TOETS- EN ITEMANALYSE MET TIA

Toelichting bij het lezen en interpreteren van toets- en itemanalyses voor gesloten en/of open vragen

P. Goldebeld

Arnhem, maart 1992
eerste druk

© Cito Arnhem 1992
Uit deze uitgave mogen zonder toestemming van de auteur korte aanhalingen met volledige bronvermelding worden overgenomen.

VOORWOORD

Het lag oorspronkelijk in de bedoeling om de verouderde toelichting op TIA (Interne Documentatie 269 en 270) enigszins te bewerken. Gaandeweg is de bewerking echter uitgedijd tot een publikatie die én de ins en outs van TIA bevat (de hoofdstukken 2 en 3) én een handreiking is aan degenen die behoefte hebben aan meer informatie (de overige hoofdstukken). Het betreft informatie op het gebied van onder andere de klassieke testtheorie, de cesuurbepaling, cijfergeving en normhandhaving. Bij het schrijven van de toelichting is gebruik gemaakt van de vakliteratuur en van de op het Cito aanwezige kennis.

INHOUDSOPGAVE

1 Inleiding

In 1987 zijn er voor het eerst gemengde-toetsen afgenomen bij de centrale eindexamens LBO/MAVO. Gemengde-toetsen bevatten zowel gesloten als open vragen en daarom was het noodzakelijk een standaardprogramma voor dit soort toetsen te ontwikkelen. De Dienst Automatisering en de afdeling Onderzoek en Psychometrische Dienstverlening (OPD) zijn in 1986 begonnen met het ontwikkelen van een toets- en itemanalyse-programma voor gemengde toetsen. Dit heeft geresulteerd in een programma (TIA) met vele (nieuwe) mogelijkheden die beschreven worden in hoofdstuk 2 terwijl in hoofdstuk 3 de in- en uitvoer van TIA wordt besproken aan de hand van een examen dat in 1991 is afgenomen. De in- en uitvoer van dit examen is in een aparte bijlage opgenomen.

In hoofdstuk 4 worden betrouwbaarheid en standaardmeetfout aan de orde gesteld terwijl in hoofdstuk 5 onderwerpen worden behandeld die interessant (kunnen) zijn voor personen die meer over de klassieke testtheorie willen weten. In hoofdstuk 6 worden normen voor psychometrische grootheden zoals KR-20 en r_{it}-waarde vermeld en in de hoofdstukken 7 en 8 wordt beschreven hoe de cesuur vastgesteld kan worden en hoe scores omgezet kunnen worden naar cijfers. In hoofdstuk 9 worden normhandhavingsmethoden beschreven en in hoofdstuk 10 wordt uitgebreid ingegaan op de nauwkeurigheid van diverse psychometrische grootheden. Hoofdstuk 11 bevat een overzicht van de op de Dienst Automatisering aanwezige standaard-antwoordbladen. Tenslotte is in hoofdstuk 12 een aantal boeken en tijdschriftartikelen genoemd waarbij 'Tentamineren' van Dousma en Horsten aanbevelenswaardig is. Zeker voor een niet-ingewijde op psychometrisch gebied.

Gezegd moet nog worden dat van TIA zowel een PC- als een HP-versie gemaakt is en dat er naar gestreefd is de aansturing van beide versies gelijk te maken. Hoe TIA precies aangestuurd moet worden is beschreven in een handleiding waarvan W.J. van Daal van de Dienst Automatisering de auteur is. Verder staan de ontwerpers open voor suggesties en bij een eventuele herziening van TIA zullen deze zeer zeker in overweging genomen worden.

2 De mogelijkheden van TIA

Met het toets- en itemanalyse-programma TIA is het onder andere mogelijk toetsen bestaande uit gesloten en/of open vragen te analyseren. Alle mogelijkheden worden hieronder puntsgewijs genoemd.

a. Met TIA kunnen de volgende toetsvormen geanalyseerd worden:
   - toetsen met uitsluitend gesloten vragen (z.g. gesloten-toetsen)
   - toetsen met uitsluitend open vragen (z.g. openvraag-toetsen)
   - toetsen met gesloten en open vragen (z.g. gemengde-toetsen), waarbij de gesloten en open vragen door elkaar mogen staan.

b. Indien de toets uitsluitend gesloten vragen bevat dan kan men kiezen uit een t/i-analyse voor gesloten vragen of een t/i-analyse voor open vragen. Aanbevolen wordt in dat geval de toets te analyseren met een t/i-analyse voor gesloten vragen omdat deze meer informatie oplevert.
   Bij een toets met uitsluitend open vragen en bij een toets met gemengde vragen krijgt men een t/i-analyse voor open vragen. Een gemengde-toets wordt dus behandeld alsof het een toets met allemaal open vragen is. Uiteraard kan men wel, indien er een subtoets voor de gesloten vragen wordt gespecificeerd, deze subtoets analyseren met een t/i-analyse voor gesloten vragen.

c. De toets moet minimaal 3 en mag maximaal ca. 325 vragen bevatten met de kanttekening dat het getal 325 niet te absoluut genomen moet worden. De precieze bovengrens van het aantal vragen hangt samen met de gekozen opties.

d. De behaalde scores (0 of 1 punt bij gesloten vragen en bij open vragen het door de corrector toegekende aantal punten) kunnen gewogen worden. De wegingsfactor kan een positief geheel of een gebroken getal zijn en kan van item tot item verschillen. Het wegen gebeurt direct nadat de scores zijn ingelezen en met de gewogen item-scores worden de berekeningen uitgevoerd.
   De gewogen leerlingscores worden altijd als gehele getallen gepresenteerd, dit geldt zowel voor de scores van subtoetsen als voor de scores van de gehele toets.

e. Bij een *open vraag* mogen (vóór weging) de scoremogelijkheden variëren van 0 punten tot meer dan 20 punten. Echter bij meer dan 20 punten per vraag kan de bij die vraag behorende frequentieverdeling niet afgedrukt worden.
   Verder mag de maximale score geen 0 punten bedragen en kunnen er vooraf geen kwarten en halven gegeven worden, wel kan men per vraag een wegingsfactor opgeven (zie ook punt d).
f. Een gesloten vraag moet minimaal 2 en mag maximaal 6 alternatieven (antwoordmogelijkheden) bevatten.
g. Bij een gesloten vraag moet een kandidaat/leerling op het antwoordblad één alternatief aanstrepen, niet meer en niet minder. In alle andere gevallen wordt het antwoord fout gerekend en krijgt een kandidaat 0 punten.
h. In het geval dat een vraag niet goed heeft gefunctioneerd (een z.g. 'calamiteiten-item', afgekort 'cala-item') is het bij gesloten en open vragen mogelijk om een vervangende (ongewogen) score op te geven welke kan variëren van 0 tot de maximale bij een vraag behorende score. Verder is het mogelijk een vraag te laten vervallen.
   Bij een gesloten vraag is het ook nog mogelijk om meer dan één antwoord goed te rekenen. Als alle antwoorden goed worden gerekend dan impliceert dit dat alle kandidaten 1 scorepunt krijgen, ook degenen die het item niet gemaakt hebben. Bij gesloten vragen is dus alle antwoorden goed rekenen hetzelfde als iedereen de vervangende score 1 geven.
i. Naast de toets- en itemanalyse van de totaaltoets zijn de volgende opties mogelijk:
   - toets- en itemanalyse van subtoetsen. Een subtoets moet, evenals een totaaltoets, uit minimaal 3 vragen bestaan
   - frequentietabel van toetsscores
   - histogram
   - correlatietabel, bevattende correlaties tussen (sub)toetsen
   - empirische item-response-curve
   - extra betrouwbaarheden
   - het geven van bonuspunten (alleen bij de totaaltoets)
   - het maken van een scorebestand; bij dit scorebestand zijn in de totaalscores de eventuele bonuspunten verdisconteerd

- het opvragen van het dichotome bestand behorende bij de gesloten vragen
- het opvragen van een lijst waarop de foutieve invoerrecords (antwoordpatronen van de kandidaten) staan
- het opvragen van de eerste vijf (goede) invoerrecords
- het maken van gegevenstabellen die de toets- en itemgegevens bevatten zoals de betrouwbaarheid en de *p*- en r_{it}-waarden.

Opgemerkt moet worden dat wanneer er sprake is van grenzen deze niet absoluut zijn. Bijvoorbeeld: het maximale aantal items in één toets mag circa 325 zijn en het maximale aantal subtoetsen circa 20. Wanneer in een toets sprake is van beide maxima dan kan er in combinatie met enkele opties wel eens te weinig rekencapaciteit zijn. Wil men echter alleen een analyse van de totale toets zonder subtoetsen en zonder opties, dan mag het maximale aantal items in een toets meer dan 325 zijn.

3 Beschrijving van de in- en uitvoer van TIA

Het nieuwe programma wordt toegelicht middels de uitvoer van een LBO/MAVO-D toets. Deze uitvoer staat in een aparte bijlage. Om bepaalde zaken duidelijk te maken zijn enkele sleutels aangepast: bij item 11 is het oorspronkelijk goede antwoord gewijzigd, bij item 19 is naast C ook B goed gerekend en bij de items 10 en 31 krijgen alle leerlingen de maximale score.
De LBO/MAVO-toets bestond uit 22 gesloten vragen en 9 open vragen. Elke gesloten vraag had een gewicht van 2 zodat men met de gesloten vragen maximaal 44 punten kon verdienen. Met de 9 open vragen waren maximaal 46 punten te verdienen. In totaal konden de leerlingen dus 90 punten behalen. Daarnaast kreeg elke leerling 10 bonuspunten zodat de maximaal te behalen score 100 punten bedroeg en de minimale 10.

Via het voorloopbestand, waarmee TIA wordt aangestuurd, is een toets- en itemanalyse van de totaaltoets gevraagd. Daarnaast is apart een toets- en itemanalyse gevraagd van de gesloten vragen (de items 1 t/m 22) en van de open vragen (de items 23 t/m 31).
Bovendien zijn nog de volgende opties gekozen:
- de lijst met foutieve invoerrecords (kandidaatrecords)
- het overzicht van de eerste 5 (goede) invoerrecords
- de empirische item-response-curve
- extra betrouwbaarheden
- de correlatietabel
- de scores van de leerlingen
- het gedichotomiseerde bestand van de gesloten vragen
- tabellen met toets- en itemgegevens (de zgn. gegevenstabellen)

Om een toets met calamiteiten-items te creëren hebben alle leerlingen bij item 10 een vervangende (ongewogen) score van 1 gekregen en bij item 31 een vervangende score van 7. Bovendien is bij item 19 naast C ook B goed. Let wel dat deze cala-items gefingeerd zijn en in de oorspronkelijke toets- en itemanalyse niet voorkomen.

