeilijkheid, zodat men noodgedwongen toch ook gebruik zal moeten maken van gemakkelijker en moeilijker items. Aan de eenmaal vastgestelde aantallen moet men zich houden zolang men de oude normen bij de nieuwe tentamina wil

handhaven. Hierbij kan worden opgemerkt, dat men zich bij de item-steekproefmethode niet hoeft te beperken tot het éénmaal handhaven van normen, doch dat men dit over onbeperkte tijd kan doen.

c. Men stelt de deeltoets samen door uit elke groep het betreffende aantal items op volkomen toevallige wijze te laten aanwijzen. Wanneer het nu een statisch vak betreft is hiermee het proces voltooid. Elke volgende toets is equivalent aan de vorige. Maar meestal zal de docent zijn voorraad items willen aanvullen. Oude items raken op den duur bekend. Bovendien zal de behandelde stof zich geleidelijk wijzigen. En ten slotte is het mogelijk de kwaliteit van de items op te voeren: onduidelijkheden worden opgeheven, alternatieven verbeterd, etc. Dus:

d. Men voegt aan de deeltoets nieuwe items toe, waarvan men wel verwacht, dat ze de betrouwbaarheid van het totale tentamen zullen verhogen, maar waarvan nog geen statistische gegevens bekend zijn. Deze items kunnen dus niet behoren tot de deeltoets, maar anderzijds wil men van deze items wel gegevens vergaren, om ze later aan de verzameling toe te kunnen voegen. Voorts verhogen ze bij elkaar genomen toch ook de betrouwbaarheid; ze dragen wel niet bij tot een preciese normering maar wel tot een betere differentiatie tussen de personen. En ten slotte is men tegenover de studenten wel verplicht ze bij de score mee te rekenen indien ze bij het tentamen zijn afgenomen. Evenals bij de herhaalitem-methode bestaat het tentamen dus uit een deeltoets en een rest, maar hier kan de deeltoets een veel groter percentage van het geheel vormen; eventueel kunnen deel en geheel immers identiek zijn.

e. Men normeert het nieuwe tentamen op het oude met de herhaalitem-berekening. De methode van paragraaf 1 kan gevolgd worden omdat er weinig restitems zijn, waardoor de vorm van de verdeling van de totale toets practisch identiek is aan die van de deeltoets.

f. Men voegt de nieuwe items, die voldoende correleren met het totaal, toe aan de voorraad, na ze geklassificeerd te hebben naar inhoud en moeilijkheid. Als het niveau van de laatste groep studenten afwijkt van de vorige dan levert dit nog een klein probleem op. Een item kan dan niet direct geklassificeerd worden naar zijn p-waarde (het percentage personen dat het item goed heeft beantwoord), maar deze p-waarde moet vergeleken worden met de gemiddelde p-waarde van de deeltoetsitems gegroepeerd naar (oude) moeilijkheidsklasse. Het item wordt geplaatst in de klasse waarvan de gemiddelde p-waarde bij het laatste tentamen het meest overeenkomt met die van het item.

Opgemerkt kan worden, dat de methode *niet* geheel objectief is. Het stellen van de minimum item-totaalcorrelatie is, evenals het oordeel of het item logisch en verbaal acceptabel is, onvermijdelijk subjectief. Verwacht wordt echter dat *deze* subjectiviteit niet te veel invloed zal hebben op de constantie van de norm, zolang men – hoe dan ook – een zeker kwaliteitsniveau van de items weet te handhaven. Verandering van de itemintercorrelaties zou verandering van de spreiding van de scores betekenen, en daarmee van de normen.

4. EEN VERGELIJKING TUSSEN ITEMSTEEKPROEFMETHODE EN KERNITEMMETHODE

Het belangrijkste verschil werd reeds in de aanhef vermeld: de itemsteekproefmethode pretendeert geen normen te *stellen* doch slechts te *handhaven*. Voor deze laatste taak lijkt de itemsteekproefmethode beter geëigend, immers:

a. Het aantal items, dat de norm bepaalt, is groter. Er is geen restrictie betreffende belangrijkheid. Alle items uit de verzameling kunnen deel uitmaken van de deeltest, die de normen overbrengt.

b. Er is geen verschil in inhoud (gemeten factoren) tussen deeltest en totale test. Bij de kernitemmethode is er wel een verschil; de kernitems meten een andere (belangrijker) factor dan de restitems, hetgeen bezwaren kan opleveren. Men kan zich bijvoorbeeld indenken, dat student X, die goed is in de belangrijke factor maar slecht in de andere, deel uitmaakt van een groep, die slecht is in de eerste maar goed in de tweede factor. De groep wordt door de kernitemmethode laag beoordeeld, dus er slagen weinig personen. Onder de afgewezenen zal zich echter X bevinden, die juist goed is in het beantwoorden van de belangrijke kernitems. [1]

c. De moeilijkheidsgraad van de items van de deeltoets is om psychometrische redenen gunstiger dan bij de kernitems, die gemiddeld door meer personen goed worden beantwoord.

