13

Schalen, normen en cijfers

Een toets hoort te worden afgesloten met een heldere en duidelijke presentatie van het toetsresultaat, die de ruimte voor misverstanden tot een minimum beperkt. In het voorliggende hoofdstuk worden manieren besproken waarmee dit doel dichterbij kan worden gebracht. We gaan we ervan uit dat de antwoorden van de persoon op de items op papier zijn gecodeerd als itemscores en dat we dus de beschikking hebben over een vector van itemscores, ook kortweg antwoordpatroon genoemd. Zolang het antwoordpatroon niet expliciet met een of enkele kwaliteitsoordelen is samengevat, is het antwoordpatroon op zich niet erg informatief over het niveau van de geleverde prestatie. Zo'n samenvattend kwaliteitsoordeel noemen we een schaalwaarde of liever een cijfer. Het cijfer moet snel een zo nauwkeurig mogelijke indruk geven van het niveau van het resultaat. Voor een correcte interpretatie van het cijfer moet het natuurlijk duidelijk zijn waarvoor de toets valide is. De validiteit van toetsscores is eerder afdoende aan de orde geweest, zodat in de volgende vijf paragrafen aandacht kan worden geschonken aan andere aspecten van het rapporteren van toetsresultaten. In paragraaf 13.1 wordt het schaalniveau van cijfers behandeld. Het schaalniveau van de cijfers, bijvoorbeeld ordinaal of interval, moet worden vermeld, en verantwoord, om te voorkomen dat er onjuiste conclusies aan cijfers worden verbonden. Men kan niet volstaan met alleen te vermelden welk schaalniveau de cijfers hebben. Ook aan de manier waarop dit schaalniveau is bereikt en met welke veronderstellingen hoort aandacht te worden besteed. In paragraaf 13.2. behandelen we cijfers waarmee het niveau van de prestatie gemakkelijk kan worden vergeleken met prestaties in een of meer groepen. In paragraaf 13.3 behandelen we beheersingsschalen. Dit zijn cijferschalen waarmee het niveau van een prestatie wordt weergegeven als de mate waarin een vaardigheid wordt beheerst. De nauwkeurigheid van het cijfer kan op meerdere manieren in de rapportage worden verwerkt. In paragraaf 13.4 worden daarvoor enige suggesties gedaan. Het nemen van beslissingen op grond van cijfers is het onderwerp van paragraaf 13.5. De manier waarop dit gebeurt moet in de rapportage worden verantwoord. Bij de beslissing of een leerling slaagt of zakt voor een examen

moet bijvoorbeeld duidelijk zijn waarom een bepaald cijfer is aangewezen als de laagste voldoende.

13.1 Het niveau van de schaal

Cijfers winnen aan informatieve waarde naarmate de schaal waarop wordt gerapporteerd een hoger meetniveau heeft. In hoofdstuk 2 zagen we dat naarmate het meetniveau hoger is, de verzameling transformaties naar equivalente schalen kleiner is. Stel dat bijvoorbeeld ruwe scores zouden worden gerapporteerd op een schooltoets die wordt afgenomen voordat de leerstof is behandeld en die na de behandeling nog een keer wordt gemaakt. Kees behaalt de scores 24 en 30 en Hendrik 26 en 32. Het ligt voor de hand om te denken dat beide personen evenveel vooruit zijn gegaan. Echter, het schaalniveau van ruwe scores is lager dan intervalniveau. Daarom kunnen deze twee verschillen op verschillende plaatsen van de ruwe scoreschaal niet zonder meer met elkaar worden vergeleken. We zullen hierna evenwel zien dat met een geschikte theorie de interval-informatie die ruwe scores kunnen bevatten zichtbaar gemaakt kan worden. Het meetniveau van cijfers verkrijgen we door een psychometrische theorie of model over het ontstaan van een antwoordpatroon. Zonder enige theorie hebben we van een groep personen alleen hun antwoordpatronen op de toets of, nog erger, op verschillende toetsen. Twee personen met een verschillend antwoordpatroon op dezelfde toets, bijvoorbeeld 111000 en 001110, worden daarom verschillend beoordeeld. We weten echter niet of het eerste antwoordpatroon een betere, een slechtere of een gelijke prestatie weer-spiegelt als het tweede antwoordpatroon. Zelfs is niet duidelijk of het antwoordpatroon 111100 een grotere prestatie weergeeft dan 000111. Alleen als de antwoordpatronen van twee personen op dezelfde toets gelijk zijn, dan worden hun toetsprestaties gelijk beoordeeld. Indien dat niet het geval is dan moeten de oordelen over hun prestaties verschillen. Zonder enige veronderstelling komen we dus met antwoordpatronen niet verder dan een nominale schaal. Twee antwoordpatronen van verschillende toetsen maken natuurlijk geen enkele onderlinge vergelijking mogelijk. Zonder enige verdere veronderstelling over antwoord-patronen is hun informatieve waarde dus zeer beperkt.

In de klassieke testtheorie wordt dit probleem opgelost door simpelweg te stellen dat de toetsprestatie wordt weergegeven door de som van de itemscores of de ruwe score. De persoon wordt gekarakteriseerd met een ware score op de toets en de ruwe score is daarvan een schatter. Hoe hoger de ruwe score des te groter de toetsprestatie. Alle antwoordpatronen met dezelfde ruwe score zijn daarmee equivalent verklaard en de ruwe score geeft ordinale informatie over de toetsprestatie. De twee eerder genoemde

antwoordpatronen 111000 en 001110 vertegenwoordigen voor de klassieke testtheorie dus een gelijke toetsprestatie, en 011101 een hogere. Door deze afspraak is score 4 hoger dan score 3, en score 3 is hoger dan score 2. Echter, het verschil in niveau tussen de scores 2 en 3 en dat tussen 3 en 4 is niet vergelijkbaar. Immers, de ordinale cijfers 2, 3 en 4 zijn equivalent met bijvoorbeeld 1, 2, 100 en ook met 1, 99, 100. Maar toch, een aanzienlijke winst in de informatieve waarde van het ordinale cijfer ten opzichte van alleen het antwoordpatroon. Het is wel vreemd dat door af te zien van de rijke variëteit aan antwoordpatronen, en grote groepen daarvan als equivalent te beschouwen, het niveau van nominaal naar ordinaal stijgt, en dat we dus aan informatie winnen.

Voorwaarde voor de ordinale informatie van ruwe scores is dat ze op dezelfde toets behaald zijn. Scores op verschillende toetsen zijn niet zonder meer vergelijkbaar. Het ligt voor de hand dat een persoon met een ware score 7 op een toets van 10 items, een hogere ware score heeft op een toets van 20 ongeveer even moeilijke items. Dat zal ongeveer 14 zijn. Voor het probleem van de onderlinge vergelijkbaarheid van scores op verschillende toetsen zijn in het kader van de klassieke testtheorie vele equivaleringsmethoden ontwikkeld (zie hoofdstuk 8).

De introductie van itemresponsmodellen in de psychometrie kan als een belangrijke kwaliteitsimpuls worden beschouwd. We vatten de voordelen van de latente variabele in een itemresponsmodel ten opzichte van de ware score in de klassieke testtheorie nog eens kort samen. Om te beginnen is de waarde van de latente variabele exclusief gekoppeld aan de persoon en niet afhankelijk van de toets zoals de ware score. De toets waarmee de latente vaardigheid wordt geschat, is niet van belang voor de interpretatie van de waarde van de schatter maar alleen voor de nauwkeurigheid daarvan. Voorwaarde is wel dat de items alle-maal afkomstig zijn uit dezelfde verzameling gecalibreerde items of itembank. De geschatte vaardigheden van personen die zijn geschat met hun toetsresultaten op verschillende toetsen uit zo'n verzameling zijn direct vergelijkbaar. Bovendien is het bereikte meetniveau hoger dan het ordinale niveau van de toetsscore. Hoe moeten we begrijpen dat het ordinale niveau van de ruwe score wordt verhoogd naar het intervalniveau van de latente variabele? In de eerste plaats is er het formele argument dat alleen lineaire transformaties van de latente schaal equivalent zijn met de gekozen latente schaal. In de tweede plaats volgt hieruit de meer informele interpretatie dat een bepaalde verhoging van de latente vaardigheid overal op de schaal dezelfde interpretatie toelaat. Gegeven (een verhoging van) de latente vaardigheid kennen we van ieder item (de verandering van) de verdeling van de itemscores, en daarmee bijvoorbeeld ook (van) de verwachte itemscore. Het lijkt erop dat we daarmee niet erg veel opschieten. De itemscores zijn immers van ordinaal

niveau. Lood om oud ijzer dus? We proberen hierna aan te tonen waarom deze vraag ontkennend moet worden beantwoord.

Eerder gaven we het voorbeeld dat de itemscores 1, 2, 3 equivalent zijn met 1, 2, 100, maar ook met 1, 99, 100. Intuïtief voelt iedereen wel aan dat hiermee informatie in de item-scores wordt genegeerd. Bij de introductie van itemscores werd gesteld dat zij in principe ordinaal zijn, evenals toetsscores. Maar toetsconstructeurs kennen bij het opstellen van de scoringsvoorschriften wel degelijk ook informatie toe aan het verschil tussen itemscores. Voor hen zijn 1, 2, 3 en 1, 2, 100 niet hetzelfde. Evenwel, het ontbreekt op het moment van de constructie van de scoringsvoorschriften nog aan een theorie om deze verschillen tussen itemscores meettheoretische betekenis te geven. Daarom kunnen itemscores op dat moment alleen nog maar ordinaal worden geïnterpreteerd. Niet omdat itemscores geen interval-informatie bevatten, maar omdat die er nog niet kan worden uitgehaald. Als er vanaf het begin geen informatie in de verschillen tussen itemscores had gezeten, dan had geen enkele theorie die er uit kunnen halen. Itemresponsmodellen, zoals het Raschmodel of OPLM, kunnen de informatie in de verschillen tussen toetsscores zichtbaar maken.

De parameters in het Raschmodel of OPLM zijn van intervalniveau, of, na een exponentiële transformatie van de modelparameters van log-intervalniveau. Schalen die via een transformatie in elkaar over te voeren zijn, bijvoorbeeld log-interval en interval, worden isomorf genoemd (Stine, 1989). Dit betekent dat zij dezelfde informatieve waarde hebben. Wanneer voor een verzameling items het Raschmodel geldt, kan een transformatie $\hat{\theta}(r)$ worden vastgelegd van toetsscores naar een variabele θ van intervalniveau. Deze transformatie is maar ten dele bepaald door de keuze van het Raschmodel. De schattingsprocedure voor de itemparameters (CML, MML) en de schattingsprocedure voor de persoonsparameters (ML, WML, EAP) zijn mede bepalend voor deze transformatie van toetsscores naar een latente variabele van intervalniveau. We moeten derhalve concluderen dat, wanneer het Raschmodel geldt, ruwe scores isomorf zijn met een schaal van intervalniveau, en derhalve informatie van dit niveau bevatten. Dit betekent echter ook dat de itemscores interval-informatie bevatten. Immers, kies een willekeurig item. Zij r de score van een persoon op de toets zonder het item. Gegeven de score r, wordt de intervalinformatie tussen score 0 en 1 op het item, zichtbaar gemaakt in het verschil tussen $\hat{\theta}(r)$ en $\hat{\theta}(r+1)$.

De eerstvolgende betekenisvolle verhoging van het schaalniveau wordt verkregen door de introductie van een vast nulpunt. Echter, zolang er geen natuurlijk absoluut nulpunt van vaardigheid of itemmoeilijkheid wordt ontdekt, zal het niveau van de schalen in de psychometrie niet boven het intervalniveau uitstijgen.

13.2 Normschalen

Door het cijfer voor een toetsprestatie te laten afhangen van een vergelijking van deze prestatie met de prestaties van een belangrijke groep personen kan de relatieve waarde van de prestatie beter worden beoordeeld. De vergelijkingsgroep wordt een normgroep of referentiepopulatie genoemd, en een cijferschaal waarop de prestaties van een normgroep zijn af te lezen heet een normschaal. De cijfers op een normschaal noemen we normcijfers ter onderscheiding van de cijfers op basis waarvan de normschaal wordt geconstrueerd. Dit kunnen ruwe of gewogen scores zijn, maar ook latente vaardigheidsschattingen. We veronderstellen dat deze cijfers minimaal van ordinaal niveau zijn.

Voor de constructie van een normschaal moet een zogenaamd normeringsonderzoek worden uitgevoerd. Hiertoe moet in de eerste plaats een normgroep ondubbelzinnig worden afgebakend. Een normgroep is bijvoorbeeld alle kinderen in Nederland in groep 8 die niet hebben gedoubleerd. Het is belangrijk dat een normgroep nauwkeurig is omschreven, zodat precies duidelijk is wie er wel en wie er niet toe behoort. Verder moet zij betekenisvol zijn in relatie tot de toetsresultaten. Als de toets bijvoorbeeld is gericht op het meten van de rekenvaardigheden in groep 5 van de basisschool voor de kerstvakantie, dan kan de normgroep precies deze groep bevatten. Echter, als de normschaal beter interpreteerbaar zou worden door alleen de leerlingen te nemen die niet zijn blijven zitten, dan verdient dit de voorkeur.

Vervolgens vereist de constructie van een normschaal dat de frequentieverdeling van de cijfers in de normgroep wordt geschat. Hiertoe moet een representatieve steekproef uit de normgroep worden getrokken. De schatting van de frequentieverdeling is het uitgangspunt voor een ruime keuze aan normschalen. We bespreken vier hoofdtypen van normschalen: cumulatieve verdelingen, genormeerde lineaire transformaties, genormaliseerde schalen en ontwikkelingsschalen.

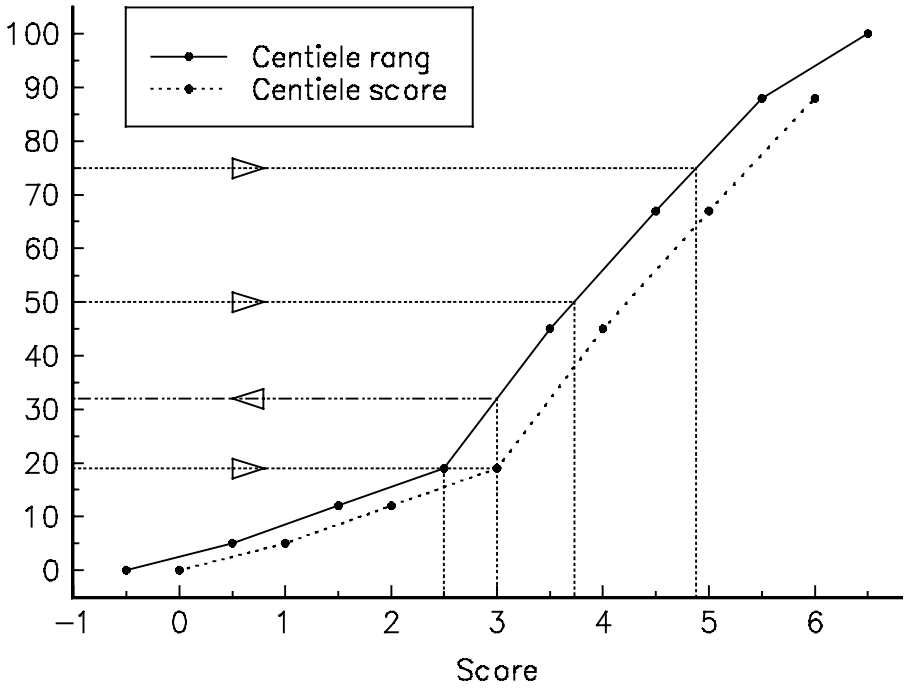
13.2.1 Cumulatieve verdelingen

Afgezien van de onenigheid onder de geleerden over de terminologie is de eenvoudigste normschaal de centiel- of percentielschaal. Uitgangspunt voor een centielschaal is een tabel met (schattingen van) de cumulatieve percentages van de scores op een toets in een normgroep, zoals bijvoorbeeld weergegeven in tabel 13.1.