De volgende uitvoer wordt nu gepresenteerd en besproken:
in 3.1 : Het voorloopbestand (bijlage I)
in 3.2 : Foutenlijst invoerrecords (bijlage II)
in 3.3 : Overzicht invoerrecords (bijlage III)

in 3.4 : Frequentietabel van de totale toets (bijlagen IV en V)
in 3.5 : Histogram van de totale toets (bijlage VI)
in 3.6 : Toets- en itemanalyse van de totale toets (bijlage VII)
in 3.7 : Empirische item-response-curve van de totale toets (bijlage VIII)
in 3.8 : Toets- en itemanalyse van de gesloten vragen (bijlage XI)
in 3.9 : Correlatietabel (bijlage XVII)
in 3.10: Het dichotome bestand en het scorebestand (bijlage XVIII)

Verder is in de bijlagen XIII tot en met XVI terwille van de volledigheid de uitvoer van de 9 open vragen gepresenteerd. Deze uitvoer lijkt op de uitvoer van de totale toets en wordt daarom niet nader toegelicht.

3.1 Het voorloopbestand (bijlage I)

Met het voorloopbestand (ook wel stuurbestand genoemd) wordt het toets- en itemanalyseprogramma aangestuurd. In het voorloopbestand staat exact vermeld hoe de samenstelling van de toets is en wat de gebruiker wel en niet wil.

Als voorbeeld zullen we enkele regelnummers verduidelijken:
regel 014: elke leerling krijgt 10 bonuspunten
regel 016: de items 1 t/m 22 zijn gesloten vragen
regel 018: de gesloten vragen worden met de factor 2 vermenigvuldigd
regel 020: de items 2, 11 en 20 bestaan uit 5 alternatieven
regel 022: de sleutel van de gesloten vragen
regel 023: de items 23 t/m 31 zijn open vragen
regel 025: de maximaal te behalen score op elk van de open vragen
regel 033: elke kandidaat krijgt op vraag 10, een gesloten vraag, een vervangende (ongewogen) score van 1
regel 034: van item 19 is zowel B als C goed
regel 035: elke leerling krijgt op vraag 31, een open vraag, een vervangende score van 7

Voor een volledige toelichting op het aansturen van TIA wordt verwezen naar de gebruikershandleiding van W.J. van Daal (1992) van de Dienst Automatisering.

3.2 Foutenlijst invoerrecords (bijlage II)

Op deze lijst staan de records (leerlingantwoordpatronen) waarvan één of meerdere antwoorden c.q. scores van een leerling niet zijn toegestaan. De eerste 9 cijfers van een record vormen de toets/leerlingidentificatie, de volgende 22 vormen de antwoordcodes van de 22 gesloten vragen en de laatste 9 cijfers zijn de scores behaald op de open vragen 23 t/m 31.
In het bestand van 2596 records zijn er 3 signaleringen. In leerlingrecord 00001 staat bij het laatste item (een openvraag-item) een score van 9, terwijl de maximaal te behalen score 7 is. Dit record wordt uit het bestand verwijderd en doet in de analyse niet mee. De betreffende leerling krijgt geen score.
In het leerlingrecord 00002 staat bij itemnummer 010 code 7 (F is op het antwoordblad aangestreept) terwijl het item een vierkeuze-item is. Dit record wordt echter niet verwijderd; normaliter zou deze leerling 0 punten krijgen bij item 010 maar omdat het item een cala-item is (zie ook bij regelnummer 033 in het voorloopbestand) krijgt deze leerling 1 punt evenals de andere leerlingen. In record 02111 tenslotte is bij twee items iets vergelijkbaars aan de hand als in record 00002.

3.3 Overzicht invoerrecords (bijlage III)

In dit overzicht staan de eerste vijf goede leerlingrecords (record 1 is niet afgedrukt omdat er bij item 31 een score staat die hoger is dan de maximale score). In het eerste (goede) leerlingrecord zien we vóór de eerste verticale streep het toets/leerlingidentificatienummer en tussen de eerste en tweede streep het antwoord dat de leerling op elke gesloten vraag heeft gegeven waarbij de codes het volgend betekenen:
code 1: een leerling heeft A aangestreept op het antwoordblad
code 2: een leerling heeft B aangestreept op het antwoordblad
code 3: een leerling heeft C aangestreept op het antwoordblad
code 4: een leerling heeft D aangestreept op het antwoordblad
code 6: een leerling heeft E aangestreept op het antwoordblad

code 7: een leerling heeft F aangestreept op het antwoordblad
code 0: een leerling heeft of een item overgeslagen of twee of meer antwoorden aangestreept of een niet bestaand antwoord aangestreept.

Achter de tweede verticale streep staat de score die de leerling op elk van de open vragen heeft behaald.

3.4 Frequentietabel van de totale toets (bijlagen IV en V)

In de kop van de tabel staan kenmerkende gegevens als schooltype, tijdvak, jaar en toetsnummer en indien van toepassing: subtoetsnummer, subtoetsnaam en de itemnummers van de subtoets.
Voor een deel kan men zelf bepalen wat er in de kop van de tabel komt te staan (zie regelnummers 002 en 010 van het voorloopbestand). Het is verstandig om zoveel mogelijk informatie in de kop mee te nemen mede omdat deze kop terugkomt bij andere tabellen.

De frequentietabel bestaat uit 4 kolommen die de volgende gegevens bevatten.

SCORE	De kolom met het kopje SCORE bevat alle mogelijke scores die op een toets te behalen zijn. De kolom begint bij deze toets met de score 10 vanwege de 10 bonuspunten en eindigt bij 100. De scores zijn dus van laag naar hoog gerangschikt.
FREQ. (frequentie)	In deze kolom staat de frequentie, d.w.z. het aantal kandidaten dat een bepaalde toetsscore heeft behaald. Bij deze toets hebben bijvoorbeeld 27 kandidaten de score van 41 behaald.
CUM.FREQ. (cumulatieve frequentie)	In deze kolom staat bij elke score de cumulatieve frequentie, d.w.z. het aantal kandidaten dat een bepaalde score of een lagere behaald heeft. Bij deze toets blijken 176 kandidaten een score van 41 of een lagere behaald te

hebben. De cumulatieve frequentie van 176 wordt verkregen door de frequenties behorende bij de scores 10 t/m 41 op te tellen.

CUM.PERC. (cumulatief percentage)

Deze kolom bevat de cumulatieve percentages en deze worden verkregen door de cumulatieve frequenties te delen door het aantal kandidaten dat de toets gemaakt heeft en het resultaat van de deling te vermenigvuldigen met 100.

Dus: $CUM.PERC = \frac{CUM.FREQ}{AANTAL\ KAND} * 100$

Aan de hand van de kolom CUM.PERC. kan bekeken worden hoeveel procent van de kandidaten een onvoldoende krijgt bij een bepaalde cesuur (grens tussen onvoldoende en voldoende). Indien bij deze toets de cesuur op 52/53 was gelegd dan zou 27.7% van de kandidaten een onvoldoende gekregen hebben.

Onderaan de tabel vinden we nog de volgende gegevens:
- AANTAL KANDIDATEN
- GEMIDDELDE SCORE (inclusief bonuspunten)
- STANDAARDDEVIATIE

Bij AANTAL KANDIDATEN staat het aantal kandidaten waarop de toets- en itemanalyse gebaseerd is. Bij GEMIDDELDE SCORE is de gemiddelde toetsscore vermeld (inclusief 10 bonuspunten) en bij STANDAARDDEVIATIE is de standaarddeviatie van de toetsscores vermeld. Bijgaande toets- en itemanalyse heeft betrekking op 2595 kandidaten die een gemiddelde score van 60.51 behaald hebben. De standaarddeviatie bedraagt 12.83.

3.5 Histogram van de totale toets (bijlage VI)

Het histogram is een grafische afbeelding van de behaalde toetsscores van de kandidaten. Langs de horizontale as staat de toetsscore en langs de verticale as staat het percentage kandidaten dat een bepaalde toetsscore heeft behaald, de zogenaamde relatieve frequentie.

3.6 Toets- en itemanalyse van de totale toets (bijlage VII)

De toets- en itemanalyse in bijlage VII is het voorbeeld van een zogenaamde gemengde-toets, de items 1 t/m 22 zijn namelijk meerkeuzevragen en de items 23 t/m 31 open vragen. We zullen de uitvoer weer kolom voor kolom bespreken.

ITEM — Onder ITEM staan 2 kolommen. De ene kolom bevat de nummers van de vragen welke in dit voorbeeld van 1 t/m 31 lopen en de andere kolom bevat de itemcodes. Item nr. 1 heeft als code MK0001. Men kan de code zelf bepalen via het voorloopbestand (zie regelnummers 017 en 024).

MAX.SCORE (maximale score) — Deze kolom bevat de maximale score die per vraag behaald kan worden. In ons voorbeeld kan men bij de meerkeuzevragen vanwege de weging (zie kolom GEW.) maximaal 2 punten per vraag verdienen. Bij de open vragen varieert het aantal te verdienen punten van minimaal 3 tot maximaal 7.

GEM.SCORE (gemiddelde score) — Per vraag is de gemiddelde score van de kandidaten vermeld.

p′ (p'-waarde) — De p'-waarde is een percentage tussen 0 en 100 dat de moeilijkheidsgraad van een vraag en/of de vaardigheid van de kandidaten weergeeft. De p'-waarde wordt als volgt berekend:

$$P' = \frac{GEM.SCORE}{MAX.SCORE} * 100,$$

waarbij GEM.SCORE de gemiddelde score op een vraag is en MAX.SCORE de maximaal op die vraag te behalen score.

SD (standaarddeviatie)

De standaarddeviatie is een maat voor de spreiding van de scores binnen een vraag. Hoe hoger de standaarddeviatie hoe groter de spreiding.

Indien alle kandidaten dezelfde score hebben behaald dan is de standaarddeviatie 0. De standaarddeviatie is maximaal als de ene helft van de kandidaten de vraag fout heeft en de andere helft de vraag goed.