1. Op het bezwaar van mogelijke factorische discrepantie bij de kernitemmethode werd mijn aandacht gevestigd door een nog niet gepubliceerde psychometrische studie van Drs. E. E. Ch. I. Roskam. Het probleem der impliciete veronderstellingen is nogal netelig. Zo moet bij de onder 1 en 2 beschreven methoden verondersteld worden dat de groepen personen op alle andere dan de te meten factoren steekproeven vormen uit eenzelfde populatie.

d. De itemsteekproefmethode staat vermoedelijk op een hechtere theoretische basis.

Hier staat natuurlijk tegenover dat de kernitemmethode gemakkelijker is toe te passen. Het zal van de situatie afhangen of het loont om zich de moeite te getroosten van het nauwgezet categorizeren van de items voor het bereiken van constante normen.

5. APPENDIX

$$(5)\quad z_{t_1} = \frac{X_{t_1} - M_{t_1}}{s_{t_1}} = z_{d_1} = \frac{X_{d_1} - M_{d_1}}{s_{d_1}}.$$

$$(6)\quad X_{d_1} = z_{d_1} s_{d_1} + M_{d_1} = \frac{s_{d_1}}{s_{t_1}} (X_{t_1} - M_{t_1}) + M_{d_1} = X_{d_2}.$$

$$(7)\quad \frac{s_{d_1}}{s_{t_1}} (X_{t_1} - M_{t_1}) + M_{d_1} = z_{d_2} s_{d_2} + M_{d_2}.$$

$$(8)\quad z_{d_2} = \frac{\frac{s_{d_1}}{s_{t_1}} (X_{t_1} - M_{t_1}) + M_{d_1} - M_{d_2}}{s_{d_2}} = z_{t_2} = \frac{X_{t_2} - M_{t_2}}{s_{t_2}}.$$

$$(9)\quad X_{t_2} = z_{t_2} s_{t_2} + M_{t_2} = \frac{s_{t_2}}{s_{d_2}} \left[\frac{s_{d_1}}{s_{t_1}} (X_{t_1} - M_{t_1}) + M_{d_1} - M_{d_2} \right] + M_{t_2}.$$

$$(10)\quad X_{t_2} = \frac{s_{t_2} s_{d_1}}{s_{d_2} s_{t_1}} X_{t_1} + M_{t_2} - \frac{s_{t_2} s_{d_1}}{s_{d_2} s_{t_1}} M_{t_1} + \frac{s_{t_2}}{s_{d_2}} (M_{d_1} - M_{d_2}).$$

SAMENVATTING

In aansluiting op de Groot's kernitemmethode voor het vaststellen van normen bij tentamens worden hier drie methoden behandeld om eenmaal vastgestelde normen te handhaven. Eerst worden de lineaire en de niet-lineaire normeringsmethode aangepast aan de tentamensituatie. Ten slotte wordt voor het normale gebruik van objectieve toetsen een methode voorgesteld, die gebaseerd is op het trekken van gelaagde itemsteekproeven en waarbij ook de lineaire normeringsmethode wordt benut.

Literatuurverwijzingen:

DE GROOT, A. D. 1964. *De kernitemmethode voor de bepaling van de caesuur voldoende/onvoldoende.* Paedagogische Studiën 41, 425–440.
GULLIKSEN, H. 1950. *Theory of mental tests*, John Wiley, New York.

Dr. R. F. van Naerssen. Geboren 1922 te Surakarta. Studeerde, na aanvankelijk een militaire loopbaan gevolgd te hebben, van 1951 tot 1957 psychologie aan de Rijksuniversiteit te Leiden. Was daarna 2 jaar verbonden aan het Nederlands Instituut voor Praeventieve Geneeskunde, en 4 jaar aan het Ministerie van Defensie. Promoveerde bij Prof. Dr. A. D. de Groot op een proefschrift getiteld: „Selectie van Chauffeurs – Onderzoekingen ten behoeve van de selectie van chauffeurs bij de Koninklijke Landmacht". Sindsdien verbonden aan het Psychologisch Laboratorium der Universiteit van Amsterdam. Publicaties op het gebied van studietoetsen en psychometrika.