Tabel 13.1

Cumulatieve percentages van de scores op een toets met zes dichotome items

Scores	Cumulatieve percentages
0	5
1	12
2	19
3	45
4	67
5	88
6	100



Figuur 13.1
Cumulatieve verdelingen en centielschalen bij discrete scores als continue variabele

Op basis van tabel 13.1 zijn er in figuur 13.1 met behulp van lineaire interpolatie twee grafieken voor de verdeling van de scores getekend. De score wordt hier als een continue variabele opgevat en kan derhalve worden gerepresenteerd met een horizontale lijn. De percentages worden op de verticale as afgezet. In figuur 13.1 laten we zien dat voor het tekenen van een verdeling van continue scores meerdere keuzes mogelijk zijn. Het is gebruikelijk in de statistiek om in verband met de zogenaamde correctie voor continuïteit, bijvoorbeeld het percentage 19 bij score 2 op de score-as af

te beelden op 2.5, precies tussen de bijbehorende score en zijn eerstvolgende waarde in. In figuur 13.1 is deze procedure weergegeven met de linker doorgetrokken lijn. Deze lijn wordt gebruikt voor het berekenen van de centiele rang. In figuur 13.1 kan men zien hoe de centiele rang bij score 3 door lineaire interpolatie wordt bepaald. We vinden dat de centiele rang bij score 3 gelijk is aan 19 + (45-19)/2 = 32. De centiele rang wordt ook wel centiele score genoemd (Drenth & Sijtsma, 1990). Hoewel niet de belangrijkste, is een van de oorzaken van de eerder genoemde verwarring het feit dat er in de psychometrie nog een tweede methode wordt gebruikt. Met deze tweede methode beeldt dan het percentage 19 af op de score 3.0. Hieraan wordt wel de naam verbonden van centiel of ook weer centiele score. Een andere benaming is percentiel. Uit tabel 13.1 zien we 19 als cumulatief percentage bij score 2. Dat het centiel 19 bij score 3 hoort, betekent derhalve dat 19% in de normgroep lager scoort dan 3. In de figuur is het enige effect van dit tweede alternatief dat de eerste curve een half scorepunt op de schaal naar rechts is verschoven. Een zekerder interpretatie kan als excuus worden aangevoerd om toch van deze tweede mogelijkheid gebruik te maken. Als het centiel bij score 3 gelijk is aan 19 dan weet men zeker dat men hoger heeft gescoord dan 19% van de normgroep. Bij de centiele rang van 32 bij score 3 is de interpretatie minder duidelijk. Bij een meer gedifferentieerde scoreschaal dan die in het voorbeeld van 0 tot 6 weegt dit voordeel minder zwaar, omdat de afstand tussen de curven voor de centiele score en de centiele rang kleiner is, en gaat het nadeel van een grotere kans op verwarring zwaarder tellen.

Men zegt ook wel dat een score in het zoveelste centiel ligt. Dit woordgebruik verdient enige toelichting. Het eerste centiel loopt van de centielen 0.0 tot 1.0, het tweede van 1.0 tot 2.0, enzovoort. Omdat het centiel van score 2 gelijk is aan 12.0 ligt score 2 dus in het dertiende centiel. Behalve de indeling van de verdeling van de scores in 100 gelijke stukjes, gebruikt men ook andere indelingen. Decielen bijvoorbeeld hebben een vergelijkbare betekenis als centielen, behalve dat de eenheid 10% is in plaats van 1%. In figuur 13.1 delen we de verticale as in tien gelijke delen in. De waarde van het deciel verkrijgen we door de laatste 0 van de getallen langs de verticale as in figuur 13.1 weg te laten. Bij score 2 met centiel gelijk aan 12.0 hoort dan een deciel gelijk 1.2. Omdat het eerste deciel loopt van deciel 0.0 tot 1.0 en het tweede deciel van deciel 1.0 tot 2.0, zegt men ook wel dat score 2, met deciel 1.2, in het tweede deciel ligt. Bij kwartielen is de eenheid 25%. Delen we het centiel van een score door 25 dan verkrijgen we het kwartiel. Het kwartiel bij centiel 12.0 is derhalve 0.48. Ronden we het kwartiel af dan zeggen we dat score 2 in het eerste kwartiel ligt. De algemene benaming voor centielen, decielen enzovoort is quantielen. Het Leerlingvolgsysteem rapporteert bijvoorbeeld in kwartielen per afnamemoment

(normgroep), waarbij het laagste kwartiel nog eens is onderverdeeld in de laagste 10% en de overige 15%. Beelden we bij het verkrijgen van centielen en centiele rangen continue scores af op percentages, voor centiele scores (terminologie van Guilford & Fruchter, 1978), ook wel centiel, centiel punt of centiele rang genoemd, gaan we de andere kant op. Dus van de percentage-schaal naar de continue scoreschaal. We kiezen eerst een percentage *p*, bijvoor-beeld *p* = 75, en zoeken, zoals in figuur 13.1 door lineaire interpolatie, de bijbehorende score. Meestal gebruikt men hiervoor de linker curve voor de centiele rang. Dit is in figuur 13.1 afgebeeld met de lijn die begint bij cumulatief percentage 75. De centiele score bij cumulatief percentage 75 is gelijk aan 4.5 + (75-67)/(88-67) = 4.88. Een andere centiele score is de mediaan. Hiervoor doet men hetzelfde als zojuist bij het percentage 75, maar nu voor het percentage 50. We beginnen dus bij de lijn die begint bij het cumulatieve percentage 50 en vinden dan dat de mediaan gelijk is aan 3.5 + (50-45)/(67-45) = 3.73. Voor de bepaling van de centiele scores wordt ook wel de andere curve genomen.

Uit het voorgaande blijkt dat de naamgeving bij deze schalen in de literatuur onzorgvuldig is. De hoofdbron van de verwarring lijkt te zijn dat er onvoldoende rekening mee wordt gehouden dat een transformatie een relatie tussen twee verzamelingen definieert: een element uit het domein wordt afgebeeld op een element uit de beeldverzameling. Men moet zich derhalve steeds goed realiseren welke twee verzamelingen bij de transformatie zijn betrokken en of bijvoorbeeld de scores op percentages worden afgebeeld of andersom. Hier is hoofzakelijk de terminologie aangehouden zoals gegeven in Guilford en Fruchter (1978). Door de rommelige terminologie bij deze schalen is het geen overbodige luxe om bij een rapportage op een dergelijke schaal zich goed te realiseren wat er is bedoeld. Gelukkig zijn de gehanteerde concepten eenvoudig, zodat de context en de gehanteerde waarden mogelijk de gevraagde helderheid verschaffen. Om misverstanden te voorkomen zou men er goed aan doen termen als centiel, centiele score en centiele rang te vermijden en gewoon te beschrijven hoe de waarden van een schaal zijn berekend.

13.2.2 Genormeerde lineaire transformaties

De algemene gedaante van een lineaire transformatie *s* van een cijfer *r* naar een cijfer *s(r)* is $s(r) = ar + b$. Het cijfer *s* is een normcijfer wanneer de transformatieconstanten *a* en *b* op basis van de frequentieverdeling van *r* zo zijn gekozen dat de prestatie bij een normcijfer gemakkelijk met de prestaties in de normgroep kan worden vergeleken. Omdat met een lineaire transformatie alleen het gemiddelde en de schaaleenheid van

de oorspronkelijke cijferschaal kunnen worden veranderd, worden alleen het gemiddelde en de standaarddeviatie van de frequentieverdeling van de cijfers in de normgroep gebruikt. Een eenvoudig te interpreteren transformatie is die naar standaardscores. De transformatieconstanten *a* en *b* worden zodanig gekozen dat in de normgroep het gemiddelde van de normcijfers *s* gelijk is aan 0 en de standaarddeviatie gelijk is aan 1. Het gemiddelde en de standaarddeviatie van *r* in de normgroep noteren we respectievelijk met μ_r en σ_r. Het is eenvoudig na te gaan dat $a = 1/\sigma_r$ en $b = -\mu_r/\sigma_r$ het gewenste resultaat geven, zodat $s_{(0,1)}(r) = (r-\mu_r)/\sigma_r$. Een standaardscore van $s = 1.0$ betekent derhalve dat men een standaarddeviatie boven het gemiddelde van de normgroep heeft gescoord.

Behalve een gemiddelde van 0 en een standaarddeviatie van 1, zijn vele andere waarden in gebruik, bijvoorbeeld een gemiddelde van 250 en een standaarddeviatie van 10. De waarden voor *a* en *b* die dit bewerkstelligen, verkrijgt men door $s_{(1,0)}$ met 10 te vermenig-vuldigen en er 250 bij op te tellen:

$$s_{(250,\,10)}(r) = 10\frac{(r-\mu_r)}{\sigma_r} + 250 \;\Rightarrow\; a = \frac{10}{\sigma_r},\; b = -\frac{10\mu_r}{\sigma_r} + 250.$$

Toetsscores worden ook vaak lineair getransformeerd naar de nederlandse schoolcijferschaal van 1 tot 10. Ook hier kan de frequentieverdeling van een normgroep aan ten grondslag liggen. Een voorbeeld. Op de cijferschaal wordt de grens tussen voldoende en onvoldoende meestal gelegd bij 5.5. Nemen we aan dat de cijfers worden gerapporteerd met een decimaal. Als men vindt dat 25% van de normgroep hoort te zakken, dan moet de centielscore bij 25%, zeg 17.83, worden afgebeeld op het normcijfer 5.5 - 0.05 = 5.45. Hiermee hebben we het eerste van de twee punten gevonden die de gezochte lineaire transformatie bepalen. Het tweede punt kunnen we vinden door bijvoorbeeld vast te stellen dat niet meer dan 25% van de normgroep een normcijfer 8.0 of hoger mag krijgen. Dan moeten we derhalve zorgen dat centielscore bij 75%, zeg 46.12, wordt afgebeeld op 8.0 - 0.05 = 7.95. De gewenste transformatie krijgen we door het volgende stelsel van twee vergelijkingen op te lossen:
$7.95 = a \times 46.12 + b$ en $5.45 = a \times 17.83 + b$. We vinden dan $a = (7.95 - 5.45)/(46.12 - 17.83)$ en $b = 5.45 - a \times 17.83$. Als de normcijfers niet lager dan 1.0 en niet hoger dan 10.0 mogen zijn, dan rapporteert men 1.0 voor alle cijfers die beneden 1.0 worden afgebeeld en 10.0 voor alle cijfers die boven 10.0 worden afgebeeld.

Een bekend voorbeeld is de ’standaardscore’ die de Eindtoets Basisonderwijs rapporteert voor een leerling (Uiterwijk & Engelen, 1993). Dit zijn geen standaardscores zoals zojuist vermeld, met gemiddelde 0.0 en standaarddeviatie 1.0. De (Eindtoets)standaardscores van een standaardjaar, voor de Eindtoets van 1990 is het standaardjaar 1985, hebben een gemiddelde van 535 en een standaarddeviatie van 10.

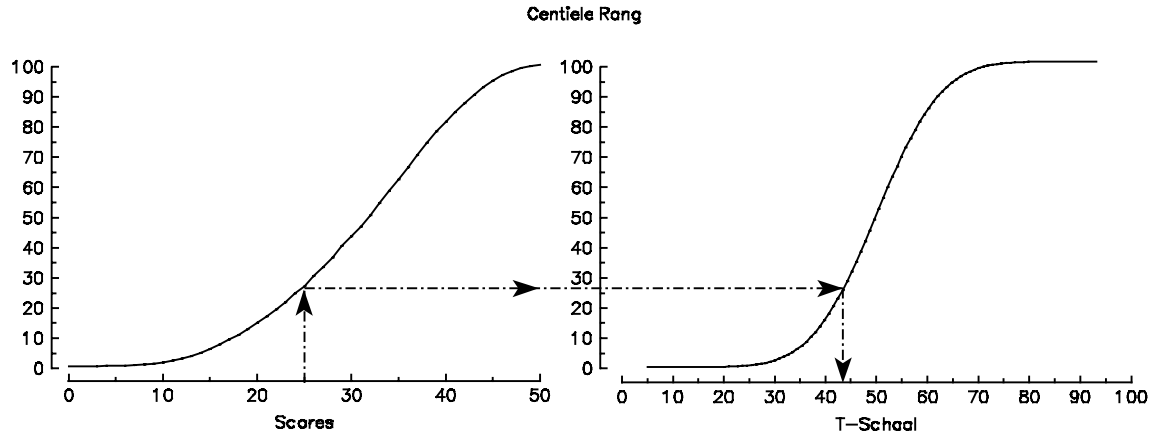
De toetsen na 1985 zijn middels een lineaire equivaleringsprocedure naar de schaal van het standaardjaar getransformeerd.

13.2.3 Genormaliseerde schalen

Tot nu toe werd geen enkele aanname gedaan over de vorm van de verdeling van de normcijfers in de normgroep. Dit lijkt misschien minder relevant, maar het is goed te beseffen dat daardoor de interpretatie van de waarde van een toetsresultaat er flink naast kan zitten. Neem bijvoorbeeld aan dat de cijfers volgens de Beta-verdeling in figuur 13.3 erg scheef naar links verdeeld zijn . De schaal van deze verdeling loopt van 0.0 tot 1.0 en de verdeling heeft een gemiddelde van 0.65 en een standaarddeviatie van 0.23. Stel dat we van een leerling in dit geval een standaardscore van 1.52 zouden rapporteren ($0.65 + 1.52 \times 0.23 \approx 1.00$, dus hoger kan niet). Over het algemeen zal dit worden geïnterpreteerd, weliswaar onterecht maar toch met de normale verdeling in het achterhoofd, als een goede prestatie, behorend tot het hoogste deciel. Deze interpretatie is weliswaar niet onjuist, maar miskend dat de prestatie tot het hoogste centiel van de Beta-verdeling behoort. Deze onjuiste interpretatie wordt vermeden door een genormaliseerde schaal te kiezen. De cijfers op een genormaliseerde schaal zijn verdeeld volgens de normale verdeling. Niet omdat de vaardigheid op de toets zo verdeeld zou zijn in de normgroep, maar eenvoudig omdat de schaal zo is geconstrueerd. Bijvoorbeeld, op een genormaliseerde standaardschaal betekent 1.52 dat precies 94% van de normgroep gelijk of lager scoorde. Het hoogste centiel op een genormaliseerde schaal is pas bereikt bij een cijfer 2.62. Bovendien is het aardige van een aanname over de vorm van de verdeling, dat daarmee een intervalschaal wordt gecreëerd, wanneer men daarbij tenminste ook een dichtheidsfunctie veronderstelt. Een ééndimensionale verdeling en een daarbij behorende dichtheidsfunctie veronderstellen noodzakelijkerwijs een lengtemaat op de intervallen van zijn domein. Was dat niet het geval, dan zou de dichtheids-functie niet zijn gedefinieerd. De dichtheidsfunctie is immers de afgeleide van de verdeling naar de maat op het domein. Wanneer het niveau van de oorspronkelijke cijfers niet van intervalniveau is, dan is men vrij om een dergelijke aanname te maken omdat zij op geen enkele manier getoetst en verworpen kan worden. Wanneer de oorspronkelijke schaal wel van intervalniveau is, dan is een hypothese over de verdeling wel toetsbaar. We komen hier nog op terug.