RSK (relatieve standaarddeviatie)

De RSK is een relatieve spreidingsmaat die dient om de standaarddeviaties van de verschillende vragen vergelijkbaar te maken. De RSK wordt berekend door van een vraag de standaarddeviatie (SD) te delen door de maximaal te behalen score op die vraag (MAX.SCORE).

Dus: $RSK = \frac{SD}{MAX.SCORE}$

We zien bijvoorbeeld dat van alle vragen vraag 30 absoluut gezien de grootste spreiding heeft (SD = 2.15), relatief gezien is de spreiding echter niet afwijkend (RSK = .36).
De naam voor de relatieve standaarddeviatie is ontleend aan oud-Cito-medewerker Wiel Knops.

RIT (r_{it}-waarde)

De r_{it} is de correlatie (produkt-momenten-correlatie en bij gesloten vragen ook wel point-biserial genoemd) tussen de score van een vraag en de totaalscore van een toets inclusief de score op de vraag zelf. De r_{it} geeft aan in hoeverre men er met een vraag in geslaagd is te differentiëren

tussen ’goede’ en ’slechte’ kandidaten. Onder goede kandidaten verstaan we kandidaten die een hoge toetsscore hebben behaald en onder slechte kandidaten verstaan we kandidaten die een lage toetsscore hebben behaald. De r_{it} is een discriminatie-index. Een hoge r_{it} betekent dat veel kandidaten met een hoge toetsscore de vraag goed en veel kandidaten met een lage toetsscore de vraag fout hebben beantwoord.
Verder betekent een hoge r_{it} dat de vraag relatief veel bijdraagt aan de betrouwbaarheid van de toets. Nog anders gezegd: een hoge r_{it} betekent dat een vraag een representatief onderdeel is van de totale toets.
N.B.
Een liggend streepje in deze kolom betekent dat de r_{it} rekentechnisch niet bepaald kan worden.

RIR (r_{ir}-waarde)

De r_{ir} is een soortgelijke index als de r_{it}. Gaat het bij de r_{it} het om de samenhang tussen een vraag en de hele toets, inclusief de vraag zelf, bij de r_{ir} gaat het om de samenhang van de vraag en de rest van de toets, dat is de gehele toets minus de betreffende vraag.
Let op dat de r_{it} ’s en r_{ir} ’s in de uitvoer met 100 vermenigvuldigd zijn. Per definitie liggen zij tussen -1.00 (perfect negatief lineair verband) en +1.00 (perfect positief lineair verband).

We willen er nog op wijzen dat zowel aan de r_{it} als aan de r_{ir} bezwaren kleven. De r_{it} geeft namelijk een (enigszins) geflatteerd beeld van de samenhang tussen de score op een vraag en de totaalscore omdat de score op de vraag in de totaalscore is verdisconteerd en we dus de vraag voor een deel met zichzelf

correleren. De r_{ir} ondervangt dit bezwaar maar heeft een ander bezwaar; de resttoets waarmee een vraag gecorreleerd wordt, varieert met de vraag (De Gruijter, 1982). Bovengenoemde bezwaren impliceren dat we voorzichtig moeten zijn met het vergelijken van r_{it}- en r_{ir}-waarden van items indien die afkomstig zijn uit toetsen waarvan de lengtes veel verschillen. Verder is het beter om binnen eenzelfde toets de r_{it}-waarde als discriminatie-index te gebruiken en niet de r_{ir}-waarde.

Tenslotte zij opgemerkt dat genoemde bezwaren geen rol meer spelen bij toetsen met meer dan 40 items (Thorndike, 1982). In Thorndike wordt bovendien een correctieformule voor de r_{it} vermeld die ontwikkeld is door Henrysson (1963). Met deze correctieformule wordt de r_{it} gecorrigeerd voor de invloed van de toetslengte.

EFF
(effectief gewicht)

Onder effectief gewicht verstaan we de bijdrage van een vraag aan de spreiding in de totale toets. Hoe hoger het effectief gewicht van een vraag des te meer spreiding in de totale toets toegeschreven kan worden aan de betreffende vraag. In ons voorbeeld heeft vraag 30 het grootste effectief gewicht.
Het effectief gewicht wordt als volgt bepaald:

$$\text{effectief gewicht} = \frac{RIT * SD}{s_x},$$

waarbij RIT en SD respectievelijk de r_{it}-waarde en standaarddeviatie van een vraag zijn terwijl S_x de standaarddeviatie van de totale toets is.

N.B. De som van de effectieve gewichten is 1.

AR (alpha-rest)

Alpha-rest is de betrouwbaarheid (vermenigvuldigd met 100) van een toets minus het betreffende item. Dus als een item uit een toets wordt verwijderd dan is de AR de betrouwbaarheid van de resterende toets. In ons voorbeeld zien we dat bij verwijdering van één item de betrouwbaarheid licht daalt of gelijk blijft.
De AR is, naast bijvoorbeeld de r_{it}, een maat om de psychometrische kwaliteit van een item te karakteriseren.

D (D-waarde)

D = ALPHA - AR.
In deze kolom is het verschil vermeld tussen de betrouwbaarheid van de totale toets en de AR van het betreffende item. Hoe positiever de D-waarde hoe groter de bijdrage van een item aan de betrouwbaarheid. Een item met een D-waarde van 0 of met een negatieve D-waarde draagt niet of in negatieve zin bij aan de betrouwbaarheid.
Item 11 bijvoorbeeld heeft een negatieve D-waarde. Zouden we dit item uit de toets verwijderen dan stijgt de betrouwbaarheid. (Omdat we met afgeronde getallen werken is deze stijging niet te zien.)

GEW. (gewicht)

Deze kolom bevat per vraag de wegingsfactor waarmee de oorspronkelijke scores vermenigvuldigd zijn. In ons voorbeeld zijn de vragen 1 t/m 22 meerkeuzevragen waarop men niet de traditionele 0 of 1 scoring heeft toegepast. In de kolom GEW zien we dat de scores met een factor 2 vermenigvuldigd zijn. Dus het antwoord fout betekent een score van 0 en het

antwoord goed een score van 2. Zie ook kolom MAX.SCORE.

RELATIEVE FREQUENTIES VAN ITEMSCORES (IN %)

De kolommen met de kop RELATIEVE FREQUENTIES VAN ITEMSCORES (IN %) bevatten per vraag de relatieve frequenties. De scores per vraag kunnen variëren van 0 tot meer dan 20 waarbij zij aangetekend dat het ongewogen scores betreft. Indien bij een vraag meer dan 20 punten te behalen zijn dan wordt er bij die vraag geen frequentieverdeling afgedrukt.

In ons voorbeeld heeft bij vraag 1 18% van de kandidaten 0 punten behaald en 82% 1 punt. Anders gezegd 18% had deze meerkeuzevraag fout en 82% goed. Omdat de wegingsfactor bij deze vraag 2 is betekent dit dat na weging de kandidaten die de vraag fout hadden 0 punten blijven houden en dat de kandidaten die de vraag goed hadden 2 punten krijgen.

Bij vraag 24 (een open vraag) heeft 25% van de kandidaten 0 punten behaald, 19% 1 punt, 19% 2 punten en 37% 3 punten. Bij deze vraag wordt niet gewogen dus blijven de scores zoals ze zijn.

Algemene gegevens

Onderaan de toets- en itemanalyse staan nog de volgende gegevens:

- AANTAL KANDIDATEN
- GEMIDDELDE SCORE
- STANDAARDDEVIATIE
- GEMIDDELDE P′-WAARDE
- BETROUWBAARHEID (ALPHA)
- STANDAARDMEETFOUT

Bij AANTAL KANDIDATEN staat het aantal kandidaten waarop de toets- en itemanalyse gebaseerd is. Bij GEMIDDELDE SCORE is de gemiddelde toetsscore vermeld (zonder bonuspunten) en bij STANDAARDDEVIATIE is de standaarddeviatie van de toetsscores vermeld. De GEMIDDELDE P′-WAARDE is een maat (uitgedrukt in %)

voor de moeilijkheidsgraad van een toets en/of de vaardigheid van de kandidaten en wordt als volgt berekend:

$$\text{GEMIDDELDE } p'\text{-WAARDE} = \frac{\textit{GEMIDDELDE SCORE}}{\textit{MAXIMALE TOETSSCORE}} * 100$$

Vullen we in bovenstaande formule de gemiddelde toetsscore in van 50.51 en de maximale toetsscore van 90 dan is het resultaat een gemiddelde p'-waarde van 56.1. Opgemerkt moet nog worden dat in genoemde formule de bonuspunten niet in de maximale toetsscore zijn verdisconteerd. Er moet verder op gewezen worden dat de gemiddelde p'-waarde berekend moet worden met bovenstaande formule en *niet* via de p'-waarden van de individuele vragen. Met de twee berekeningsmethoden krijgen we immers niet dezelfde uitkomsten indien de maxima van de vragen verschillen. Bovendien wordt er door bovenstaande formule te gebruiken minder vaak afgerond.
Bij BETROUWBAARHEID staat de coëfficiënt alpha(α) vermeld. In hoofdstuk 4 wordt op deze betrouwbaarheidscoëfficiënt ingegaan. De STANDAARDMEETFOUT zal eveneens in hoofdstuk 4 aan de orde komen.

Verder zijn er nog 3 betrouwbaarheidscoëfficiënten vermeld (B1, B2 en BRS) die niet standaard berekend worden maar alleen als dit via een optie is aangevraagd (de optie 'XB' op regelnummer 010 in het voorloopbestand). In paragraaf 5.6 zullen we op deze optionele betrouwbaarheidscoëfficiënten ingaan. We willen echter op deze plaats waarschuwen spaarzaam gebruik te maken van deze optie omdat enerzijds de rekentijd van TIA exponentieel toeneemt en anderzijds deze coëfficiënten (nog) geen gemeengoed zijn in de testtheorie.

3.7 Empirische item-response-curve van de totale toets (bijlage VIII)

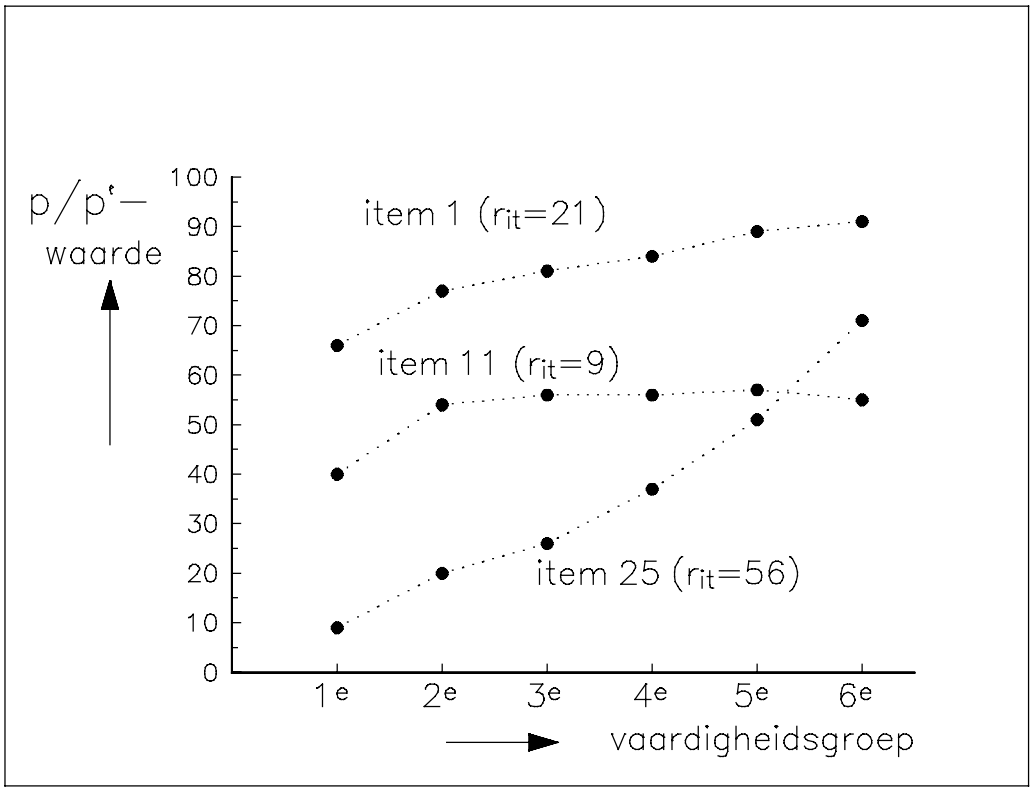
De tabel in bijlage VIII is tot stand gekomen door de 2595 kandidaten op te splitsen in 6 groepen waarin zich ongeveer evenveel kandidaten bevinden (regelnummer 015 en de optie 'EMPIC' op regelnummer 011 in het voorloopbestand). In de 1e groep bevinden zich de zwakste kandidaten, in ons voorbeeld zijn dat 411 kandidaten met een score variërend van 10 t/m 47. In de 2e groep zitten 446 kandidaten met een score lopend van 48 t/m 54. In deze groep zitten dus 'betere' kandidaten dan in de 1e groep. Hoe hoger het groepsnummer des te hoger zijn de

scores van de kandidaten in de groep. In elke groep is nu per item de p'-waarde vermeld van de groep. Als een item in psychometrisch opzicht goed heeft gefunctioneerd dan zal met het oplopen van het groepsnummer de p'-waarde toenemen.

Voorbeelden:

1. Item 1 (code MK001) heeft gemiddeld over alle kandidaten een p'-waarde van 82. Kijken we naar de afzonderlijke groepen dan loopt de p'-waarde op van 66 naar 91. De minst vaardige groep heeft de laagste p'-waarde en de meest vaardige groep de hoogste. Het item heeft in psychometrisch opzicht goed gefunctioneerd.
2. Item 25 is een nog beter voorbeeld van een item dat goed gediscrimineerd heeft; de p'-waarde in de 1e groep is 9 en in de hoogste 71.
3. Item 11, heeft op de 1e groep na, in elke groep een vrijwel gelijke p'-waarde. Met dit item is men er niet in goed geslaagd om de goede en slechte kandidaten te scheiden. Dit item is in psychometrische opzicht niet geslaagd.

Per item kunnen we de p'-waarden grafisch weergeven en het resultaat is een empirische item-response-curve. Van de items 1, 11 en 25 zijn de curves afgebeeld.



Figuur 1. Enkele empirische item-response-curves.

Uit bovenstaande figuur blijkt dat item 25 de steilste curve heeft. Zoals reeds vermeld is men er met dit item heel goed in geslaagd de bokken van de schapen te scheiden. De r_{it} van dit item is dan ook 56 terwijl item 1 een r_{it} van 21 heeft en item 11 een r_{it} van 9.

Via het stuurbestand (zie regelnummer 015) kunnen we aangeven, met als grenzen 2 en 14, hoeveel groepen we willen. Programmatisch is verder ingebouwd dat het gemiddeld aantal kandidaten in een groep nooit minder dan 25 kan zijn. Bij minder dan 25 kandidaten in een groep is de steekproeffout te groot (zie ook hoofdstuk 10).

Rest ons nog te melden dat bij de gesloten vragen voor de p'-waarde de p-waarde gelezen moet worden.

3.8 Toets- en itemanalyse van de gesloten vragen (bijlage XI)

Er zijn in ons voorbeeld 2 subtoetsen aangevraagd, te weten: één subtoets met uitsluitend de gesloten vragen (regelnummer 026) en één

subtoets met uitsluitend de open vragen (regelnummer 029). De uitvoer van de subtoets met open vragen ziet eruit zoals in voorgaande hoofdstukken beschreven is. De beschrijving van de uitvoer van de toets- en itemanalyse van de subtoets met gesloten vragen volgt hieronder. We behandelen de uitvoer weer kolom voor kolom.

ITEM — Onder ITEM staan 2 kolommen. De ene kolom bevat de nummers van de vragen welke in dit voorbeeld van 1 t/m 22 lopen en de andere kolom bevat een itemcode. Item nr. 1 heeft als code MK0001. Men kan de code zelf bepalen via het voorloopbestand (zie regelnummer 017).

SLEUTEL — Deze kolom geeft aan welk alternatief (antwoordmogelijkheid) goed is of welke alternatieven er goed zijn. Het uitroepteken (!) bij item 10 duidt op een calamiteiten-item; elke leerling heeft bij deze vraag 1 punt (ongewogen) gekregen. Ook de leerling die of het item heeft overgeslagen of twee of meer alternatieven heeft aangestreept of een niet bestaand alternatief heeft aangestreept.

GEW. (gewicht) — Deze kolom bevat per item de wegingsfactor waarmee de oorspronkelijke scores vermenigvuldigd zijn. Bij deze gesloten-toets is bij elk item de wegingsfactor 2, dus het antwoord fout betekent een score van 0 en het antwoord goed een score van 2.

P- EN A-WAARDEN — Bij elk alternatief is het percentage leerlingen vermeld dat dit alternatief heeft gekozen. Het percentage waarbij een asterisk (*) staat is het percentage leerlingen dat het antwoord goed heeft. Dit percentage noemen we de *p*-waarde. De bij de afleiders (foute antwoorden) vermelde percentages worden *a*-waarden genoemd. Bij item 2 in ons voorbeeld zien we het volgende patroon:

P- EN A-WAARDEN					
A	B	C	D	E	F
7	7	82*	2	1	

Dit betekent dat C het goede alternatief is met een *p*-waarde van 82%. De *a*-waarden zijn respectievelijk 7% (A), 7% (B), 2% (D) en 1% (E). Bij F staat niets vermeld, er waren 5 antwoordmogelijkheden bij item 2 (regelnummer 020 in het voorloopbestand).
Bij item 19 is zowel alternatief B als alternatief C goed (regelnummer 034). De respectievelijke percentages zijn beide dan ook van een asterisk voorzien. In de kolom met *p*-waarden is de som van beide *p*-waarden vermeld terwijl de r_{it}-waarde in de kolom RIT berekend is door B en C als één goed antwoord te beschouwen.

O/D (overgeslagen/ dubbel)
: In deze kolom staat het percentage kandidaten dat of het item heeft overgeslagen of twee of meer alternatieven heeft aangestreept of een niet bestaand alternatief heeft aangestreept. Bij item 1 staat een percentage van 2 in de kolom O/D. De som der *p*- en *a*-waarden is dan ook 98% in plaats van 100%.

ITEM NR (itemnummer)
: Om de leesbaarheid van de toets- en itemanalyse te verhogen wordt de kolom met itemnummers herhaald.

P (*p*-waarde)
: In deze kolom staan de *p*-waarden een keer apart vermeld.

RIT (r_{it}-waarde)
: De r_{it} is de correlatie (produkt-momenten-correlatie en bij gesloten vragen ook wel point-biserial genoemd) tussen de score van een vraag en de totaalscore van een toets inclusief de score op de vraag zelf. De r_{it} geeft aan in hoeverre men er met een vraag in geslaagd is te differentiëren tussen 'goede' en 'slechte' kandidaten. Onder goede kandidaten verstaan we kandidaten die een hoge toetsscore hebben behaald en onder slechte kandidaten verstaan we kandidaten die een lage toetsscore hebben behaald. De r_{it} is een discriminatie-index. Een hoge r_{it} betekent dat veel kandidaten met een hoge toetsscore de vraag goed en

veel kandidaten met een lage toetsscore de vraag fout hebben beantwoord.
Verder betekent een hoge r_{it} dat de vraag relatief veel bijdraagt aan de betrouwbaarheid van de toets. Nog anders gezegd: een hoge r_{it} betekent dat een vraag een representatief onderdeel is van de totale toets.

N.B. Een liggend streepje in deze kolom betekent dat de r_{it} rekentechnisch niet bepaald kan worden.

RIR (r_{ir}-waarde)

De r_{ir} is een soortgelijke index als de r_{it}. Gaat het bij de r_{it} om de samenhang tussen een vraag en de hele toets, inclusief de vraag zelf, bij de r_{ir} gaat het om de samenhang van de vraag en de rest van de toets, dat is de gehele toets minus de betreffende vraag.
Let op dat de r_{it} 's en r_{ir} 's in de uitvoer met 100 vermenigvuldigd zijn. Per definitie liggen zij tussen -1.00 (perfect negatief lineair verband) en +1.00 (perfect positief lineair verband).

We willen er nog op wijzen dat zowel aan de r_{it} als aan de r_{ir} bezwaren kleven. De r_{it} geeft namelijk een (enigszins) geflatteerd beeld van de samenhang tussen de score op een vraag en de totaalscore omdat de score op de vraag in de totaalscore is verdisconteerd en we dus de vraag voor een deel met zichzelf correleren. De r_{ir} ondervangt dit bezwaar maar heeft een ander bezwaar; de resttoets waarmee een vraag gecorreleerd wordt, varieert met de vraag (De Gruijter, 1982).
Bovengenoemde bezwaren impliceren dat we voorzichtig moeten zijn met het vergelijken van r_{it}- en r_{ir}-waarden van items indien die afkomstig zijn uit toetsen waarvan de lengtes veel verschillen. Verder is het beter om binnen eenzelfde toets de r_{it}-waarde als discriminatie-index te gebruiken en niet de r_{ir}-waarde. (Bij eenzelfde item zijn de r_{ir}- en r_{ar}-waarden onderling vergelijkbaar.)