In de sociale wetenschappen gebruikt men graag de normale verdeling. Het hoeft ons dan ook niet te verbazen dat vaak wordt verondersteld dat de normcijfers normaal zijn verdeeld met een vrij te kiezen gemiddelde en standaarddeviatie (μ,σ). Veel

voorkomende genormaliseerde schalen zijn de T-schaal, de C-schaal en de Stanine schaal. Voor de T-schaal kiest men $(\mu, \sigma) = (50,10)$, voor de C-schaal en de Stanines $(\mu, \sigma) = (5,2)$. Voor deze laatste twee schalen komt daar nog bij dat alleen gehele getallen worden gerapporteerd. Voor de C-schaal lopen die getallen van 0 tot 10. Stanines zijn identiek aan de C-schaal, behalve dat de C-schaalcijfers 0 en 1 worden samengevoegd tot Stanine 1 en de C-schaalcijfers 9 en 10 tot Stanine 9.



Figuur 13.2

Bepaling van T-schaal bij een toetsscore. Links staan centiele rangen van een referentiepopulatie bij de toetsscores. Rechts is de cumulatieve normale verdeling N(50, 10^2) weergegeven

Het algemene principe voor de berekening van genormaliseerde schalen is als volgt (zie figuur 13.2). Zij G een cumulatieve verdelingsfunctie, bijvoorbeeld de cumulatieve normale verdeling $N(50, 10^2)$. Dan is het genormaliseerde cijfer $n(r)$ van cijfer r met centiele rang $c(r)$ gelijk aan $n(r) = G^{-1}(c(r))$, dus $G(n(r)) = c(r)$. Oftewel de cumulatieve verdelingsfunctie met als argument het genormaliseerde cijfer is gelijk aan de centiele rang van het cijfer. De linker grafiek representeert de centiele rangen bij de cijfers, de functie $c(r)$. De rechter grafiek toont de cumulatieve normale verdelingsfunctie met gemiddelde 50 en standaarddeviatie 10. In figuur 13.2 is de bepaling van de T-score bij cijfer 25 grafisch weergegeven. Daartoe zoeken we eerst de centiele rang p_{25} bij cijfer 25. Dit is weergegeven in het linkerdeel van figuur 13.2. Daar kunnen we zien dat p_{25} ongeveer gelijk is aan 26. Vervolgens zoeken we bij p_{25} de T-schaalwaarde, zoals weergegeven in het rechterdeel van figuur 13.2. Daar zien we dat de T-schaalwaarde bij score 25 ongeveer gelijk is aan 43.

T-schaalcijfers worden niet altijd gebaseerd op centiele rangen. Men gebruikt ook wel het cumulatieve percentage van de toetsscore lager dan de betreffende toetsscore (centiel), en soms ook wel inclusief de betreffende toetsscore zelf.

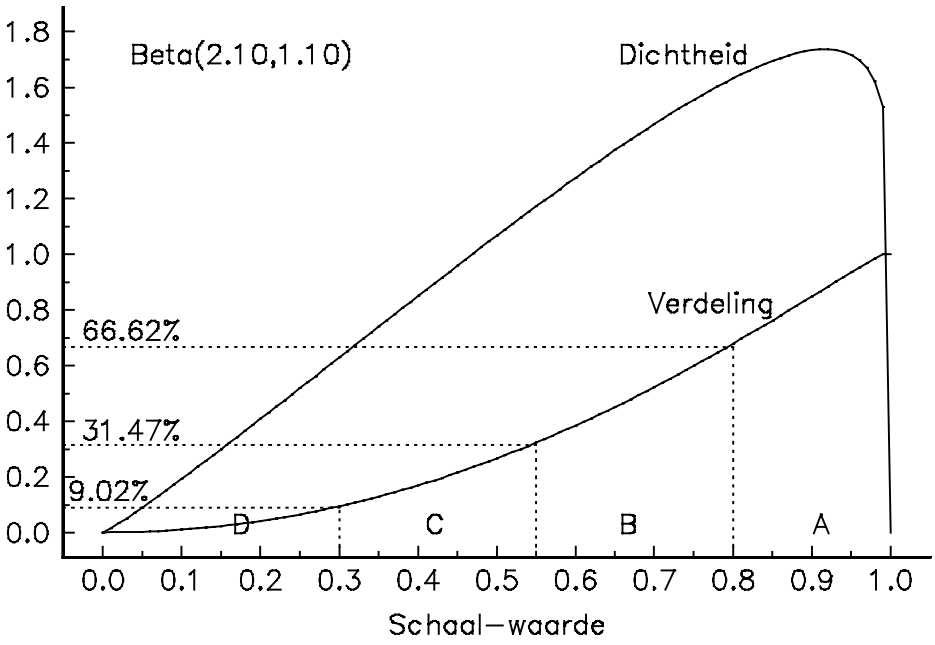
Tabel 13.2

Bovengrenzen van genormaliseerde standaardscores en centiele rangen voor de C-schaal

C-schaal waarde	Genormaliseerde standaardscore	Centiele rang
0	-2.25	1.2
1	-1.75	4.0
2	-1.25	10.6
3	-0.75	22.7
4	-0.25	40.1
5	0.25	59.9
6	0.75	77.3
7	1.25	89.4
8	1.75	96.0
9	2.25	98.8
10	∞	100.0

In tabel 13.2 zijn de bovengrenzen van de centiele rangen opgenomen voor de C-schaal. Het C-schaalcijfer van een cijfer wordt gevonden bij de kleinste bovengrens groter dan de centiele rang van het cijfer. Als bijvoorbeeld cijfer 25 een centiele rang heeft van 26.5, dan is het C-schaalcijfer voor cijfer 25 gelijk aan 4, omdat 40.1 de kleinste bovengrens is groter dan 26.5. De onderlinge afstand tussen C-schaalcijfers komt overeen met 0.50 standaarddeviatie. Natuurlijk kunnen we de C-schaalcijfers door een lineaire transformatie afbeelden naar een schaal met gemiddelde 0 en standaarddeviatie 1.0. Daartoe trekken we van het C-schaalcijfer 5 af en delen het resultaat door 2. We verkrijgen dan de genormaliseerde versie van de eerder genoemde standaardscores. Genormaliseerde standaardscores zijn per definitie normaal verdeeld. Daarentegen heeft de verdeling van de eerder genoemde standaardscores dezelfde vorm als die van de oorspronkelijke cijfers. Let wel dat tabel 13.2 de bovengrenzen van de genormaliseerde standaardscores bij de C-schaal bevat. Bij de C-schaalwaarde 5 hoort bijvoorbeeld een genormaliseerde standaardscore van 0.0, de bovengrens is echter 0.25. Een niet onbelangrijk voorbeeld van een genormaliseerde schaal is de deviatie-IQ-schaal. Dit IQ is in iedere normgroep (leeftijdsgroep) normaal verdeeld met een gemiddelde van 100 en een standaarddeviatie van 15. De gemiddelde intelligentie, voor zover gemeten door de Stanford-Binet IQ-tests, neemt na het vijftiende levensjaar niet meer toe (Linn, 1989). Een willekeurige steekproef uit de populatie van volwassenen en een willekeurige steekproef van vijftienjarigen hebben dezelfde gemiddelde ruwe score op de Stanford-Binet test. Linn vermeldt niet of de variantie boven deze leeftijd

onveranderd blijft. Voor de SON (Snijders-Oomen Non-verbale intelligentietest, 1991) zijn normschalen gepubliceerd voor de nederlandse populatie voor leeftijden van 5.5 tot 16.5 jaar. Deze schalen laten nog een progressie zien tot en met de hoogste leeftijdsgroep.



Figuur 13.3
De Beta-getransformeerde schaal van de Entreetoets

Een vergelijkbare procedure is gevolgd bij de Entreetoets van het Cito (Moelands, 1988). Dit is overigens net als de Eindtoets, een hele batterij van toetsen die samen een groot deel van de leerstof van het laatste jaar van de basisschool dekken. Voor de schalen van de toetsen in de Entreetoets werd echter geen normale verdeling gekozen maar de Betaverdeling B(2.10, 1.10). Voor de verdeling in figuur 13.3 kan men de cijfers 0.0 t/m 1.0 langs de verticale as lezen als centiele rangen gedeeld door honderd. Figuur 13.3 is dan een Beta-equivalent van het rechterdeel van figuur 13.2. We hebben hier dus geen genormaliseerde schaal maar een 'Beta-getransformeerde' schaal. Deze verdeling werd gekozen omdat zij redelijk aansloot bij de wens de totale schaal in vier hoofdcategorieën (A, B, C, D) in te delen die respectievelijk de 30% hoogste scoorders bevat (A), de middelste 40% (B), 20% lagere (C) en de 10% laagste (D). Verder wenste men de Beta-schaal in te delen in 20 intervallen ter grootte van 0.05, zodanig dat de verdeling van deze intervallen over de hoofd categorieën D t/m A gelijk is aan 6, 5, 5, 4. Hoofdcategorie D bevat de eerste 6 van deze eenheden, B en C ieder 5, en

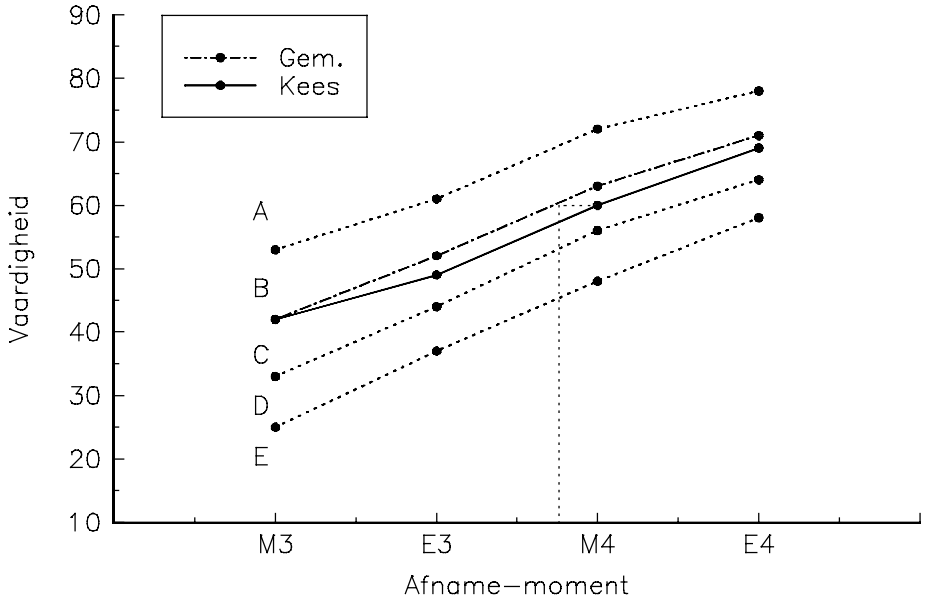
A de hoogste 4 (17 t/m 20). De genoemde B(2.10, 1.10) voldeed ongeveer aan deze merkwaardig gedetailleerde wensen. Zoals in figuur 13.3 te zien is, leidt deze transformatie tot een aan de onderkant enigszins uitgerekte, maar overigens bijna lineaire transformatie van de percentielschaal. Door deze aan de onderkant gerekte schaal wordt bereikt dat de cijfers op de twintigpuntsschaal vooral differentiëren tussen zwakkere leerlingen: de eerste elf van de twintig punten wordt verdeeld onder de 31.5% laagst scorende leerlingen. Dit laatste is in overeenstemming met het doel van de Entreetoets om vooral te letten op het lagere deel van de schaal: het detecteren van zorgwekkend lage niveaus in het vaardigheidsprofiel van een leerling.

13.2.4 Ontwikkelingsschalen

De intelligentietests van Binet-Simon (Drenth & Sijtsma, 1990) rapporteerden intelligentie als het quotiënt van mentale leeftijd en kalenderleeftijd maal 100: het intelligentiequotiënt. De mentale leeftijd van een kind met cijfer *r* is de leeftijdsgroep met gemiddeld cijfer *r*. De mentale leeftijd is een voorbeeld van een ontwikkelingsschaal. De constructie van een ontwikkelingsschaal vereist grootschalig onderzoek. Men kiest een normgroep met een voldoende range aan leeftijden, bijvoorbeeld de populatie van het basisonderwijs. Men groepeert de leeftijden in deze normgroep in een aantal categorieën. Bijvoorbeeld de leeftijdscategorie 6 bevat alle leerlingen die op het moment van de toetsafname tussen de 5½ en 6½ jaar oud zijn. De leeftijdsgroep 6 zouden we dan een deelnormgroep kunnen noemen. Uit alle leeftijdsgroepen trekt men een representatieve steekproef. Voor iedere leeftijdsgroep wordt het gemiddelde cijfer bepaald, eventueel de mediaan. Vervolgens wordt bijvoorbeeld door lineaire interpolatie een regressiefunctie van de cijferschaal naar de leeftijdsschaal verkregen. Deze regressiefunctie geeft bij ieder cijfer een leeftijdsaanduiding, bijvoorbeeld bij cijfer 25 de leeftijd 5;7 jaar. Zou men de Binet-Simon manier van rapporteren kiezen en stel dat het kind met score 25 de leeftijd heeft van 5;5 jaar, dan is de quotiëntscore (5 7/12) : (5 5/12) x 100 = 103.

Het rapporteren van toetsresultaten op een ontwikkelingsschaal is tamelijk problematisch en de rapportage op een quotiëntschaal dus ook. Verschillende vaardigheden kunnen zich met verschillende snelheid ontwikkelen ten opzichte van de spreiding in een normgroep. Het gemiddelde verschil in leesvaardigheid tussen zeven en negen jaar kan bijvoorbeeld maar een standaarddeviatie op de schaal van zevenjarigen groot zijn, terwijl dit voor rekenen gelijk zou kunnen zijn aan twee standaarddeviaties. Rekenen is voor achtjarigen bijvoorbeeld al een standaarddeviatie

hoger. Dergelijke verschillen in ontwikkelingssnelheid leiden tot oneven-wichtigheid in de rapportage. Neem een kind van zeven jaar dat zowel op een leestoets als op een rekentoets een standaarddeviatie boven het gemiddelde van zijn leeftijdsgroep scoort. Dit kind krijgt voor lezen het leeftijdscijfer 9 en voor rekenen een 8. Dit wekt de indruk dat het kind met lezen meer presteert dan met rekenen. Het is gemakkelijk dit voorbeeld zo extreem te maken dat men wel moet concluderen dat deze indruk onterecht is.