Tenslotte zij opgemerkt dat genoemde bezwaren geen rol meer spelen bij toetsen met meer dan 40 items

(Thorndike, 1982). In Thorndike wordt bovendien een correctieformule voor de r_{it} vermeld die ontwikkeld is door Henrysson (1963). Met deze correctieformule wordt de r_{it} gecorrigeerd voor de invloed van de toetslengte.

AR (alpha-rest)

Alpha-rest is de betrouwbaarheid (vermenigvuldigd met 100) van een toets minus het betreffende item. Dus als een item uit een toets wordt verwijderd dan is de AR de betrouwbaarheid van de resterende toets. In ons voorbeeld zien we dat bij verwijdering van één item de betrouwbaarheid licht daalt of gelijk blijft. Alleen als item 11 verwijderd zou worden dan zou de betrouwbaarheid licht stijgen.
De AR is, naast bijvoorbeeld de r_{it}, een maat om de psychometrische kwaliteit van een item te karakteriseren.

D (D-waarde)

D = KR-20 - AR.
In deze kolom is het verschil vermeld tussen de betrouwbaarheid van de totale toets en de AR van het betreffende item. Hoe positiever de D-waarde, hoe groter de bijdrage van een item aan de betrouwbaarheid. Een item met een D-waarde van 0 of met een negatieve D-waarde draagt niet of in negatieve zin bij aan de betrouwbaarheid.

RIR- EN RAR-WAARDEN

Per item is één r_{ir}-waarde vermeld en net zoveel r_{ar}-waarden als er afleiders (foute antwoorden) zijn. De r_{ir} is hierboven besproken; de r_{ar} is eenzelfde maat als de r_{ir} en wordt ook op dezelfde manier berekend. Dat wil zeggen de kandidaten die het betreffende foute antwoord hebben gekozen krijgen bij de berekening van de r_{ar} een score van 1 en de anderen een score van 0. Toetstechnisch gezien is het wenselijk om naar positieve r_{ir}-waarden te streven en naar negatieve r_{ar}-waarden. Een positieve r_{ar}-waarde kan duiden op een sleutelfout, immers relatief veel goede kandidaten hebben de

betreffende afleider (foute antwoord) als het goede antwoord aangemerkt.
Bij item 11 bijvoorbeeld heeft alternatief C een positieve r_{ar}. Omdat deze waarde groter is dan de r_{ir} is er alle aanleiding om dit item nader te bestuderen (zie ook in de kolom CODE). De nadere bestudering is bij dit item terecht omdat het goede antwoord eigenlijk C is. Maar om in de kolom CODE een melding te genereren is B als sleutel (goede antwoord) ingevoerd. Was C het goede antwoord geweest dan was de bijbehorende r_{ir}-waarde 13 geweest.

CODE

In deze kolom staat een signalering vermeld indien het item toetstechnisch gezien niet naar behoren heeft gefunctioneerd.
De codes A, B en C kunnen er in voorkomen en hebben de volgende betekenis.

A: Een r_{ar}-waarde van een item is groter dan of gelijk aan de r_{ir}-waarde.
B: De r_{ir}-waarde is kleiner dan of gelijk aan 0.
C: Een r_{ar}-waarde van een item is groter dan of gelijk aan 10.
Bij item 11 staan de codes A en C. Met code A wordt gesignaleerd dat de r_{ar} van de afleider C groter is dan de r_{ir} van het 'goede' antwoord B. Code C geeft aan dat de r_{ar}-waarde van C groter dan 10 is, in ons geval 13.

Algemene gegevens

Onderaan de toets- en itemanalyse staan de volgende gegevens:

- AANTAL KANDIDATEN
- GEMIDDELDE SCORE
- STANDAARDDEVIATIE
- GEMIDDELDE P-WAARDE
- BETROUWBAARHEID (KR-20)
- STANDAARDMEETFOUT

Bij AANTAL KANDIDATEN staat het aantal kandidaten waarop de toets- en itemanalyse gebaseerd is. Bij GEMIDDELDE SCORE is de

gemiddelde toetsscore vermeld en bij STANDAARDDEVIATIE is de standaarddeviatie van de toetsscores vermeld. Bij GEMIDDELDE P-WAARDE staat het gemiddelde van de p-waarden. De gemiddelde p-waarde ($\overline{p}$) is een maat voor de moeilijkheidsgraad van een toets en/of de vaardigheid van de kandidaten. De gemiddelde p-waarde kan berekend worden door alle p-waarden op te tellen en de som te delen door het aantal p-waarden (K) of, wat nauwkeuriger is, de gemiddelde score te delen door het maximaal aantal te behalen punten op de toets en het quotiënt te vermenigvuldigen met 100.
In formulevorm:

$$1.\ \overline{p} = \frac{\sum_{i=1}^{k} p_i}{k} = \frac{82 + 82 + 44 + \ldots + 82}{22} = 60.7 \qquad \text{of}$$

$$2.\ \overline{p} = \frac{GEMIDDELDE\ SCORE}{MAXIMALE\ TOETSSCORE} * 100 = \frac{26.67}{44} * 100 = 60.6$$

waarbij:
$\overline{p}$ = gemiddelde p-waarde (in %)
p_i = p-waarde van item i
k = aantal items

Bij BETROUWBAARHEID staat de KR-20 vermeld. Wij zullen in hoofdstuk 4 ingaan op deze betrouwbaarheidscoëfficiënt evenals op de STANDAARDMEETFOUT.

Verder zijn er nog 3 betrouwbaarheidscoëfficiënten vermeld (B1, B2 en BRS) die niet standaard berekend worden maar alleen als dit via een optie is aangevraagd (de optie 'XB' op regelnummer 027 in het voorloopbestand). In paragraaf 5.6 zullen we op deze optionele betrouwbaarheidscoëfficiënten ingaan. We willen echter op deze plaats waarschuwen spaarzaam gebruik te maken van deze optie omdat enerzijds de rekentijd van TIA exponentieel toeneemt en anderzijds deze coëfficiënten (nog) geen gemeengoed zijn in de testtheorie.

3.9 Correlatietabel (bijlage XVII)

De correlatietabel bevat de (produkt-momenten-)correlaties tussen de totaaltoets en de subtoetsen. Zoals wellicht bekend zegt een correlatie iets over de lineaire samenhang tussen twee variabelen, in ons geval twee (sub)toetsen. Indien een toets opgesplitst kan worden in subtoetsen kan het zinvol zijn

om de mate van samenhang tussen de subtoetsen te berekenen. Hoe groter het verband tussen de subtoetsen hoe groter de betrouwbaarheid van de toets (zie ook paragraaf 5.5). Indien een subtoets laag correleert met andere subtoetsen dan komt dat de betrouwbaarheid van de totale toets niet ten goede. Er zijn in ons voorbeeld 3 correlatiecoëfficiënten in de tabel afgedrukt; zo heeft de correlatie tussen de totaaltoets en subtoets 01 (het meerkeuze-deel) de waarde van 0.86 en de correlatie tussen subtoets 01 en subtoets 02 (het gesloten deel) de waarde van 0.54. Onderaan de tabel staan van de totaaltoets en van de subtoetsen de gemiddelde score en de standaarddeviatie vermeld. Bovendien is het aantal kandidaten vermeld waarop de berekeningen in de correlatietabel gebaseerd zijn.

3.10 Het dichotome bestand en het scorebestand (bijlage XVIII)

Op regel 007 in het voorloopbestand kunnen we zien dat gevraagd is het dichotome bestand (EXDICH) van de gesloten vragen te maken terwijl tevens gevraagd is het scorebestand (EXSCOR) te maken (regel 006). Het dichotome bestand kunnen we nodig hebben voor onderzoeksdoeleinden. Het scorebestand bevat de scores van de leerlingen.
Van het dichotome bestand en van het scorebestand zijn de eerste 5 records afgedrukt en terwille van de inzichtelijkheid zijn ook de eerste 5 (goede) invoerrecords (leerlingantwoordpatronen) nogmaals afgedrukt. Deze 5 records komen uit het bestand STP320D, een bestand bestaande uit 2596 records.
Bekijken we het eerste invoerrecord van het bestand STP320D dan zien we na de toets/leerlingidentificatie van 9 cijfers de respectievelijke codes 0,3,6 enz. Code 0 slaat op de eerste vraag en dit betekent dat deze vraag (een gesloten vraag) niet beantwoord is of dat er twee of meer streepjes op het antwoordblad hebben gestaan of dat een niet bestaand alternatief

is aangestreept. Bij de tweede vraag staat code 3 en dit betekent
dat de leerling C heeft aangestreept terwijl de 6 bij vraag 3 betekent dat er een E is aangestreept (zie ook paragraaf 3.3 blz.13). De vragen 1 t/m 22 zijn gesloten vragen en daar staan

dus codes terwijl vanaf vraag 23 (de open vragen) de op die vragen behaalde scores zijn vermeld.

De codes in het invoerbestand zijn in het dichotome bestand omgezet naar een 1 of een 0 waarbij 1 betekent dat de gesloten vraag goed is beantwoord en 0 dat de gesloten vraag fout is beantwoord. *(De code 0 (item overgeslagen etc.) in het invoerbestand wordt niet omgezet naar een 0 maar naar een spatie in het dichotome bestand).*

Kijken we nu naar het scorebestand dan zien we dat de eerste leerling een score van 78 heeft behaald op de totale toets. Deze score kunnen we reproduceren door in het dichotome bestand de score (ongewogen) op de gesloten vragen te bepalen, in ons geval 17, en in het invoerbestand de score op de open vragen te bepalen, in ons geval 34. Om deze score van 34 te kunnen reproduceren moeten we ons in herinnering roepen dat iedereen op de laatste open vraag de vervangende score van 7 punten heeft gekregen.
De totaalscore, inclusief 10 bonuspunten, van de eerste leerling bedraagt: 2 * 17 + 34 + 10 = 78 punten, waarbij de factor 2 de wegingsfactor van de gesloten vragen is.

3.11 De bestanden met gegevenstabellen

Als we de toets- en itemgegevens niet verloren willen laten gaan of voor andere doeleinden willen gebruiken zoals voor itembanken, dan kunnen we deze in gegevenstabellen vastleggen via een optie in het voorloopbestand (regelnummers 008 en 009). De gegevens van deze toets zijn in de bestanden EXTOET, EXFREQ en EXITEM vastgelegd. Voor een beschrijving van deze tabellen verwijzen we naar W.J. van Daal (1992).

Tot zover de mogelijkheden van en de toelichting op TIA. In de volgende hoofdstukken zullen we onderwerpen uit de klassieke testtheorie aan de orde stellen waarbij we van TIA uitgaan. Verder worden nog onderwerpen behandeld als cesuurbepaling, cijfergeving en normhandhaving. Leidraad bij de komende hoofdstukken is geweest om enerzijds de laatste ontwikkelingen op het gebied van de klassieke testtheorie aan te stippen en anderzijds antwoord te geven op veel voorkomende vragen.

4 Betrouwbaarheid en standaardmeetfout

De betrouwbaarheid is de mate waarin men staat kan maken op de meetresultaten, dat wil zeggen de mate waarin de scores consistent, nauwkeurig en reproduceerbaar zijn, kortom vrij van

meetfouten. De beste manier om de betrouwbaarheid te bepalen is die waarbij de correlatie tussen de scores op twee parallelle (gelijkwaardige) toetsen berekend wordt. In de praktijk is dit (vaak) niet mogelijk en zijn er methoden ontwikkeld om de betrouwbaarheid bij één toetsafname te bepalen. In paragraaf 4.1 worden indien er sprake is van één toetsafname de meest bekende schattingen van de betrouwbaarheid behandeld namelijk de KR-20 en α. De in deze paragraaf behandelde schattingen van de betrouwbaarheid geven aan in hoeverre de vragen van een toets onderling samenhangen in statistisch opzicht. Met andere woorden de KR-20 en α zijn maten voor de interne consistentie van een toets en als zodanig schattingen van de betrouwbaarheid. De betrouwbaarheid wordt uitgedrukt in een getal dat (theoretisch) tussen 0 en 1 ligt waarbij 0 de ondergrens is en 1 de bovengrens.

Naast de betrouwbaarheid is de standaardmeetfout een maat voor de psychometrische kwaliteit van een toets. Met de standaardmeetfout s_e kan men aangeven binnen welk interval de ware score van een leerling ligt (zie ook paragraaf 5.1). Heeft een leerling bijvoorbeeld een score van 70 behaald en de standaardmeetfout is 6.2 dan kan men met vrij grote zekerheid zeggen dat de ware score van de betreffende leerling tussen de 58 en 82 punten ligt.

4.1 Een schattingsmethode voor de betrouwbaarheid en de standaardmeetfout

De betrouwbaarheid van een toets (r) wordt vaak met de volgende formule geschat:

$$r = \frac{k}{k-1}\left[\frac{s_x^2 - \sum_{i=1}^{k} s_i^2}{s_x^2}\right] = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{s_x^2}\right] \qquad (1)$$

waarbij:
k = aantal vragen
s_i^2 = variantie van de scores bij vraag i
s_x^2 = variantie van de totaalscores

De via bovenstaande formule berekende waarde van r heet coëfficiënt alpha (α) bij polytoom gescoorde vragen zoals in openvraag-toetsen en KR-20 bij dichotoom gescoorde vragen zoals in gesloten-toetsen.

Bij dichotoom gescoorde vragen kan de formule, omdat $s_i^2 = p_i q_i$, als volgt geschreven worden:

$$KR-20 = \frac{k}{k-1}\left[\frac{s_x^2 - \sum_{i=1}^{k} p_i q_i}{s_x^2}\right] \qquad (2)$$

waarbij:

p_i = fractie van de kandidaten die item i goed heeft

q_i = fractie van de kandidaten die item i fout heeft ($q_i = 1 - p_i$)

Ter informatie:
De KR-20 formule is door Kuder en Richardson (1937) ontwikkeld en is de 20e in een reeks van 22.

Een andere formule om de betrouwbaarheid te schatten is:

$$r = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{\left(\sum_{i=1}^{k} r_{it} s_i\right)^2}\right] \qquad (3)$$

waarbij:

k = aantal vragen

s_i^2 = variantie van de scores bij vraag i

s_i = standaarddeviatie van de scores bij vraag i ($s_i = \sqrt{s_i^2}$)

r_{it} = correlatie tussen de scores van vraag i en de scores van de totaaltoets

Formule (3) is algebraïsch gelijk aan de formules (1) en (2); formule (3) laat echter duidelijk het verband zien tussen de r_{it} en de betrouwbaarheid, namelijk hoe hoger de r_{it} van de vragen hoe hoger de betrouwbaarheid van de toets.

De standaardmeetfout (s_e) wordt met de volgende formule berekend:

$$s_e = s_x\sqrt{1-r} \qquad (4)$$

waarbij:
s_e = standaardmeetfout
s_x = standaarddeviatie van de totaalscores
r = betrouwbaarheid

4.2 De interpretatie van de betrouwbaarheid

Alvorens we overgaan tot de interpretatie van de betrouwbaarheid zullen we eerst een theoretisch kader schetsen (zie ook Traub en Rowley, 1991).
We gaan bij de klassieke testtheorie ervan uit dat de toetsscore (X) van een kandidaat uit twee componenten bestaat, namelijk uit een component ware score (T) en een component meetfout (E). In formulevorm:

$$X = T + E \qquad (5)$$

De toetsscore (X) is de score die een kandidaat behaald heeft. De hoogte van de score wordt grotendeels bepaald door de vaardigheid van de kandidaat en de moeilijkheidsgraad van de toets. De score van een kandidaat wordt verder nog beïnvloed door andere factoren zoals mentale en lichamelijke conditie van de kandidaat, de kwaliteit van de vragen, pech of geluk bij raden, de verdeling van de items over de stof enz. Omdat genoemde storende invloeden altijd aanwezig zullen zijn wordt X als een schatting beschouwd van de ware score T en het verschil tussen T en X is dan de meetfout.

Als we nu uitgaan van formule (5) dan kan het volgende afgeleid worden:

1. $\overline{X} = \overline{T}$ (6)

2. $s_x^2 = s_T^2 + s_e^2$ (7)

Uit formule (6) blijkt dat het gemiddelde van de toetsscores van alle kandidaten ($\overline{X}$) gelijk is aan het gemiddelde van de ware scores ($\overline{T}$). Uit formule (7) blijkt dat de variantie van de toetsscores (s_x^2) gelijk is aan de som van de variantie van de ware scores (s_T^2) en de variantie van de meetfouten (s_e^2).

Nu is de betrouwbaarheid als volgt gedefinieerd:

$$r = \frac{s_T^2}{s_X^2} = \frac{s_X^2 - s_e^2}{s_X^2} = 1 - \frac{s_e^2}{s_X^2} \qquad (8)$$

waarbij:

r = betrouwbaarheid
s_T^2 = variantie van de ware scores
s_X^2 = variantie van de toetsscores
s_e^2 = variantie van de meetfouten

Uit formule (8) blijkt dat een betrouwbaarheidscoëfficiënt aangeeft welk deel van de variantie van toetsscores uit ware variantie bestaat en welk deel uit toevalsvariantie. Indien bijvoorbeeld de betrouwbaarheid 0.77 is (geschat bijv. via **α** of KR-20) dan kunnen we zeggen dat 77% van de variantie van toetsscores uit ware variantie bestaat en 23% uit toevalsvariantie. Een hoge betrouwbaarheid betekent dus dat de toevalsvariantie vrijwel afwezig is en dat de toetsscore (X) een goede afspiegeling is van de ware score (T). De betrouwbaarheid van 0.77 kan verder geïnterpreteerd worden als de mate waarin de scores van een toets correleren met de scores van een (denkbeeldige) paralleltoets.

5 Speciale Onderwerpen

5.1 Nauwkeurigheid van de score van een kandidaat

We zullen twee methoden beschrijven om de nauwkeurigheid van de score van een kandidaat te bepalen. De eerste methode is in hoofdstuk 4 al aangestipt en bij deze methode wordt de toetsscore X als een schatting beschouwd van de ware score met $s_e = s_x\sqrt{1-r}$ als de daarbij behorende standaardmeetfout.

Samengevat:

$$\hat{T} = X \text{ en } s_e = s_x\sqrt{1-r} \qquad (9)$$

waarbij:
$\hat{T}$ = een schatting van de ware score van een kandidaat
X = toetsscore van een kandidaat
s_e = standaardmeetfout
s_x = standaarddeviatie van de toetsscores
r = betrouwbaarheidscoëfficiënt

Op grond van bovenstaande formules kunnen we nu een interval bepalen waarbinnen de ware score (T) van een kandidaat ligt. Het 95% betrouwbaarheidsinterval bijvoorbeeld kunnen we als volgt berekenen:

$$\hat{T} - 1.96\ s_e < T < \hat{T} + 1.96\ s_e$$

Bovengenoemde methode om de nauwkeurigheid van een score van een kandidaat te bepalen wordt vaak gehanteerd hoewel hij theoretisch gezien niet de meest correcte is. Een betere methode is om, naast de toetsscore van een kandidaat en de betrouwbaarheid, uit te gaan van de gemiddelde toetsscore van alle kandidaten (zie o.a. Dousma en Horsten, 1980; Drenth en Sijtsma, 1990).
De formules zien er als volgt uit:

$$\hat{T} = r(X - \overline{X}) + \overline{X} \text{ en } s_e = s_x\sqrt{1-r} * \sqrt{r} \qquad (10)$$

waarbij:

$\hat{T}$ = een schatting van de ware score van een kandidaat

r = betrouwbaarheidscoëfficiënt

X = toetsscore van een kandidaat

$\overline{X}$ = gemiddelde toetsscore van alle kandidaten

s_e = standaardmeetfout

s_x = standaarddeviatie van de toetsscores

Met behulp van $\hat{T}$ en s_e kunnen we ook hier het 95% betrouwbaarheidsinterval voor de ware score bepalen.

We zullen ter verduidelijking een rekenvoorbeeld van beide methoden presenteren. Stel een kandidaat heeft op de toets een score van 70. De gemiddelde toetsscore is 60.5, de betrouwbaarheid 0.77 en de standaardmeetfout 6.2.