Figuur 13.4
Het grafische LVS rapport van de ontwikkeling van Kees

De bovengenoemde problemen kunnen worden opgelost door een rapportagevorm te vinden waarin zowel de ontwikkeling van de normgroep, als de plaats van de persoon in zijn huidige normgroep tot zijn recht komt. Nog beter is het wanneer ook de ontwikkeling van de persoon kan worden weergegeven. Deze vorm heeft men in het Leerlingvolgsysteem (Jansen e.a., 1992) weten te realiseren, hoewel een adequate schatting van de ontwikkeling van de persoon technische problemen oplevert (zie hoofdstuk 10). Figuur 13.4 laat het grafische rapport zien van de prestaties van Kees op de rekentoetsen voor de afnamemomenten Medio Groep 3 (M3) tot en met Eind Groep 4 (E4). De gebieden A, B en C bevatten de drie bovenste kwartielen van de centielschaal, waarvan A (boven de bovenste lijn) het hoogste deel. D en E bevatten samen het onderste kwartiel, waarvan E de laagste 10%. Voor Kees zijn in de grafiek niet alleen zijn positie binnen zijn groep duidelijk, maar ook zijn 'Groepsequivalenten'. Bijvoorbeeld, het snijpunt van de horizontale lijn door zijn positie op M4 met de lijn

voor het gemiddelde levert zijn Groepsequivalent op M4. Dit ligt ongeveer op een kwart van de afstand (E3, M4) onder M4 (figuur 13.4). Nemen we aan dat de tijd tussen E3 en M4 een half leerjaar bedraagt, dan zou men kunnen zeggen dat hij op M4 een vaardigheid heeft die gelijk is aan het gemiddelde in de normgroep van ongeveer een achtste leerjaar geleden, of dat hij op M4 ten opzichte van het gemiddelde in zijn groep een achtste leerjaar achterloopt. De bepaling van dit snijpunt lukt natuurlijk niet voor alle gevallen. Voor een leerling die op M3 beneden het gemiddelde scoort, bestaat zo'n snijpunt niet. Dit is echter een probleem dat aan alle ontwikkelingsschalen kleeft en is niet uniek voor de schalen van het Leerling-volgsysteem.

13.2.5 De nauwkeurigheid van normschalen

Normschalen zijn gebaseerd op een schatting van de frequentieverdeling in een normgroep. De schatting van deze frequentieverdeling is natuurlijk behept met steekproeffouten. Met name wanneer er nonrespons te verwachten is die samenhangt met de te normeren schaal kan de schattingsfout van de frequentieverdeling aanzienlijk zijn. Wanneer bijvoorbeeld in een normeringsonderzoek van een rekentoets vooral de slecht presterende scholen niet meedoen, dan zal de resulterende normschaal een te somber beeld geven van de prestaties van de leerlingen. De schatting van het gemiddelde cijfer van de toets in de normgroep zal dan bijvoorbeeld hoger uitvallen dan in werkelijkheid het geval is. Een leerling die in werkelijkheid gemiddeld scoort, zal een normcijfer krijgen dat aangeeft dat hij beneden het gemiddelde presteert. De steekproeffouten kunnen worden verkleind door een gestratificeerde steekproef te trekken waarin bijvoorbeeld de percentages jongens en meisjes gelijk zijn aan die in de normgroep. Een belangrijke overweging voor de keuze van stratificatievariabelen is de beschikbaarheid van de verdeling uit een andere bron, bijvoorbeeld het Centraal Bureau voor de Statistiek (CBS). De tweede overweging voor de keuze van een stratificatievariabele is een verwachte samenhang met een dreigende nonrespons. Wanneer de stratificatievariabelen aan beide voorwaarden voldoen, dan kan de representativiteit van de steekproef en de mogelijke invloed van nonrespons worden ingeschat en eventueel worden gecorrigeerd. Angoff (1971) bespreekt overwegingen rond steekproef-trekking en vereiste nauwkeurigheid van normschalen. Zijn aanbevelingen komen erop neer dat de steekproeffouten van de normschaal ten opzichte van de meetfouten van de normcijfers verwaarloosbaar horen te zijn. In de rapportage over een normschaal mag een verslag over de representativiteit van de steekproef niet ontbreken. Hierin wordt de verdeling van belangrijke

achtergrondvariabelen in de steekproef vergeleken met de verdeling in de normgroep, voor zover bekend uit bijvoorbeeld CBS-publikaties.

13.3 Beheersingsschalen

Hoewel voor veel schoolvakken een normcijfer een belangrijke indicatie is voor het niveau van de prestatie, zijn er ook situaties waar het er minder toe doet welk percentiel van een normgroep aan de prestatie van een persoon gehecht moet worden. Piloten moeten een vliegtuig veilig aan de grond zetten. Het doet er niet toe of 90% van de kandidaten daartoe in staat is of maar 1%. Zoiets geldt ook voor loodgieters en bruggenbouwers. Hun produkten moeten gewoon voldoen aan de eisen die daaraan gesteld moeten worden. In dit soort gevallen geeft een normschaal niet de gewenste informatie. Een normcijfer geeft geen inzicht in het niveau van de prestatie. Hoe goed kan een persoon rekenen die een centiel van 80 scoort in groep 4? Hoeveel procent van de aftreksommen met getallen van vier cijfers maakt zo'n leerling goed? Hoeveel procent van de deelsommen? Dit type informatie wordt gegeven door een beheersingsschaal. Het kan zowel gaan om een indicatie van de huidige beheersing, alsook voor een te verwachten beheersing in de toekomst. Beheersingsschalen geven een cijfer betekenis door dit te transformeren naar een maat die aangeeft in welke mate de persoon een leerstofonderdeel beheerst of zal beheersen. We noemen deze maat verder het beheersings- cijfer. De psychometrie van beheersingsschalen werd met name in de jaren 70 ontwikkeld. Men noemt beheersingsschalen ook wel criterium-georiënteerde schalen (Van der Linden, 1982).

Het eerste probleem bij de constructie van een beheersingsschaal is het afbakenen van het leerstofdomein. Zolang hierover onduidelijkheid bestaat kan aan geen enkel beheersingscijfer een ondubbelzinnige betekenis worden gegeven. Het probleem voor de afbakening is de veelal grote keuze aan invalshoeken en begrenzingen. Deze kunnen leerstofgericht zijn of gebaseerd zijn op cognitief psychologische onderscheidingen. Ook het onderscheid tussen kennis, toepassing en inzicht wordt hier vaak gehanteerd. Daar komt nog bij dat vele van deze onderscheidingen erg vaag zijn. Het lijdt bijvoorbeeld geen twijfel, dat een toepassing toch vaak ook inzicht vereist. En kan een leerling inzicht hebben zonder dat deze evidente toe-passingen ziet? Ook een inhoudelijke afbakening laat echter vaak meerdere interpretaties toe. Zo hebben bijvoorbeeld de schoolvakken aardrijkskunde en wiskunde de laatste decennia grote veranderingen ondergaan. Maar niet duidelijk is of leerstofonderdelen die nu expliciet tot de leerstof worden gerekend, er tevoren, impliciet of in de praktijk, ook al niet toe behoorden.

Het probleem van de afbakening van een leerstofdomein is concreter wanneer men niet alleen over tamelijk abstracte leerdoelen praat, maar ook over een concrete verzameling items. Eerst maakt men afspraken waarover de items zullen gaan, maar daarna kan worden volstaan met de vraag of een bepaald item nu wel of niet tot het domein kan worden gerekend. Bovendien kan men lacunes in de itemverzameling opsporen, daar weer items bij schrijven, enzovoort. Zo kan een itembank ontstaan waar over men het gemakkelijker over eens kan worden dat hiermee een leerstofdomein kan worden gemeten. Een groot voordeel van de constructie van een dergelijke itembank is de duidelijke betekenis die daarmee aan een beheersingscijfer kan worden gehecht. Men kan bijvoorbeeld rapporteren welk percentage van deze verzameling naar verwachting correct beantwoord zal worden. Binnen de klassieke testtheorie is dit zonder groot verlies van nauwkeurigheid (generaliseren) echter niet goed mogelijk. Daar beperkt men zich vaak tot het percentage van de items in de toets zelf. Als schatter van dit percentage neemt men dan eenvoudig $r/m \times 100$ %, waarin r de toetsscore en m de maximaal te behalen score op de toets. Deze oplossing heeft het bezwaar dat twee verschillende toetsen uit dezelfde itemverzameling kunnen verschillen in moeilijkheid. Een percentage beheersing op een gemakkelijke toets is dan een overschatting van het percentage beheersing van de itemverzameling en een percentage op een moeilijke toets een onderschatting. Binnen het kader van IRT vervalt dit bezwaar doordat voor iedere schatting van de latente vaardigheid het verwachte percentage correct op de complete gecalibreerde itemverzameling kan worden berekend.

Ook de Eindtoets rapporteert beheersingscijfers. Wegens het ontbreken van een gecalibreerde itembank hebben deze beheersingscijfers echter alleen betrekking op de gemaakte toetsen. Men rapporteert het percentage items uit de toets dat goed is beantwoord. Bij het Leerlingvolgsysteem wordt een fraai grafisch overzicht gepresenteerd van de beheersingsgraad van een leerling op de vaardigheidsschaal, waarop ook het interval tussen 50% en 80% kans op correct voor een selectie van de items is aangegeven.

13.4 Het rapporteren van meetnauwkeurigheid

Voor een goede interpretatie van cijfers is het belangrijk als de nauwkeurigheid gemakkelijk is af te lezen. Een algemeen raamwerk hiervoor wordt beschreven in Kolen (1986, 1988). Men kiest een cijferstap h en een $\gamma \times 100$ % betrouwbaarheidsinterval. Vervolgens wordt een transformatie $s(r)$ van de cijfers r geconstrueerd zodat bij iedere

s het interval $[s - h, s + h]$ een tweezijdig $\gamma\times100$ % betrouwbaarheidsinterval is. Kiest men bijvoorbeeld $h = 1.0$ en $\gamma = 0.50$, dan is voor een getransformeerd cijfer *s(r)* het interval [*s* - 1.0, *s* + 1.0] een 50% betrouwbaarheidsinterval rond *s*.

Als de standaardmeetfout σ_E van de cijfers *r* constant is over het bereik van *r*, dan is de transformatie *s* lineair. De coëfficiënt *b* van de lineaire transformatie *s(r) = ar + b* kan arbitrair worden gekozen terwijl *a* als volgt wordt bepaald. Laat z_γ het getal zijn waarvoor

$$(2\pi)^{-\frac{1}{2}} \int_{-z_\gamma}^{z_\gamma} \exp\left(\frac{-t^2}{2}\right) dt = \gamma, \tag{13.1}$$

dan is $a = h/(\sigma_E z_\gamma)$. Let wel dat het gebruik van (13.1) een normaal verdeelde meetfout veronderstelt.

Als de standaardmeetfout niet constant wordt verondersteld, maar een functie σ_E is van het cijfer *r*, dan wordt het ingewikkelder. Kolen (1986, 1988) behandelt de arcsinus-transformatie (Freeman & Tukey, 1950; Lord & Novick, 1968). De variantie van de arcsinustransformatie van de ruwe score is onder het binomiale of compound binomiale foutenmodel ongeveer constant. Het is op zich een interessant probleem om bij een willekeurige standaardmeetfout als functie van *r* een variantiestabiliserende transformatie te bedenken. Zij daarom de meetfouten van cijfer *r* verdeeld volgens G_r met standaarddeviatie $\sigma_E(r)$. Het meest voor de hand ligt om de functie $\sigma_E(r)$ te zien als een te corrigeren transformatie T^{-1} van de maat van de intervallen tussen de opeenvolgende cijfers *r*. Door de inverse transformatie te nemen kan de variabele standaarddeviatie constant worden gemaakt:

$$T(r) = \int_{r_0}^{r} \frac{1}{\sigma_E(v)}\, dv$$

waarin r_0 een willekeurig cijfer is. Hierna volgt een schets van het bewijs dat de meetfout van $T(r)$ ongeveer constant is. Het kwadraat van de standaardmeetfout van $T(r)$ is

$$\begin{aligned}
\sigma_E(T(r))^2 &= \int_R (T(u) - T(r))^2\, d\,G_r(u) \\
&= \int_R \left(\int_{r_0}^{u} \frac{1}{\sigma_E(v)}\, dv - \int_{r_0}^{r} \frac{1}{\sigma_E(v)}\, dv \right)^2 d\,G_r(u) \\
&= \int_R \left(\int_{u}^{r} \frac{1}{\sigma_E(v)}\, dv \right)^2 d\,G_r(u),
\end{aligned}$$

waarin R het domein van r. Veronderstellen we nu dat de standaardmeetfout $\sigma_E(v)$ voor v 'in de buurt van' r ongeveer gelijk is aan $\sigma_E(r)$, dan blijkt dat

$$\sigma_E(T(r))^2 \approx \int_R \left(\frac{u-r}{\sigma_E(r)} \right)^2 dG_r(u)$$

$$= \frac{\sigma_E(r)^2}{\sigma_E(r)^2} = 1,$$

ongeveer constant is. De uitdrukking 'in de buurt van r' moet men zien in relatie tot G_r. Het 'ongeveer gelijk aan' is in samenhang met 'in de buurt van r' preciezer te maken, maar dat is hier niet zo relevant.

Deze transformatie maakt het mogelijk om voor iedere cijferschaal waarvan de standaardmeetfout bekend is een schaal te construeren volgens het recept van Kolen. Zij bijvoorbeeld het cijfer een schatting $\hat{\theta}$ van de latente vaardigheid θ op een Raschschaal, geschat met een toets met informatiefunctie $I(\theta)$. Dan is de transformatie $T(\hat{\theta})$:

$$T(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}} \sqrt{I(v)}\, dv. \tag{13.2}$$

Is het cijfer de ruwe score op deze toets, dan krijgen we de transformatie $T(r)$:

$$T(r(\hat{\theta})) = \int_{-\infty}^{\hat{\theta}} \frac{1}{\sqrt{I(v)}} dr(v), \tag{13.3}$$

waarin $r(\hat{\theta})$ de verwachte score op de toets voor latente vaardigheidsschatter $\hat{\theta}$. Deze transformatie kan voor toetsen, die aan het Raschmodel voldoen, in plaats van de bovengenoemde Freeman-Tukey arcsinustransformatie worden gekozen. Uiteraard leiden (13.2) en (13.3), als functie van $\hat{\theta}$, tot hetzelfde resultaat. Dit is ook als volgt in te zien. De informatiefunctie is gedefinieerd als:

$$I(\theta) \doteq \frac{\left(\frac{\partial r(\theta)}{\partial \theta} \right)^2}{\sigma_r^2(\theta)}.$$

Omdat in het Raschmodel $I(\theta) = \sigma_r^2(\theta)$, volgt dat $dr(\theta) = I(\theta)\, d\theta$, waarmee de identiteit van (13.2) en (13.3) is aangetoond.

Een reden die kan worden aangevoerd om te kiezen tussen bijvoorbeeld T-schaal, C-schaal of Stanines is de meetnauwkeurigheid. De algemene regel is om met de

rapportage van het cijfer geen grotere nauwkeurigheid te suggereren dan de standaardmeetfout van het cijfer toelaat. Deze enigszins vage regel wordt dan geconcretiseerd tot de vuistregel dat de cijfers moeten oplopen in stappen van ongeveer een standaardmeetfout. Kolen (1986, 1988) wijst erop dat deze procedure niet goed te verdedigen is. Immers, bij toepassing van de vuistregel voegt men dan door afronden maximaal een halve standaardmeetfout toe aan de meetfout, gemiddeld dus ongeveer een kwart van de standaardmeetfout. Natuurlijk moet er geen betekenisloze precisie worden gerapporteerd, maar een kwart van de standaardmeetfout lijkt te veel. Een betere richtlijn zou zijn om voor de rapportage een precisie te kiezen waarbij de door afronden toegevoegde meetfout verwaarloosbaar is ten opzichte van de meetfout. Men kan natuurlijk een kwart toegevoegde meetfout verwaarloosbaar vinden. Dit is evenwel niet goed te rijmen met de moeite en kosten die gepaard gaan met de constructie van zo nauwkeurig mogelijke meetinstrumenten. Dit betekent ook dat meetnauwkeurigheid minder belangrijk is voor de keuze tussen de zojuist genoemde drie schalen. Hoewel dit niet gebruikelijk is, kan men bijvoorbeeld C-schaalwaarden op een decimaal nauwkeurig rapporteren.