Dus: $X = 70$; $\overline{X} = 60.5$; $r = 0.77$; $s_e = 6.2$

Volgens de eerste methode kunnen we zeggen dat de score van 70 een schatting van de ware score is ($\hat{T}$) en dat de ware score (T) met een betrouwbaarheid van 95% tussen de waarden 57.8 en 82.2 ligt.

Dus: $\hat{T} = 70$ en $57.8 < T < 82.2$

Volgens de tweede methode heeft een kandidaat met een score van 70 een geschatte ware score ($\hat{T}$) van 67.8 (0.77 * (70 - 60.5) + 60.5 = 67.8) en een standaardmeetfout van 5.4 (6.2 * $\sqrt{0.77}$ = 5.4). De ware score (T) ligt met een betrouwbaarheid van 95% tussen de waarden 57.1 en 78.5.

Dus: $\hat{T} = 67.8$ en $57.1 < T < 78.5$

Uit de resultaten blijkt dat met de tweede methode een kleiner interval voor de ware score wordt gevonden. Vooral bij betrouwbaarheden lager dan 0.75 en bij toetsscores die ver van het gemiddelde af liggen is de winst aanzienlijk. Tenslotte willen we naar een artikel van Harvill (1991) verwijzen dat specifiek gewijd is aan de in deze paragraaf behandelde materie.

5.2 Factoren die de betrouwbaarheid beïnvloeden

Tot op zekere hoogte kunnen we de betrouwbaarheid van een toets beïnvloeden (Frisbie, 1988). De volgende factoren hebben effect op de betrouwbaarheid:

- toetslengte
- de effectiviteit van de vragen
- samenstelling van de groep kandidaten
- objectiviteit van de scoring

In de volgende hoofdstukken zullen we aandacht schenken aan elk van de afzonderlijke factoren.

5.2.1 Toetslengte

Het ligt voor de hand dat we bijvoorbeeld met 50 vragen de vaardigheid van een kandidaat betrouwbaarder kunnen bepalen dan met 10 vragen. De invloed van de toetslengte op de betrouwbaarheid wordt duidelijk als we formule (3) uit paragraaf 4.1 omwerken:

$$r = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{\left(\sum_{i=1}^{k} r_{it} s_i\right)^2}\right] =$$

$$= \frac{k}{k-1}\left(1 - \frac{s_1^2 + s_2^2 + \ldots + s_k^2}{(s_1 r_{1t} + s_2 r_{2t} + \ldots + s_k s_{kt})^2}\right) \approx$$

$$\approx \frac{k}{k-1}\left(1 - \frac{k(\overline{s}_i)^2}{k^2 * (\overline{s_i r}_{it})^2}\right) \approx$$

$$\approx \frac{k}{k-1}\left(1 - \frac{1}{k(\overline{r}_{it})^2}\right)$$

Indien we $\frac{k}{k-1}$ nog verwaarlozen wat probleemloos kan bij niet te kleine toetsen, dan komen we tot:

$$r \approx 1 - \frac{1}{k(\overline{r}_{it})^2}. \qquad (11)$$

waarbij:

r = betrouwbaarheid

k = aantal vragen

$\overline{r}_{it}$ = het gemiddelde van de r_{it}-waarden

Dus hoe meer vragen (k wordt groter) hoe groter de betrouwbaarheid (r nadert 1).

Door Spearman en Brown (zie ook Gulliksen, 1950) is eveneens in formulevorm een verband weergegeven tussen de betrouwbaarheid van een toets en de toetslengte. Bij toetsverlenging (of verkorting) wordt verondersteld dat de toegevoegde items qua eigenschappen overeenkomen met de oorspronkelijke items.
De Spearman en Brown formule voor het verlengen of verkorten van een toets heeft de volgende vorm:

$$r_v = \frac{\frac{k_v}{k_o} * r_o}{1 + (\frac{k_v}{k_o} - 1) * r_o} \qquad (12)$$

waarbij:

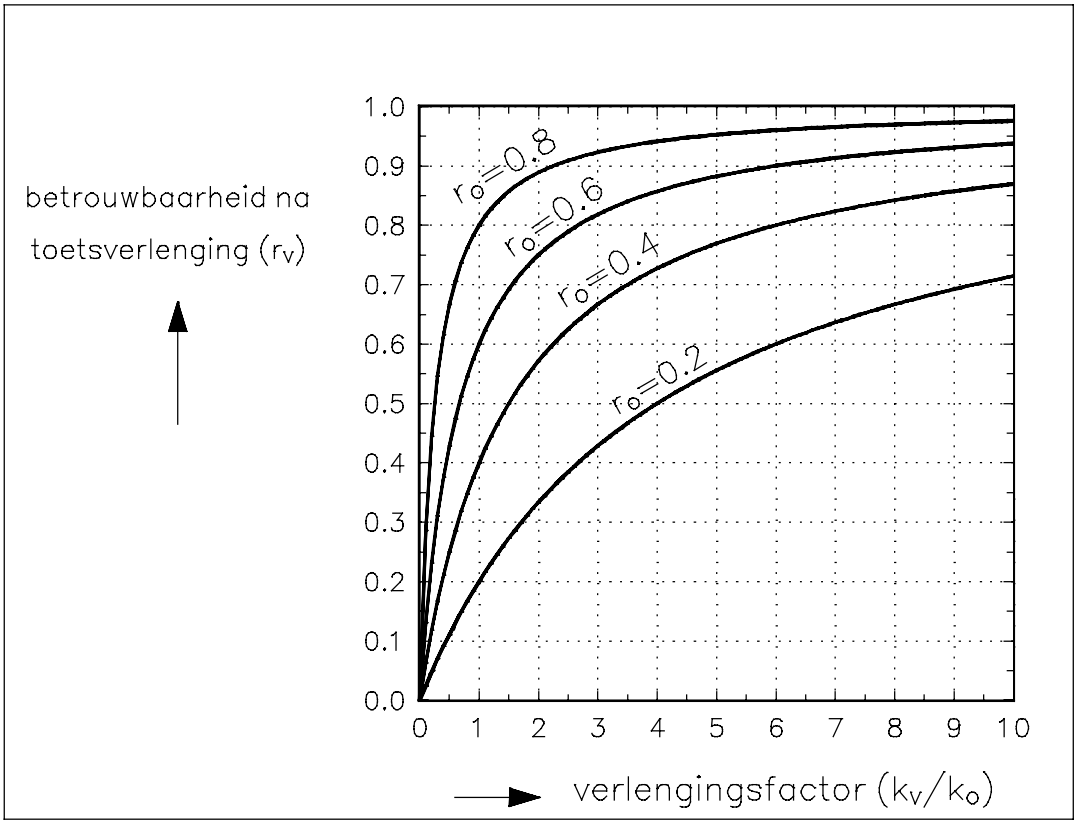
r_v = betrouwbaarheidscoëfficiënt van de verlengde of verkorte toets

r_o = betrouwbaarheidscoëfficiënt van de oorspronkelijke toets

k_v = aantal items van de verlengde of verkorte toets

k_o = aantal items van de oorspronkelijke toets

Het verband tussen de verlengingsfactor $\frac{k_v}{k_o}$ en r_v is in onderstaande grafiek voor enkele waarden van r_o weergegeven.



Figuur 2. Betrouwbaarheid (r_v) na toetsverlenging van toetsen met oorspronkelijke betrouwbaarheden (r_o) van 0.2, 0.4, 0.6 en 0.8.

Als bijvoorbeeld een toets van 25 items en met een betrouwbaarheid van 0.60 verlengd wordt tot een toets van 50 items dan zien we dat de betrouwbaarheid ongeveer 0.75 wordt.

We kunnen formule (12) zodanig herschrijven dat $\frac{k_v}{k_o}$ in het linkerlid komt. Formule (12) komt er dan als volgt uit te zien:

$$\frac{k_v}{k_o} = \frac{r_v(r_o-1)}{r_o(r_v-1)} \qquad (12a)$$

Met bovenstaande formule kunnen we gemakkelijk uitrekenen met hoeveel items een toets verlengd moet worden als we de betrouwbaarheid tot een bepaalde waarde willen laten stijgen. Het meerkeuze-deel in ons voorbeeld heeft een betrouwbaarheid van 0.64 (Bijlage XI). Willen we de betrouwbaarheid verhogen naar 0.70 dan substitueren we 0.64 en 0.70 in formule (12a).

$$\frac{k_v}{k_o} = \frac{0.70(0.64-1)}{0.64(0.70-1)} = 1.3125$$

Om een betrouwbaarheid van 0.70 te bereiken moet de toets met de factor 1.3125 verlengd worden. Dat wil zeggen het aantal items moet toenemen van 22 naar 1.3125 * 22 = 29.

5.2.2 De effectiviteit van de vragen

Een tweede factor die de betrouwbaarheid van een toets beïnvloedt is de effectiviteit van de vragen. Wil een vraag in psychometrisch opzicht goed zijn en dus bijdragen aan de betrouwbaarheid dan moet een vraag niet te moeilijk en niet te gemakkelijk zijn voor de groep van kandidaten. Met de vraag moeten we dus spreiding kunnen genereren tussen de kandidaten. Verder moet de vraag discrimineren, d.w.z. de vaardige kandidaten moeten de vraag beter kunnen maken dan de minder vaardige.
De effectiviteit van een vraag wordt als volgt berekend:

$$\text{effectief gewicht} = \frac{RIT * SD}{s_X} \qquad (13)$$

waarbij:
RIT = r_{it}-waarde van een vraag
SD = standaarddeviatie van een vraag
s_X = standaarddeviatie van de totale toets

N.B. De teller in formule (13) wordt in de Engelstalige literatuur de 'item reliability index' genoemd (Gulliksen, 1950).

5.2.3 De samenstelling van de groep kandidaten

De homogeniteit van de groep kandidaten, de derde factor die de betrouwbaarheid beïnvloedt, bepaalt de spreiding van de toetsscores. Hoe groter de verschillen in vaardigheid, hoe groter de spreiding in toetsscores en uit formule (8) in paragraaf 4.2 blijkt dat de spreiding in toetsscores (s_x^2) effect heeft op de betrouwbaarheid. Grotere varianties hebben hogere betrouwbaarheden tot gevolg. De grootste variantie wordt verkregen als de ene helft van de groep kandidaten alle vragen goed heeft en de andere helft van de groep kandidaten alle vragen fout.