Duidelijker is het om het betrouwbaarheidsinterval van bijvoorbeeld een standaardmeetfout op de schaal zelf af te beelden (zie tabel 13.3). Dit verdient de voorkeur boven het kiezen van de schaaleenheid op basis van de meetnauwkeurigheid. Deze procedure wordt onder andere gevolgd bij de Eindtoets door het betrouwbaarheidsinterval van het cijfer van een leerling met enkele aaneengesloten sterretjes op de cijferschaal weer te geven.

Tabel 13.3

Rapportage van toetsresultaat en de nauwkeurigheid op een reeks van schalen

*	: puntschatting					
+++*+++	: 50% betrouwbaarheidsinterval					
---+++*+++---	: 90% betrouwbaarheidsinterval					
Aantal items goed:	10	12	14	16	18	20
Standaardscore	25	29	33	37	41	45
Percentiel	46	51	56	61	65	70
Groeps-equivalent	5:4	5:8	5:12	6:4	6:8	6:12
Beheersing %	50	59	67	77	86	93
Cijfer	5.2	5.5	6.5	7.5	8.5	9.5
Resultaat Kees	---------+++++++*+++++++---------					

Dit kan, met enige voorzichtigheid, in een keer voor meerdere typen schalen tegelijk. Stel dat de toetsresultaten worden gerapporteerd op de ruwe score-schaal *r*, een genormeerde lineaire transformatie-schaal, standaardscore genoemd, $s(r) = a \times r + b$, een centielschaal, een ontwikkelingsschaal waarop de basisschoolgroep en het aantal maanden van het schooljaar wordt weergegeven, een beheersingsschaal, en op een cijferschaal van 1 tot 10 die wordt verkregen met twee lineaire transformaties met 'de knik' bij het cijfer 5.5. Hoe het rapport er dan kan uitzien is in tabel 13.3 weergegeven. Hoe moeten we nu naar een dergelijk uitgebreid rapport kijken? De puntschatting geeft het behaalde resultaat weer, in eerste instantie de ruwe score, want dat is de schaal waarvan de overige schalen zijn afgeleid. Kees had 16 items goed en de puntschatting weergegeven met * moet dus precies onder het getal 16 in de ruwe scoreschaal staan. Neem aan dat de beide betrouwbaarheidsintervallen bepaald zijn met de standaardmeetfout van de ruwe score. Uit de tabel is te lezen dat het 50%-betrouwbaarheidsinterval van de score van Kees loopt van ongeveer 14 tot 18, het 90% betrouwbaarheidsinterval van ongeveer 12 tot 20. In een overzicht als het bovenstaande geldt dat het betrouwbaarheidsinterval voor alle lineaire transformaties eenvoudig kan worden afgelezen. Stel dat de ondergrens van het 50% interval iets boven de 14 ligt, bijvoorbeeld 14.5, dan ligt deze ondergrens voor de standaardscore ook precies op een kwart van het interval [33,37] vanaf 33, dus op 34. In principe moet men voorzichtiger zijn met niet-lineaire transformaties, omdat men eigenlijk volgens de transformatie zelf moet interpoleren. De bovenstaande schalen wijken over het algemeen, tussen de gespecificeerde cijfers in, zo weinig af van lineariteit dat lineaire interpolatie binnen de gespecificeerde intervallen geen foute interpretaties tot gevolg zal hebben. Bijvoorbeeld, bij een ondergrens van de ruwe score op 14.5, ligt de ondergrens op de beheersingsschaal ongeveer op 67+(77-67)/4 = 69.5. Wanneer een intervalgrens zich precies op een gespecificeerd cijfer bevindt maakt men, ook bij niet-lineaire schalen, geen interpretatie-fout. Als bijvoorbeeld de ondergrens van het 50%-betrouwbaarheidsinterval van de ruwe score precies gelijk is aan 14, dan is deze ondergrens voor de schaal met groepsequivalenten precies gelijk aan 5:12. Dit is ook het geval wanneer groeps-equivalenten niet lineair zijn met de ruwe scores.

Op deze plaats is ook een waarschuwing op zijn plaats in verband met de interpretatie van een score op een ontwikkelingsschaal en de nauwkeurigheid van het meetinstrument. Als de normgroep slechts langzaam groeit op het meetinstrument, kan men grote betrouwbaarheids-intervallen verwachten op de ontwikkelingsschaal, ook bij een relatief nauwkeurig meet-instrument. Kijken we in dit verband weer eens naar het rapport van Kees in figuur 13.4. Nemen we weer zijn resultaat op M4. Daarvan werd beschreven dat zijn resultaat impliceerde dat hij ongeveer een achtste leerjaar op zijn

normgroep achterloopt. Nemen we aan, wat niet onwaarschijnlijk is, dat het 50%-betrouwbaarheidsinterval van zijn meting op E4 ongeveer loopt van de helft van het interval B tot de helft van het interval D, dan is het 50%-betrouwbaarheidsinterval op de groepsequivalenten schaal ongeveer gelijk aan [E3,E4], oftewel een heel leerjaar. Erg veel zekerheid over de vermeende achtste jaar achterstand hebben we dus niet.

13.5 De cesuur voldoende/onvoldoende en andere normen voor cijfergeving

Onder cesuur verstaan we hier het laagste voldoende cijfer. Omdat de cesuur de grens markeert tussen voldoende en onvoldoende, is zij daarmee het belangrijkste cijfer van een schooltoets. Geen wonder dat daarover reeds veel is nagedacht en geschreven (Berk, 1986). De methodes voor cesuur bepaling die ons uit de literatuur bekend zijn, stammen grotendeels uit de zeventiger jaren waarin de beschikking van interactieve computerprogrammatuur niet vanzelfsprekend was, noch het beheer van gecalibreerde itembanken. Deze twee nieuwe mogelijkheden mogen bij de zo belangrijke cesuurbepaling niet worden genegeerd. Hetzelfde geldt evenwel voor de traditie. Daarom is het van belang een vruchtbare synthese tot stand te brengen tussen de concepten die ten grondslag liggen aan de traditionele methoden en de nieuwe mogelijkheden.

We behandelen om te beginnen de methoden die bekend zijn uit de literatuur. Ook de werkwijze bij de centrale eindexamens van het voortgezet onderwijs krijgt enige aandacht omdat die afwijkt van de bekende methoden en, wegens het belang van de examens, hier niet gemist mag worden. Daarna wordt onderzocht hoe de nieuwere mogelijkheden ons in staat stellen deze methoden verder te ontwikkelen. In het laatste deel van de paragraaf besteden we tevens aandacht aan andere onderscheidingen die in een cijferschaal kunnen worden aangebracht, zoals het onderscheid tussen (ruim) voldoende en goed.

13.5.1 Traditionele methoden van cesuurbepaling

Alle methoden voor cesuurbepaling steunen op het gecombineerde oordeel van een groep van 'deskundigen'. Deze deskundigen kunnen uit meerdere groepen afkomstig zijn. Natuurlijk uit het betreffende onderwijs zelf, maar ook de groepen die belang hebben bij het niveau en het aantal geslaagde kandidaten, zoals werkgevers, de overheid, de beroepsgroep, of het vervolg-onderwijs. De methoden voor cesuurbepaling

leveren de deskundigen een methode voor het systematisch specificeren van hun oordelen en het combineren daarvan voor het verkrijgen van een cesuur. Berk (1986) beschrijft 38 methoden voor cesuurbepaling. Hier bespreken we de meest bekende methoden. Al deze methoden hebben betrekking op een toets, dus niet op een itembank of itemdomein.

De methoden voor cesuurbepaling kan men indelen in een groep die alleen gebruik maakt van de 'grenspersoon' en de rest die de hele verdeling van cijfers in de populatie in het proces betrekt. Met een grenspersoon wordt een kandidaat bedoeld die zich precies op de grens tussen zakken en slagen bevindt. De methoden van Angoff, Ebel, Nedelsky en de 'borderline group' methode van Livingston en Zieky behoren tot de eerste groep die zich alleen op de grenspersoon richt. De methoden van Beuk, Hofstee en de 'contrasting groups' methode van Livingston en Zieky maken gebruik van de verdeling van de cijfers in de populatie.

Besliskunde

Omdat de cesuur het criterium is op grond waarvan men beslist of iemand slaagt of zakt, is het zinvol de vaststelling van een cesuur ook vanuit besliskundig oogpunt te bekijken (Hambleton & Novick; 1973, Van der Linden, 1982). De besliskundige benadering van de cesuurbepaling houdt expliciet rekening met het toevallige karakter van het toetscijfer, dat slechts een onnauwkeurig beeld van de ware vaardigheid van een persoon kan geven. Daarom moet er in de eerste plaats een conceptueel onderscheid worden gemaakt tussen de cesuur of het grenscijfer en de grensvaardigheid. Met het grenscijfer of de cesuur x_g bedoelen we de grens op de cijferschaal bijvoorbeeld de ruwe sore of $\hat{\theta}$. Een cijfer lager dan het grenscijfer betekent dat de kandidaat is 'gezakt'. Het onderliggende ware cijfer van een persoon v noemen we zijn vaardigheid en noteren we met ξ_v. De ware score τ is een voorbeeld van een vaardigheid, evenals de persoonsparameter θ op een Raschschaal. De grensvaardigheid wordt genoteerd als ξ_g. Een persoon v met vaardigheid $\xi_v < \xi_g$ verdient te zakken. Heeft persoon v een hogere vaardigheid dan verdient hij te slagen. Het is de bedoeling een cesuur zo te kiezen dat zo goed mogelijk onderscheid wordt gemaakt tussen degenen die verdienen te slagen en degenen die verdienen te zakken. Maar, omdat het (geobserveerde) cijfer niet alleen van de vaardigheid afhangt, maar behept is met een meetfout, lukt het niet altijd om een juiste beslissing te nemen. Zelfs met een optimaal gekozen cesuur kan het voorkomen dat iemand ondanks een vaardigheid $\xi < \xi_g$ toch een voldoende cijfer $x \geq x_g$ behaalt. Zo iemand slaagt onterecht. Als het omgekeerde

het geval is, zakt men onterecht. Beide foute beslissingen kan men in verschillende mate schadelijk vinden. Zo kan men het erger vinden om een ongeschikte kandidaatpiloot te laten slagen dan een geschikte te laten zakken. Ook kan men het erger vinden om een kandidaat met een vaardigheid ruim boven de grensvaardigheid te laten zakken, dan een kandidaat wiens vaardigheid vlak boven de grensvaardigheid ligt. De besliskunde levert een raamwerk om, gegeven een grensvaardigheid ξ_g, een grenscijfer x_g te vinden met een zodanige verhouding tussen de twee soorten verkeerde beslissingen, dat de beslissingen in een bepaalde zin optimaal zijn.

Een eerste stap naar de bepaling van een cesuur is derhalve het vaststellen van de grensvaardigheid ξ_g, de vaardigheid op de grens tussen geslaagd en gezakt. Daarna kan dan het optimale grenscijfer x_g worden bepaald. Helaas zijn veel methoden voor cesuurbepaling tot stand gekomen zonder besliskundige overwegingen. Dit ziet men alleen al daaraan dat het onderscheid tussen cesuur en grensvaardigheid niet wordt gemaakt. Die twee worden min of meer als identiek beschouwd. Toch is meestal duidelijk welke van de twee een bepaalde methode oplevert, een grenscijfer of een grensvaardigheid. We zullen daar steeds op wijzen.

Grensgroepmethoden

De grensgroepmethoden van Angoff, Ebel en Nedelsky, verlangen van deskundigen zich een idee te vormen over een grenspersoon. Vervolgens moeten zij voor ieder item in de toets een oordeel geven over de kans op een correct antwoord voor een grenspersoon. In de methode van Angoff (1971) wordt dit precies zo gevraagd, terwijl Ebel (1972) dit oordeel over items opbouwt in twee stappen. Eerst moet de deskundige de items groeperen volgens een tweeweg-classificatie naar moeilijkheid (makkelijk, gemiddeld, moeilijk) en relevantie voor de te meten vaardigheid (essentieel, belangrijk, acceptabel, twijfelachtig). Daarna wordt voor ieder van de twaalf categorieën items bepaald welk percentage een grenspersoon hiervan goed moet beantwoorden. Nedelsky's (1954) methode is alleen toepasbaar op meerkeuzevragen. De deskundigen moeten voor ieder item aangeven welke afleiders een grenspersoon als fout moet kunnen aanwijzen. Door de aanname dat het antwoord volgens toeval uit de overblijvende alternatieven wordt gekozen, volgt dan de kans op een goed antwoord voor een grenspersoon. Over het algemeen wordt aanbevolen om de deskundigen met elkaars oordelen te confronteren en erover te discussiëren. Daarna krijgen zij de gelegenheid eventueel hun oordelen te herzien.

Ieder van deze drie methoden leidt zo voor iedere deskundige, tot een score op de toets die zij verwachten van een grenspersoon. Deze scores kunnen worden gecombineerd tot de uiteindelijke cesuur door te middelen, eventueel na uitsluiting van extremen, of, door de mediaan te nemen.

Uit de beschrijving blijkt dat deze drie methoden de verwachte ruwe score en daarmee de ware score van een grenspersoon opleveren. Dit is derhalve een grensvaardigheid. Een kandidaat met een vaardigheid beneden de vaardigheid van een grenspersoon, de grens-vaardigheid, hoort te zakken. Deze oorspronkelijke drie methoden nemen echter zonder verdere besliskundige overwegingen de laagste score die niet kleiner is dan de grensvaardigheid als de cesuur. Deze cesuur is over het algemeen in besliskundige zin niet optimaal.

De borderline group methode vereist alleen dat een deskundige de grenspersonen aanwijst, zonder hun toetsresultaat te kennen. De mediaan van de toetsscores van deze groep is de cesuur voor deze deskundige. Noch Zieky (1987), noch Livingston en Zieky (1982) vermelden hoe de cesuren van de deskundigen worden samengevoegd. Men zou ook de mediaan kunnen nemen van de cijfers van alle grenspersonen, waarbij het cijfer van een persoon die door k deskundigen als grenspersoon is aangewezen, k keer meetelt. Een nadeel van deze methode is dat de groep grenspersonen meestal klein is.

Dit nadeel heeft de contrasting group methode niet. Een deskundige geeft voor iedere kandidaat aan of hij moet slagen of zakken, eventueel zonder zijn cijfer te kennen. Men mag echter hopen dat de kans om als voldoende te worden geclassificeerd sterk positief samenhangt met het cijfer. Voor ieder cijfer c telt men het aantal foute beslissingen: het aantal voldoende personen met een cijfer lager dan c en het aantal onvoldoende personen met een cijfer hoger dan c. De cesuur voor deze deskundige is het cijfer met het kleinste aantal foute beslissingen. Deze methode heeft als bijkomend voordeel dat kan worden meegewogen hoeveel erger men het vindt om iemand onterecht te laten slagen dan iemand onterecht te laten zakken. Stel dat men onterecht zakken (een voldoende persoon scoort lager dan c) tweemaal zo erg vindt als onterecht slagen. Men geeft dan de personen die de deskundige als voldoende beoordeelde het gewicht 2, de andere personen het gewicht 1, en summeert de gewichten van de personen, die bij een bepaalde cesuur onterecht als voldoende of onvoldoende worden geklassificeerd.

Uit deze laatste eigenschap blijkt een bepaalde besliskundige benadering. Zoals Van der Linden (1984) opmerkt, wordt hier dan ook een echte cesuur gekozen. Men kan dat als volgt zien. Het oordeel van de deskundige over een kandidaat identificeren we met het gegeven dat de (ware) vaardigheid van de beoordeelde persoon groter of kleiner is dan ξ_g, evenwel zonder dat er expliciet een ξ_g is gekozen. Bij de hier impliciet

gevolgde besliskundige procedure, gebaseerd op drempelutiliteit, is dat echter niet meer relevant zodra bekend is of de vaardigheid onder of boven ξ_g ligt. Drempelutiliteit wordt gebruikt wanneer men vindt dat de afstand van de vaardigheid van een persoon tot de grensvaardigheid voor het nemen van een beslissing van geen belang is. Het wordt bijvoorbeeld even erg geacht iemand onterecht te laten zakken ongeacht of deze nu een vaardigheid heeft net boven de grens-vaardigheid, of ver daarboven. Dit klinkt misschien vreemd, maar men dient hierbij wel te bedenken dat iemand met een vaardigheid ver boven de grensvaardigheid maar zeer zelden zal zakken.

De borderline group methode levert echter, in tegenstelling tot wat Van der Linden (1984) beweert, en in overeenstemming met wat hij 'common belief' noemt, wel degelijk een grens-vaardigheid ξ_g op. De verkregen grensscore is de mediaan van de geobserveerde scores van een groep van min of meer identieke (exchangeable) personen die de deskundige een vaardigheid gelijk aan ξ_g toedicht. Onder een model met normaal verdeelde fouten gegeven de ware score is deze mediaan gelijk aan de verwachte score gegeven ξ_g en derhalve gelijk aan ξ_g.

De laatste twee methoden hebben als nadeel dat de deskundigen de personen moeten beoordelen (natuurlijk) zonder kennis van hun toetsresultaat. Dit impliceert dat de deskundigen de personen op het gebied van de te meten vaardigheid op een andere manier goed moeten kennen. In de praktijk zal het erop neerkomen dat de 'groep' deskundigen beperkt zal zijn tot de eigen (vak)docent. Geen breed samengestelde groep van deskundigen dus.

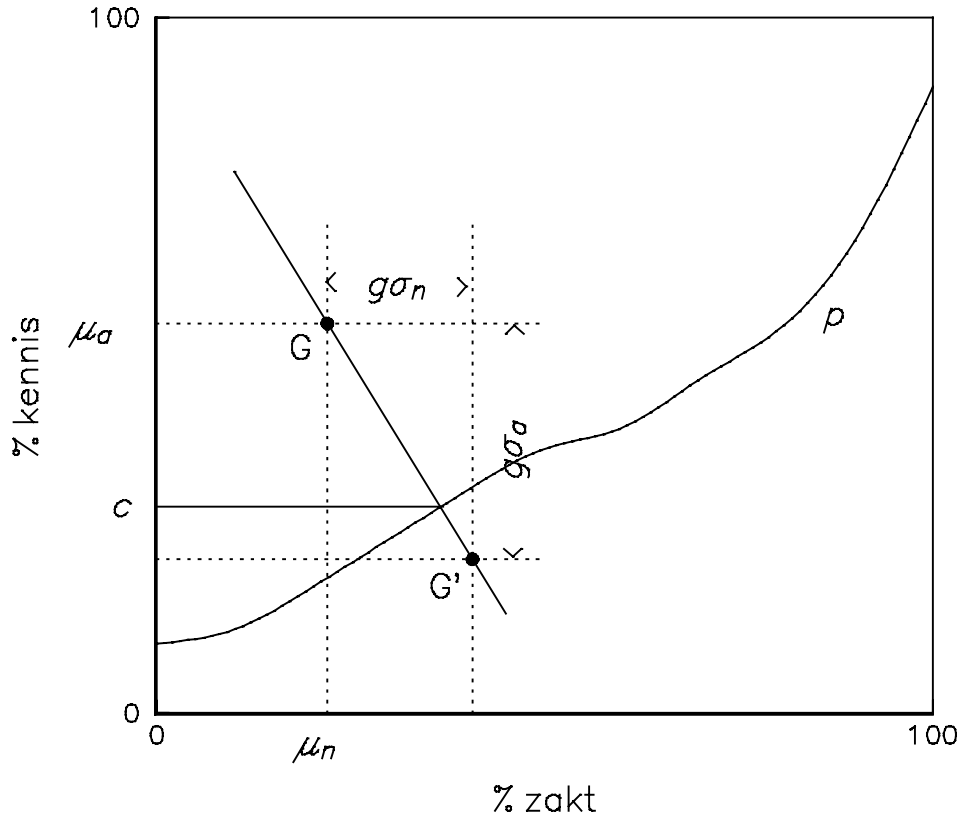
Compromismethoden

De zogenaamde compromismethoden kennen het zojuist genoemde nadeel niet. Iedereen die op de hoogte is met de betreffende vaardigheid en met de populatie van kandidaten kan hier als deskundige zijn oordeel geven. Maar het belangrijkste kenmerk van de compromis-methoden ten opzichte van al de voorgaande is dat er niet alleen naar een acceptabel prestatieniveau wordt gekeken, maar ook naar een acceptabel percentage kandidaten dat zakt. Men zoekt een compromis tussen een absolute cesuur en een normatieve cesuur. Bij een volledig normatieve cesuur telt alleen de verdeling van de cijfers. De cesuur wordt gelegd bij een vooraf bepaald percentage geslaagden, bijvoorbeeld 75%. In dat geval slagen de 75% hoogste cijfers, de overige 25% zakt. Overigens moet men zich niet voorstellen dat dit onderscheid erg hard is te maken. Bij de voorgaande methoden moesten de deskundigen zich immers een grenspersoon voorstellen. Het is haast niet te vermijden dat deze voorstelling mede wordt ingegeven

door een idee over de prestaties in de populatie. Zo spelen normatieve elementen daar ook mee. Vandaar dat we hier niet de strakke indeling volgen die wel eens wordt gemaakt tussen absoluut en normatief normeren bij het behandelen van methoden voor cesuurbepaling.

Bij de compromismethoden van Beuk en die van Hofstee worden de absolute cesuren eerst op een schaal gebracht die het percentage kennis in het getoetste domein weergeeft. Voor toetsen met open vragen is het percentage kennis bij cesuur *c* gelijk aan 100 x c/c_{max}%. Bij meerkeuzevragen wordt gecorrigeerd voor gokken. Als bijvoorbeeld het verwachte cijfer bij puur gokken gelijk is aan c_g, dan is het percentage kennis bij cesuur *c* gelijk aan $100(c\text{-}c_g)/(c_{max}\text{-}c_g)$. Op deze manier worden open vragen en meerkeuzevragen gelijk behandeld. De normatieve cesuur is het percentage van de kandidaten dat zakt.

Volgens de methode van Beuk (1984) wordt van iedere deskundige een absolute cesuur en een normatieve cesuur gevraagd. De deskundige moet de vraag beantwoorden welk percentage kennis hij precies voldoende vindt. Dit is zijn absolute cesuur. Vervolgens moet hij aangeven welk percentage hij vindt dat er moet zakken. Dit is zijn normatieve cesuur.

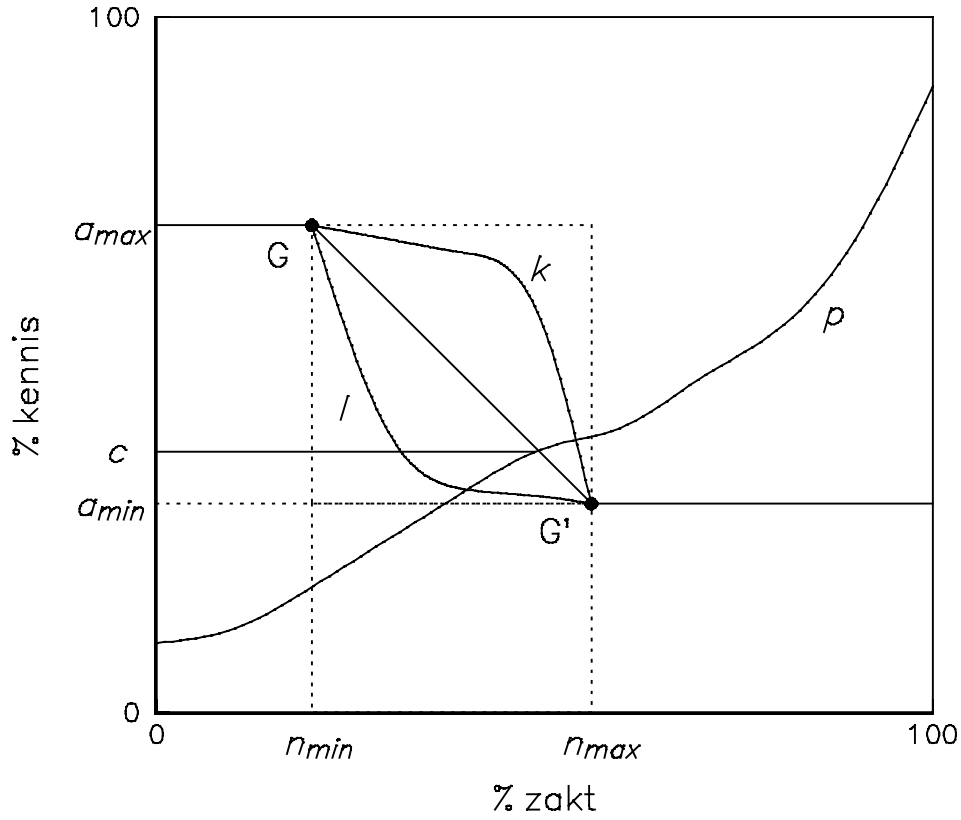


Figuur 13.5

Tabel 13.3
De gewenste absolute en normatieve cesuren van vijf fictieve deskundigen

	1	2	3	4	5	μ	σ	5σ
n% zakt	10	15	15	20	20	16	3.74	18.7
a% kennis	50	60	65	65	70	62	6.78	33.9

Daarna wordt het gemiddelde μ_a bepaald van de absolute cesuren van de deskundigen, en het gemiddelde μ_n van hun normatieve cesuren. In figuur 13.5 is op de horizontale as het percentage gezakten uitgezet en op de verticale as het percentage kennis. In de figuur is het punt (μ_n, μ_a) aangeven met de letter G. Het voorbeeld in figuur 13.5 is gebaseerd op vijf fictieve deskundigen waarvan de gegevens in tabel 13.3 zijn opgenomen. Deskundige 1 vindt bijvoorbeeld dat er 10% moet zakken en dat er minimaal 50% kennis moet worden gevraagd.
Nadat de toets is afgenomen bij de kandidatenpopulatie kent men de verdeling van de percentages kennis, zoals gemeten door de toets. Deze verdeling is in figuur 13.5 aangegeven met de lijn *p*. Een willekeurig punt (*n*, *a*) op lijn *p* betekent dat *n*% van de populatie *a*% kennis of minder heeft, en dus zou zakken als de cesuur bij *a*% zou liggen. Nu zal punt G over het algemeen niet op de lijn *p* liggen. Was dat wel het geval dan waren we klaar. Voor het verkrijgen van de cesuur moeten we vanaf G naar *p* toe schuiven in een richting waarbij de absolute en de normatieve cesuur in een bepaalde zin gelijkwaardig veranderen. Om het begrip 'gelijkwaardig' een precieze inhoud te geven, kiest Beuk voor de mate waarin de deskundigen het onderling eens zijn over beide cesuurtypen. Daartoe berekenen we de standaarddeviaties σ_n van de normatieve cesuren en σ_a van de absolute cesuren. In het voorbeeld in tabel 13.3 is σ_n = 3.74 en σ_a = 6.78. Het punt G' is nu gedefinieerd als $(\mu_n + g\sigma_n, \mu_a - g\sigma_a)$ voor een willekeurige *g* (in figuur 13.5 is *g* = 5). We bepalen vervolgens het snijpunt van GG' en *p*. Dit snijpunt bepaalt het compromis tussen absolute en normatieve cesuurwensen van de deskundigen: het minimaal geëiste kennispercentage *c* om te slagen. Het laagste cijfer op de toets dat hoort bij een kennispercentage groter of gelijk aan *c* is de laagste voldoende.



Figuur 13.6 De cesuur bepaling volgens Hofstee

De methode Hofstee (1977, 1983; De Gruijter, 1985), weergegeven in figuur 13.6, vraagt van elke deskundige twee absolute cesuren en twee normatieve cesuren. Ten eerste de minimum absolute cesuur $a_{\min}$, het percentage kennis dat minimaal wordt geëist ook al zou iedereen zakken en de maximum absolute cesuur $a_{\max}$, het percentage dat men maximaal eist ook al zou iedereen slagen. Vervolgens moet de deskundige het percentage $n_{\max}$ gezakten aangeven dat hij binnen de absolute kennisgrenzen maximaal accepteert. Als $n_{\max}$ of minder procent van de populatie $a_{\min}$ of minder kennis zou hebben dan zou hij zijn eisen tot $a_{\min}$ laten zakken. Tenslotte moet hij het percentage $n_{\min}$ opgeven dat hij minimaal accepteert binnen $a_{\min}$ en $a_{\max}$. Als het percentage gezakten bij $a_{\max}$ als cesuur lager uitvalt dan $n_{\min}$ dan wordt $a_{\max}$ als cesuur genomen. Zij nu G het punt $(n_{\min}, a_{\max})$ en G' het punt $(n_{\max}, a_{\min})$ dan noemt Hofstee het lijnstuk GG' de verzameling acceptabele compromissen. Het snijpunt van *p* en GG' levert dan het feitelijk compromis met cesuur *c*.

Drie opmerkingen over de methode Hofstee. Ten eerste zegt geen enkele van de geraad-pleegde publikaties iets over de manier waarop de oordelen van meer dan een deskundige worden gecombineerd. Men kan op beide assen het minimum van de minima en het maximum van de maxima nemen, maar ook hun gemiddelde of mediaan,

en daarmee de lijn GG' bepalen. Mocht het maximum van de minima kleiner zijn dan het minimum van de maxima, dan zou men ook daarmee de cesuur kunnen bepalen. In dat geval zijn alle deskundigen tevreden met de cesuur als *p* het lijnstuk GG' snijdt. Men zou ook voor iedere deskundige een cesuur kunnen bepalen en daarvan het gemiddelde of de mediaan kiezen. De tweede opmerking betreft de situatie die zich voordoet wanneer *p* het lijnstuk GG' niet snijdt. Mills en Melican (1987) vinden dat er dan opnieuw een cesuur moet worden vastgesteld. Echter, uit de definities van $a_{\min}$ en $a_{\max}$ blijkt dat dan, afhankelijk van heel slechte of juist heel goede prestaties, respectievelijk $a_{\min}$ of $a_{\max}$ de cesuur zal moeten zijn. De derde opmerking betreft de tamelijk willekeurige keuze van de rechte lijn GG' als verzameling acceptabele compromissen. GG' is de lijn waarin normatieve en absolute overwegingen precies gelijk worden gewogen. In principe is echter ieder punt acceptabel dat ligt in de rechthoek waarvan GG' de diagonaal is. In figuur 13.6 representeert de lijn *k* een situatie waarin men aan de absolute cesuur hogere prioriteit geeft dan aan de normatieve, terwijl dit voor de lijn *l* andersom is.

Van deze twee compromismethoden lijkt, ondanks de gesignaleerde onduidelijkheden, die van Hofstee het meest rationeel. In de methode van Hofstee geeft iedere deskundige zijn onderhandelingsruimte duidelijk aan. In de methode van Beuk, daarentegen, worden twee zaken vermengd die niet vermengd lijken te mogen worden. De 'gelijkwaardige' verandering van normatieve en absolute wensen van de deskundigen en de mate waarin zij het onderling eens zijn worden als hetzelfde beschouwd. Hoe meer zij het eens zijn over een van de twee cesuren des te kleiner de relatieve verschuiving. Over het algemeen zal een gelijkwaardige bijstelling echter door andere factoren zijn bepaald. Een klein voorbeeld kan dit verduidelijken. Stel er zijn twee deskundigen die beiden een normatieve cesuur van 25% kiezen, maar ieder een verschillende absolute cesuur, respectievelijk 60% en 70%. Volgens de methode Beuk zakt in dit geval altijd 25% van de kandidaten, ook als de absolute cesuur daarmee bijvoorbeeld op 40% of nog lager zou komen te liggen. Waarschijnlijk vinden de deskundigen 40% kennis als minimale eis niet acceptabel. Zij zouden beiden liever een groter percentage kandidaten laten zakken om zo dichter in de buurt van hun gewenste absolute cesuren te komen.

Het zou beter zijn wanneer iedere deskundige, naast zijn voorkeurspunt, ook twee richtingen van gelijkwaardige verandering zou preciseren, een richting voor een verhoging en een voor een verlaging van de absolute cesuur. Men zou dan het gemiddelde voorkeurs-punt van de deskundigen kunnen bepalen, en ook de twee gemiddelde richtingen. Vervolgens kan men de twee lijnen met deze richtingen vanuit het ideaalpunt trekken en het snijpunt met *p* bepalen voor de cesuur. Een voorbeeld

kan dit verduidelijken. De deskundige ziet het bepalen van de cesuur als een onderhandeling tussen hemzelf en een vertegenwoordiger van de kandidaten. De deskundige bepaalt zijn positie voor de onderhandelingen als volgt. Hij vindt dat 50% kennis is vereist en accepteert daarbij dat 10% van de kandidaten zakt. Mochten er evenwel bij 50% kennis meer dan 10% van de kandidaten zakken dan is hij bereid de absolute cesuur te laten zakken, maar de kandidatenpopulatie moet voor iedere 1% verlaging van de kenniseis genoegen nemen met 9% meer gezakten dan de voorgestelde 10%. Een verlaging van de kenniseis weegt dus negen keer zo zwaar als een verhoging van de normatieve eis. Mochten er bij 50% kennis minder dan 10% van de kandidaten zakken dan is er ruimte voor een kwaliteitsverhoging van het diploma. De deskundige is bereid om in ruil voor iedere 1% verhoging van de absolute cesuur 1% minder kandidaten te laten zakken dan de voorgestelde 10%.

De Gruijter (1985) doet een voorstel waar dit voorstel op het eerste gezicht enigszins op lijkt. Hij hanteert evenwel geen richtingen van verandering maar een Euclidische metriek. Deze metriek is gebaseerd op de relatieve onzekerheid die een deskundige heeft ten aanzien van beide cesuren, niet aan het relatieve belang dat wordt gehecht aan een verhoging of verlaging. In die zin lijdt zijn voorstel aan dezelfde conceptuele verwarring als de methode van Beuk. Er wordt eveneens geen onderscheid gemaakt tussen onzekerheid en bereidheid tot verandering. De Gruijter substitueert alleen een individuele onzekerheid voor de collectieve onzekerheid van Beuk. Bovendien is 'onzekerheid' symmetrisch, zodat geen onderscheid wordt gemaakt tussen verhoging en verlaging van de absolute cesuur. Doordat deze methode geen richting van verandering gebruikt maar een afstandsmaat, heeft zij de vreemde eigenschap dat het kan voorkomen dat de absolute cesuur flink wordt verlaagd, zonder dat daar een noemenswaardige verhoging van het percentage gezakten tegenover staat. Immers, als p onder het ideaalpunt doorloopt en daar niet of nauwelijks stijgt, dan kan het punt op p met de kleinste afstand tot het ideaalpunt, daar bijna loodrecht onder liggen.

Het aanwijzen van een minimaal vereist percentage kennis, komt in het besliskundig raamwerk uiteraard overeen met het aanwijzen van de grensvaardigheid ξ_g. Echter de invloed van de verdeling van de cijfers op de uiteindelijke cesuur, het normatieve element in deze methoden, is precies omgekeerd aan de invloed van het normatieve element in besliskundige procedures. Van der Linden (1984) wijst erop dat besliskundige procedures er toe leiden dat hoe hoger de prestaties in een groep zijn hoe lager de cesuur zal uitvallen. Dit is een fenomeen dat voortvloeit uit het Bayesiaanse karakter van besliskundige procedures.

De centrale eindexamens

Bij de centrale eindexamens wordt de cesuur niet met een van de eerder genoemde methoden bepaald. Hoewel er bij de examens, afhankelijk van het type vragen, zes verschillende gevallen van cesuurbepaling worden onderscheiden, wordt in essentie een enkele methode gevolgd. Om te beginnen wordt er voor ieder examen, voordat de scoreverdeling bekend is op basis van een inschatting van de moeilijkheid van het examen, de laagste voldoende ruwe score gekozen. Als de scoreverdelingen bekend zijn bespreken deskundigen hoe acceptabel deze voorafgekozen cesuur is gezien het percentage kandidaten dat zou zakken bij deze cesuur. Als het examen onverhoopt moeilijker uitvalt dan gedacht, en dus een hoog percentage gezakten zou opleveren bij de vooraf vastgestelde cesuur, dan kan de cesuur binnen bepaalde restricties worden verlaagd. Wanneer het examen makkelijker blijkt dan verwacht, en er dus veel leerlingen slagen bij de vooraf gekozen cesuur, dan mag men de voorafgekozen cesuur meestal niet verhogen.

De cesuurbepaling bij de examens komt het dichtst in de buurt van de compromismethoden. Zij mist echter een duidelijk omschreven procedure voor het afwegen van absolute en normatieve wensen. De voorafgekozen cesuur lijkt het meest op een minimaal vereist percentage kennis, een grensvaardigheid ξ_g. Ook de richting van de invloed van het niveau van de prestatie van de groep lijkt enigszinsop die van de compromismethoden. Een lage prestatie kan worden beloond met een verlaging van de cesuur. Het bestraffen van een hoge prestatie is daarentegen meestal niet toegestaan.

Naar aanleiding van een advies van het Cito over normhandhaving, is er een onderzoek gedaan (Inspectierapport, 1992) naar de gelijkwaardigheid van de examencijfers over een aantal jaren heen. Hieruit bleek dat de moeilijkheid van de examens van jaar tot jaar sterk uiteen liep. Dit is natuurlijk niet zo erg. Door equivalering kan hiervoor immers worden gecorrigeerd. Er bleek echter ook dat de cesuren van jaar tot jaar met sterk verschillende vaardigheden correspondeerden, ondanks de correcties van de cesuren door de deskundigen. Het rapport besluit dan ook met enkele suggesties voor verbetering. Pretesting en calibratie op een schaal met de eerdere examens van hetzelfde type maken er deel van uit.

Ter afsluiting van deze paragraaf behandelen we nog een aardig technisch probleem dat bijvoorbeeld bij examens ontstaat bij het toekennen van cijfers. Ruwe scores, en dus percentages goed op de toets, worden vaak afgebeeld op de gebruikelijke cijferschaal van 1 tot 10 via een of meer lineaire transformaties. De cijfers 1.0 tot en met 10.0 worden dan op een decimaal nauwkeurig gerapporteerd. Voor het vinden van de gewenste lineaire transformatie(s) gaat men als volgt te werk. Men kiest een score r_1,

die exact op het cijfer 5.5 (de laagste voldoende) moet worden afgebeeld. Verder wordt een score r_0 gekozen die op het laagste cijfer 1.0 wordt afgebeeld, en een score r_2 voor het hoogste cijfer 10.0. Dit levert twee lineaire transformaties van scores naar cijfers op, een naar de cijfers 1.0 t/m 5.5 en een naar de cijfers 5.5 t/m 10.0. Bij examens is het exacte cijfer dat men krijgt (uiteraard) erg belangrijk. Een tiende punt meer of minder kan het verschil tussen zakken of slagen uitmaken voor een bepaald vak. Bovendien is de procedure volgens welke de cijfers uit de scores worden berekend openbaar. Men kan zich dus niet veroorloven dat cijfers een tiende punt hoger of lager uitvallen door toevallige afwijkingen die ontstaan door de binaire floating point (drijvende komma) representatie van reële getallen in de computer. Deze ongewenste toevallige effecten zijn te vermijden door een algoritme voor de transformatie te gebruiken zonder floating point-getallen en -operaties. Het algoritme mag alleen met integer (gehele) getallen en integer operaties werken. Omdat de cijfers op 1 decimaal nauwkeurig worden gerapporteerd, verkrijgen we integer cijfers *f* door de oorspronkelijke cijfers met 10 te vermenigvuldigen waardoor *f* integer waarden aanneemt van 10 t/m 100. Beeld r_0 af op het cijfer f_0 en r_1 op f_1. Zij $a = f_1 - f_0$, $c = r_1 - r_o$ en $ar_1 - cf_1$, dan kan de lineaire transformatie $f = g(r)$ van scores *r* naar cijfers *f* geschreven worden met alleen integer getallen. De integer representatie *G(r)* van *g(r) = f = (ar + b)/c* is dan gegeven door:

$$cf \leq ar + b < c(f + 1). \tag{13.4}$$

Gegeven een score $r = r'$ zoekt men een f' die aan deze ongelijkheden voldoet. Als $ar' + b$ dichter bij cf' ligt dan bij $c(f' + 1)$ dan is $G(r') = f'$ anders is $G(r') = f' + 1$ ('afronden' gebeurt in het voordeel van de student). Cijfers kleiner dan het minimum (10) worden als 1.0 en cijfers groter dan het maximum (100) worden als 10.0 gerapporteerd. Bij alle overige cijfers wordt er een punt ingevoegd. Bijvoorbeeld als $f = 56$ wordt het gerapporteerde cijfer 5.6. Door gebruik te maken van integerdeling (genoteerd met \) is het eenvoudig een algoritme te construeren dat de functie *G(r)* berekent. Immers de f' die voor $r = r'$ voldoet aan de ongelijkheden in formule (13.4) is $f' = (ar' + b) \backslash c$.

13.5.2 Cesuurbepaling en overige cijfers binnen itemresponstheorie

Alle hierboven genoemde methoden voor cesuurbepaling kunnen gemakkelijk worden gegeneraliseerd naar een gecalibreerde itembank. Op het eerste gezicht lijkt deze

opmerking niet ter zake, omdat veel van de bovengenoemde methoden nu juist bedoeld waren voor de situatie dat er nog geen empirische gegevens over de items, of de toets bekend zijn. Laat staan dat men de beschikking heeft over een gecalibreerde itembank. Tegenwoordig zullen er echter bijna altijd empirische gegevens van de doelgroep beschikbaar zijn over items uit een leerstofdomein. Met deze gegevens kan men de items calibreren en de vaardigheids-verdeling van de doelgroep schatten. Op basis van deze gecalibreerde itembank kan men een grensvaardigheid ξ_g bepalen. De vaardigheidsverdeling van de doelgroep en een geschikte besliskundige procedure leveren nu voor iedere toets een optimale cesuur. Wanneer de toets of het examen moet bestaan uit nieuwe, niet eerder gebruikte items, dan kunnen die later gecalibreerd aan deze itembank worden toegevoegd.

Voor alle methoden van cesuurbepaling kiest men uit de itembank een reeks items waarvan men verwacht dat die de vaardigheid in de buurt van de nog nader te bepalen grens-vaardigheid ξ_g goed zal meten. Deze verzameling items noemen we de referentietoets. We veronderstellen dat het model voor de referentietoets een strikt monotone regressiefunctie $r(\theta)$ van de latente vaardigheid naar de verwachte ruwe score definieert. Voor het Raschmodel en OPLM is dit altijd het geval. Daarmee bestaat dus ook de inverse functie $r^{-1}(r) = \theta(r)$ van scores naar de latente vaardigheid. De methoden van Angoff, Ebel en Nedelsky leveren een verwachte ruwe score r_g voor de grenspersoon, en daarmee de minimaal voldoende vaardigheid $\theta_g = \theta(r_g)$. De borderline group methode van Livingston en Zieky is gebonden aan een groep personen die bij de deskundigen bekend zijn, echter ook deze methode kan eenmalig worden toegepast voor het vinden van een minimaal vereiste θ_g. De contrasting groups methode resulteert niet in een grensvaardigheid, maar in een echte cesuur op de referentietoets. Willen we bij deze cesuur een grensvaardigheid verkrijgen, dan moet de beslissingsprocedure worden omgekeerd. Normaal zoeken we een optimale cesuur bij een gegeven grensvaardigheid. Nu moeten we een grensvaardigheid vinden waarvoor deze cesuur op de referentietoets optimaal is.

Met een gecalibreerde itembank en een schatting van de verdeling van de vaardigheden kunnen de beide compromismethoden worden vervangen door een veel directer alternatief. Bij iedere θ is niet alleen het kennispercentage op de referentietoets bekend, maar ook het percentage kennis op de hele itembank. Bovendien staat de verdeling van vaardigheden in de doelgroep ter beschikking. Daardoor kent men bij ieder kennispercentage, dus bij iedere mogelijke grensvaardigheid, het percentage in de doelgroep dat verdient te zakken. Men kan er derhalve mee volstaan om iedere deskundige direct op de curve *p* in de figuren 13.5 en 13.6 zijn combinatie van absolute en relatieve cesuur te laten aangeven. Voor het

combineren van verschillende keuzen op de lijn *p* zijn dan meerdere voor de hand liggende oplossingen te bedenken. Een mogelijk probleem bij deze methode is, dat het percentage werkelijk gezakten bij een optimale cesuur over het algemeen niet gelijk zal zijn aan het percentage dat verdient te zakken.

Een gecalibreerde itembank kan ook worden ingezet voor het rapporteren op de schalen die behandeld zijn in paragraaf 13.2. De cumulatieve verdelingen, zoals centielen bij een geschatte vaardigheid zijn eenvoudig te berekenen. De informatiefunctie van de toets en de verdeling van de vaardigheden in de doelgroep bepalen de verdeling van de vaardigheids-schatter. Ook de genormeerde lineaire transformaties zijn daarmee eenvoudig op de latente schaal af te zetten. Alleen met de genormaliseerde schalen moeten we oppassen in verband met de eigenschap 'intervalniveau'. Hierboven werd gesteld dat de T-schaal (en de C-schaal en de Stanines) intervalniveau heeft en per definitie normaal is verdeeld in de referentie-populatie. Als de latente vaardigheidsschatter ook normaal is verdeeld, dan is de T-schaal een lineaire transformatie van de latente vaardigheidsschatter. Is deze laatste duidelijk niet normaal verdeeld, dan hebben we twee schalen van verondersteld intervalniveau, die geen lineaire transformatie van elkaar zijn. De conclusie moet zijn dat minstens een van de twee schalen er geen aanspraak op kan maken van intervalniveau te zijn.

Vele schoolgeneraties lang is het al gebruikelijk om de prestaties in ieder geval (ook) te rapporteren op een zogenaamde cijferschaal. In Nederland is dat de bekende schaal van 1 tot en met 10. Naast het rapporteren van een percentiel of T-schaalwaarde moet er dan ook een transformatie worden geconstrueerd van vaardigheidsschattingen naar de cijferschaal. We kunnen hier kort over zijn. In principe is iedere cijferovergang, bijvoorbeeld die van 7.9 naar 8.0, op een analoge manier te behandelen als de grensvaardigheid voor de cesuur. Alle methoden die men gebruikt voor het vaststellen van een grensvaardigheid, zijn ook toepasbaar voor de bepaling van een andere vaardigheidsgrens. Gelukkig hoeft niet voor alle 90 cijferovergangen op de schaal van 1.0 tot 10.0 afzonderlijk een grensvaardigheid te worden vastgesteld. Enkele belangrijke overgangen, zoals die tussen 7.9 en 8.0, of tussen 4.4 en 4.5, kan men zorgvuldig behandelen. De overige overgangen kan men vervolgens vastleggen door (lineaire) interpolatie. Is de cijferschaal eenmaal vastgelegd, dan kan vervolgens voor vele toekomstige examens die uit deze itembank worden samengesteld dezelfde automatisch geëquivaleerde cijferschaal worden gehanteerd.

Op basis van deze cijferschaal kunnen vervolgens de minimale psychometrische kwaliteiten worden gespecificeerd waaraan het examen in onze ogen moet voldoen. Uiteraard is de grens tussen voldoende en onvoldoende het punt waarnaar onze

grootste zorg zal uitgaan. Een kandidaat met een vaardigheid groter dan de minimale voldoende vaardigheid moet een zo klein mogelijke kans hebben om onvoldoende te scoren. Het is natuurlijk erger wanneer een kandidaat die een 7.0 verdient beneden de 5.5 scoort, dan wanneer dit een kandidaat overkomt die een 5.6 verdient. Zoeken we eerst het vaardigheidsinterval dat begrensd wordt door de ondergrens voor de 7.0 en de ondergrens voor de 7.1. Het midden, $\theta_{7.0}$, van dit interval representeert de vaardigheid van de kandidaten die een 7.0 verdienen. De kans dat met de vaardigheid $\theta_{7.0}$ beneden de 5.5 wordt gescoord neemt af naarmate het examen meer informatie bevat tussen de ondergrens van het interval 5.5 en $\theta_{7.0}$, terwijl tevens de informatie op $\theta_{7.0}$ zo laag mogelijk moet zijn (Verstralen & Verhelst, 1991). Als we er ook waarde aan hechten dat iemand die een 8.0 verdient een zo klein mogelijke kans heeft een 6.5 of minder te halen, dan kunnen deze twee wensen elkaar een beetje in de weg zitten. Verder kan uiteraard het aantal items niet onbeperkt groot gekozen worden. Er is programmatuur (Verschoor, 1990) die kan helpen bij het expliciteren van onze wensen met betrekking tot de lokale meetnauwkeurigheid van het examen en het vaststellen van de minimale informatiefunctie die daarbij hoort. Bij iedere informatiefunctie I kan worden gekeken hoeveel items ongeveer nodig zijn voor een toets met een informatiefunctie die groter is dan I. Bovendien kan worden beoordeeld of de conditionele verdelingsfunctie van een selectie van de cijfers gegeven θ, bijvoorbeeld $\theta = \theta_{7.0}$, acceptabel is. Als de selectie de cijfers 7.0 en 5.4 bevat, kunnen we zien hoe groot de kans is dat iemand die een 7.0 verdient, onvoldoende scoort. Hetzelfde kan ook voor andere vaardigheden worden bekeken. We kunnen bijvoorbeeld nagaan wat de kans is dat iemand die een 6.5 verdient een onvoldoende scoort. Maar ook hoe groot de kans is dat iemand die een 5.0 verdient een 6.0 of hoger haalt. Als we op deze manier onze psychometrische wensen, binnen de randvoorwaarden van het examen hebben vorm gegeven, kunnen we een examen samenstellen dat aan deze psychometrische eisen en de specificaties zoals neergelegd in een toetsmatrijs voldoet.

Gegeven een toets uit een Rasch- of OPLM-gecalibreerde itembank, kan er een functie $\hat{\theta}(s)$ van (gewogen) toetsscores naar vaardigheidsschattingen worden gevonden. We hadden met de cijferintervallen al een functie $c(\theta)$ van θ naar de cijfers van 1.0 tot en met 10.0 die θ afbeeldt op het cijfer van het interval waartoe het behoort. De samenstelling $d(s) = c(\hat{\theta}(s))$ genereert dan een transformatietabel van scores naar cijfers. Voor het bevorderen van een goed begrip van deze cijfers, kan bij ieder cijfer het centiel in een normgroep en het scorepercentage op de itembank en op de toets vermeld worden.

In de bovenbeschreven procedure voor de transformatie van scores naar cijfers is geen rekening gehouden met besliskundige aspecten. Hoewel dit in de praktijk niet

gemakkelijk zal zijn, is het principe niet ingewikkeld. Men bepaalt voor ieder van de 91 classificaties een utiliteitsfunctie $U_f(\theta)$ ($f = 1.0, \ldots, 10.0$). Met $U_f(\theta)$ geeft men aan welke waarde men eraan hecht om iemand met vaardigheid θ te classificeren als *f*. Men doet er uiteraard verstandig aan om in de serie functies U_f enige systematiek aan te brengen zodat er niet voor iedere *f* afzonderlijk nagedacht hoeft te worden. Bij iedere score *r* op de toets wordt de a posteriori verdeling g_r van θ bepaald. Vervolgens zoekt men de classificatie *f* met de grootste verwachte utiliteit over g_r. Eventueel kan men andere criteria hanteren in plaats van de grootste verwachte utiliteit (Berger, 1980).

Uiteraard hoort bij de resultaten van een meetprocedure ook een indicatie van de nauwkeurigheid. Gegeven een OPLM-gecalibreerd examen *b* en een vaardigheid θ_{vb} voor persoon *v* op deze OPLM-schaal, dan is de score op het examen een toevalsvariabele met een conditionele verdeling gegeven θ_{vb}. Omdat $\hat{\theta}_{vb} = \hat{\theta}(s_{vb})$ is ook $\hat{\theta}$ een toevalsvariabele met een conditionele verdeling gegeven θ_{vb}. De standaarddeviatie van deze verdeling is de lokale standaardschattingsfout van $\hat{\theta}_{vb}$. Deze lokale standaardschattingsfout kan ook rechtstreeks uit de informatiefunctie van het examen worden berekend als $I(\theta_{vb})^{-½} \approx I(\hat{\theta}_{vb})^{-½}$, en dus ook een 50% of 95% betrouwbaarheidsinterval. Via de hierboven genoemde transformatie *c(.)* verkrijgen we dan de overeenkomstige betrouwbaarheidsintervallen op de cijferschaal en tevens op de schalen die de interpretatie ondersteunen zoals het centiel in de referentiepopulatie. Tabel 13.5 bevat een voorbeeld van een rapportage voor de vakken Duits, Frans en Engels.

Tabel 13.5

Rapportage van cijfers en hun nauwkeurigheid van alle vakken gezamenlijk

Vak	Punt-schatting	Cijfer →					
	*	5.0	6.0	7.0	8.0	9.0	10.0
Duits	6.6		--++*++--				
Frans	6.3		--++*++--				
Engels	7.0			--++*++--			

De symbolen in tabel 13.5 hebben de volgende betekenis:

* : puntschatting (ook als getal afgedrukt onder *),

++*++ : 50% betrouwbaarheidsinterval,

--++*++-- : 95% betrouwbaarheidsinterval.

Daarna kunnen, zoals in tabel 13.3, voor ieder vak afzonderlijk, bijvoorbeeld voor Duits in tabel 13.6, de waarden van de cijfers op overige schalen, zoals norm- en

beheersingsschalen, worden gegeven waarmee de betekenis van de cijfers wordt verduidelijkt. De interpretatie van een dergelijk rapport is behandeld onder tabel 13.3.

Tabel 13.6

Rapportage per vak over meerdere schalen

Vak	Punt-schatting	Schaalwaarde →					
Duits	*						
score % itembank	72	52	66	78	86	93	99
score % examen	67	54	62	69	79	92	98
% populatie	74	63	69	77	87	98	100
cijfer	6.6	5.0	6.0	7.0	8.0	9.0	10.0
			– – + + * + + – –				

Het combineren van de resultaten op verschillende examens tot een zak/slaag-beslissing

Examens bestaan in het algemeen uit een reeks onderdelen die ieder een bepaald schoolvak als onderwerp hebben. In verband met de traditionele toekenning van diploma's, of meer in het algemeen voor een globale niveau-aanduiding, moeten de resultaten op al deze vakken worden gecombineerd tot een eindbeslissing. Over het algemeen bestaan er voor het combineren van de examenresultaten tot een beslissing over het toekennen van een bepaald diploma, allerlei compensatieregelingen. Al deze regelingen zijn echter vaak ad hoc, zodat meer gefundeerde methoden overwogen kunnen worden. Hieronder wordt een mogelijke aanpak geschetst.

Een Bayesiaanse benadering lijkt het meest aangewezen. Zij $\theta = (\theta_1,...,\theta_I)$ een vector van latente vaardigheden op de verschillende onderdelen $i, (i = 1,...,I)$ van het gehele examen. Zij $f(\theta)$, de a priori verdeling van θ, en $f(\theta \mid \boldsymbol{s})$ de a posteriori verdeling van θ, gegeven de vector $\boldsymbol{s} = (s_1,...,s_I)$ van (gewogen) scores op de I examenonderdelen. Noteer de door het model (OPLM) gegeven conditionele verdeling van de scores gegeven θ met $g(\boldsymbol{s} \mid \theta)$ en de marginale scoreverdeling met $g(\boldsymbol{s})$, dan is volgens de regel van Bayes:

$$f(\theta \mid \boldsymbol{s}) = \frac{g(\boldsymbol{s} \mid \theta) f(\theta)}{g(\boldsymbol{s})}. \tag{13.5}$$

Formule (13.5) kan als volgt uitgangspunt zijn voor het combineren van toetsuitslagen tot een beslissing over het algehele niveau.

Zij $\theta^{(5.5)} = (\theta_1^{(5.5)}, ..., \theta_I^{(5.5)})$ de vector van ondergrenzen van de intervallen voor de cijfers 5.5 op de verschillende examenonderdelen en $\Omega^{(5.5)}$ de deelverzameling van $\mathbb{R}^I$, waarin voor ieder element geldt dat alle componenten groter zijn dan het overeenkomstige element in $\theta^{(5.5)}$: $\theta \in \Omega^{(5.5)}$ als voor alle i $\theta_i > \theta_i^{(5.5)}$, dan is

$$P_{\boldsymbol{s}} = P\left(\theta > \theta^{(5.5)} \mid \boldsymbol{s}\right) = \int_{\Omega^{(5.5)}} f(\theta \mid \boldsymbol{s})\, d\theta,$$

de mate waarin we geloof kunnen hechten aan de bewering dat een persoon met scorevector $\boldsymbol{s}$ op alle onderdelen van het examen minstens een voldoende vaardigheid heeft bereikt, en $1 - P_{\boldsymbol{s}}$ dat dit op minstens een van de onderdelen niet het geval is. De ondergrens voor $P_{\boldsymbol{s}}$ waarboven tot toekenning van het diploma wordt besloten, is een subjectief besluit, waarin niet alleen de ernst van onterecht zakken of slagen moet worden verwerkt. Ook is enige ervaring met deze procedure vereist voor een afgewogen keuze.

Omdat het hier een beslissing over zakken of slagen betreft is er ook veel voor te zeggen om een besliskundige benadering te volgen, bijvoorbeeld op basis van de verwachte à posteriori utiliteit. Men kiest voor beide klassen zakken en slagen respectievelijk de utiliteitsfuncties $U_0(\theta)$ en $U_1(\theta)$ en berekent

$$U_i(\boldsymbol{s}) = \int_{\mathbb{R}^I} U_i(\theta)\, f(\theta \mid \boldsymbol{s})\, d\theta$$

voor $i = 0, 1$. Als $U_0(\boldsymbol{s}) > U_1(\boldsymbol{s})$ dan zakt een kandidaat met scorevector $\boldsymbol{s}$, anders slaagt hij. Het grootste probleem van deze benadering is de keuze van de beide utiliteitsfuncties. Men zou om te beginnen de utiliteitsfuncties kunnen bestuderen die impliciet waren in de beslisregels die bij vroegere examens zijn gehanteerd (Lord, 1983b).

Formule (13.5) kan ook de basis zijn voor nauwkeuriger puntschattingen van θ, dan wanneer de schatting per schaal afzonderlijk gebeurt. De verschillende examenonderdelen zullen immers in de a priori verdeling over het algemeen onderling gecorreleerd zijn. Het is dan evenwel beter en helderder om voor de itemcalibratie en de schattingen van persoons-parameters een multidimensioneel IRT-model te kiezen. Het is te verwachten dat dan met aanzienlijk minder dimensies kan worden volstaan

dan het aantal deelexamens, hetgeen in een overzichtelijker beschrijving van de data resulteert.

13.6 Conclusie

Over het algemeen wordt er bij de rapportage van testresultaten in voldoende mate gebruik gemaakt van de methoden en middelen die in de voorgaande paragrafen zijn besproken. Te vaak echter is het schoolrapport en de rapportage van eindexamenresultaten hierop een uitzondering. Ook de kwaliteiten van deze rapporten kunnen worden beoordeeld volgens de criteria die in het voorafgaande zijn besproken. Gezien de spaarzame informatie die het traditionele school- en eindexamenrapport biedt, valt echter niet te ontkennen dat het meten en rapporteren van het bereikte niveau van leerlingen in onze schoolcultuur geen hoge prioriteit heeft. Voor een deel is dit het gevolg van een aversie tegen het beoordelen en vergelijken van kinderen. Wat zou er echter tegen zijn om bijvoorbeeld normgegevens op te nemen met de klas, de regio, het land als normgroepen. Kinderen vergelijken hun rapportcijfers toch ook onderling. Ook beheersingsschalen zouden het informatiegehalte van schoolrapporten aanzienlijk kunnen verhogen. Met name echter, zou de meetnauwkeurigheid meer aandacht moeten krijgen. Een verandering van ruim voldoende naar zeer onvoldoende in een trimester op verschillende vakken moet bijvoorbeeld geweten worden aan een te lage betrouwbaarheid van de instrumenten, of er moet een andere reden zijn waarom de leerling niet zijn normale niveau heeft kunnen laten zien. Zo'n drastische verandering van resultaten mag echter niet zo maar worden geaccepteerd. Het rapporteren van de meetnauwkeurigheid, heeft niet alleen tot doel om ouders een betere inschatting te laten maken van de nauwkeurigheid van een resultaat. Belangrijker is dat een onderwijsinstelling meer geneigd zal zijn om de meetnauwkeurigheid van de rapportcijfers op een acceptabel niveau te houden of te krijgen.