5.2.4 Objectiviteit van de scoring

De objectiviteit van de scoring is de vierde factor die effect heeft op de betrouwbaarheid. Bij gesloten vragen is deze objectiviteit gegarandeerd. Altijd wordt immers dezelfde score gegeven, onafhankelijk van wie de toets nakijkt.

Bij <u>open vragen</u> is uit onderzoek gebleken dat de beoordeling nogal eens wil verschillen van beoordelaar tot beoordelaar en deze spreidingsbron verlaagt de betrouwbaarheid.
De beoordelaarsovereenstemming kunnen we uitdrukken in de volgende maat (Crocker en Algina, 1986; Goldebeld, 1983):

$$r_1 = \frac{\sigma_k^2}{\sigma_k^2 + \frac{1}{b}\left(\sigma_b^2 + \sigma_{kb}^2\right)} \qquad (14)$$

waarbij:
r_1 = beoordelaarsovereenstemming
σ_k^2 = variantiecomponent kandidaten
σ_b^2 = variantiecomponent beoordelaars
σ_{kb}^2 = variantiecomponent kandidaten x beoordelaars (residu)
b = aantal beoordelaars

Schattingen van bovengenoemde variantiecomponenten kunnen niet rechtstreeks uit een toets- en itemanalyse voor open vragen verkregen worden. Daarvoor is speciaal onderzoek vereist in de vorm van een aantal leerlingwerken die door meerdere beoordelaars onafhankelijk van elkaar beoordeeld worden.
Wanneer de beoordelingen van de beoordelaars perfect overeenstemmen zijn σ_b^2 en σ_{kb}^2 gelijk aan 0 en is de overeenstemming gelijk aan 1.00. Hoe lager de overeenstemming hoe groter de componenten σ_b^2 en σ_{kb}^2 in verhouding tot σ_k^2.
Door in de formule verschillende waarden van b in te vullen kan geschat worden hoe groot de overeenstemming is bij gebruikmaking van verschillende aantallen beoordelaars. Indien $b = 1$ geeft de formule een schatting van de overeenstemming tussen de beoordeling van één beoordelaar en de beoordeling van een willekeurig andere beoordelaar. Indien $b = 2$ geeft de formule een schatting van de overeenstemming tussen de gemiddelde beoordeling van 2 beoordelaars en de gemiddelde beoordeling van 2 andere willekeurige beoordelaars.

Gebleken is dat de beoordelaarsovereenstemming de betrouwbaarheid bij benadering in de volgende mate beïnvloedt:

$$\text{betrouwbaarheid} \approx \alpha * r_1 \qquad (15)$$

waarbij:
α = betrouwbaarheid van een openvraag-toets
r_1 = beoordelaarsovereenstemming

Dus de bij een openvraag-toets vermelde schatting van de betrouwbaarheid (α) geeft, wanneer de beoordelaarsovereenstemming niet maximaal is een te rooskleurig beeld van de betrouwbaarheid.
In ons voorbeeld is de betrouwbaarheid van het openvraaggedeelte 0.67. Bij volledige overeenstemming tussen de beoordelaars blijft de betrouwbaarheid 0.67 maar als de overeenstemming niet volledig is (0.90 is niet uitzonderlijk laag) dan daalt de betrouwbaarheid naar 0.60.

Opgemerkt moet nog worden dat de formule $\alpha * r_1$ een vuistregel is. In Thorndike (1982) en in Goldebeld (1983) staat een formule om de betrouwbaarheid van een openvraag-toets exact te bepalen. Beide formules zijn gebaseerd op de veronderstelling dat ieder leerlingwerk door een andere beoordelaar nagekeken wordt.

5.3 Raden en correctie voor raden

Bij gesloten vragen kan de kandidaat altijd raden indien hij het antwoord niet weet. Scores kunnen voor raden gecorrigeerd worden. In de literatuur wordt aan dit onderwerp veel aandacht besteed zonder dat men tot een eenduidige oplossing komt (zie o.a. Drenth en Sijtsma, 1990). Het effect van correctie voor raden is vaak gering op de betrouwbaarheid terwijl de gecorrigeerde scores van de kandidaten hoog tot zeer hoog correleren met de oorspronkelijke scores. Bij het Cito wordt geen correctieformule voor raden gebruikt.

5.4 Het optimale aantal alternatieven bij meerkeuze-items

Voor het realiseren van de maximale betrouwbaarheid kunnen we in z'n algemeenheid stellen dat driekeuze-items de voorkeur verdienen. Althans binnen de randvoorwaarde dat ak constant is, waarbij a het aantal alternatieven per item is en k het totale aantal items. Deze conclusie trekt ook Van den Brink (1979) waarbij hij de misschien wat meer realistische voorwaarde hanteert namelijk dat $(a + 1)k$ constant moet zijn. Hij eindigt de discussie met: 'Het wordt misschien tijd dat bijvoorbeeld het Cito het roer eens omgooit. Vierkeuze-items leveren voortreffelijke driekeuze-items op indien men het slechtste alternatief aan de vuilnisman meegeeft!'.
Melse en Mets (1984) hebben empirisch bij een C-examen Duits onderzocht wat het effect op de betrouwbaarheid is indien een oorspronkelijke vierkeuze-toets omgezet wordt naar een driekeuze-

toets door het slechtste alternatief weg te laten. Zij komen tot de conclusie dat het wat de betrouwbaarheid betreft niet uitmaakt of een toets uit vierkeuze-items of uit driekeuze-items bestaat.

Kunnen we dus aan de randvoorwaarde dat $(a + 1)k$ constant moet zijn voldoen, dan is het te prefereren om bijvoorbeeld in plaats van 50 vierkeuze-items 63 driekeuze-items af te nemen. Dit zal naast de betrouwbaarheid ook de validiteit doen toenemen. Maar zelfs als we niet aan deze randvoorwaarde kunnen voldoen hoeven driekeuze-items niet slechter te zijn dan vierkeuze-items.

N.B. 1. De opmerking van Van den Brink met betrekking tot het Cito is enigszins achterhaald. In die zin dat bij de meerkeuze-items in de centrale examens van het voortgezet onderwijs het aantal alternatieven kan variëren van 3 t/m 6. De achterliggende gedachte hierbij is dat het aantal alternatieven moet afhangen van het aantal reële antwoorden.

2. Door sommigen wordt betwijfeld of het oplossen van driekeuze-items minder tijd in beslag neemt dan het oplossen van vierkeuze-items. Het voordeel is dan gelegen in de constructiefase: men hoeft geen vierde alternatief te bedenken.

5.5 Vakbetrouwbaarheid

Met vakbetrouwbaarheid wordt het volgende bedoeld. Stel dat een kandidaat beoordeeld wordt door middel van twee toetsen, een gesloten-toets en een openvraag-toets. Gewoonlijk wordt van elke toets apart de betrouwbaarheid berekend. Een interessant gegeven is de betrouwbaarheid van beide toetsen samen, oftewel het vak (zie ook Nuttall en Willmot, 1972).

De formule voor het schatten van de vakbetrouwbaarheid ziet er als volgt uit:

$$r_{vak} \approx \frac{r_x s_x^2 + r_y s_y^2 + 2 r_{xy} s_x s_y}{s_x^2 + s_y^2 + 2 r_{xy} s_x s_y} \qquad (16)$$

waarbij:

r_{vak} = vakbetrouwbaarheid
r_x = betrouwbaarheid van toets x
s_x^2 = variantie van de toetsscores van toets x
s_x = standaarddeviatie van de toetsscores van toets x ($s_x = \sqrt{s_x^2}$)
r_y = betrouwbaarheid van toets y
s_y^2 = variantie van de toetsscores van toets y
s_y = standaarddeviatie van de toetsscores van toets y ($s_y = \sqrt{s_y^2}$)
r_{xy} = correlatie tussen de scores van toets x en toets y

Aangezien we de betrouwbaarheid en de standaarddeviatie van de afzonderlijke toetsen al hebben, hoeven we alleen nog maar de correlatie tussen de scores op beide toetsen te weten om de betrouwbaarheid van het vak te schatten.

Voorbeeld:

Toets x : r_x = 0.70 s_x = 20
Toets y : r_y = 0.80 s_y = 25
Correlatie : r_{xy} = 0.60

Toets x heeft dus een betrouwbaarheid van 0.70 en een standaarddeviatie van 20, deze waarden zijn bij toets y resp. 0.80 en 25. De correlatie tussen de scores van beide toetsen bedraagt 0.60. Substitueren we deze waarden in bovenstaande formule dan resulteert dat in 0.85 als schatting van de vakbetrouwbaarheid. De vakbetrouwbaarheid is groter dan de betrouwbaarheid van elke afzonderlijke toets en dat zal ons niet echt verbazen. Er is immers in feite sprake van toetsverlenging (zie ook paragraaf 5.2.1).

5.6 Enkele alternatieven voor de KR-20 en α

(Bron: Testtheorie, Drenth, P.J.D. en Sijtsma, K., 1990).

Guttman en Ten Berge en Zegers hebben alternatieven ontwikkeld om de betrouwbaarheid van een toets te schatten. Bij TIA is het mogelijk naast **α** of de KR-20 enkele van deze alternatieven te laten berekenen.

Guttman (1945) heeft een betrouwbaarheidscoëfficiënt ontwikkeld, de lambda-coëfficiënt (λ_2), welke in TIA gepresenteerd wordt als B1. Lambda-2 maakt samen met **α** deel uit van een oneindige lange reeks van betrouwbaarheidscoëfficiënten. (Ten Berge en Zegers, 1978). De coëfficiënten in deze reeks worden mu-coëfficiënten (μ) genoemd en ze zijn genummerd te beginnen bij nul: μ_0, μ_1, μ_2, etc. Een kenmerk van de μ-reeks is dat de coëfficiënten kunnen

worden geordend naar oplopende grootte, waarbij de nummering correspondeert met hun plaats in de ordening: $\mu_0 \leq \mu_1 \leq \mu_2$, etc. In deze reeks is μ_0 identiek aan α en aan λ_3 van Guttman. Verder is μ_1 gelijk aan λ_2 en wordt in TIA met B1 aangeduid terwijl μ_2 in TIA aangeduid wordt met B2.
Nog een andere maat voor de betrouwbaarheid is ontwikkeld door Ten Berge (1984). Deze maat ($\alpha*$) wordt in onze toets- en itemanalyse aangeduid met BRS en is volgens Ten Berge zinvol wanneer er sprake is van 'restriction of range'. Hiermee wordt bedoeld dat de populatie homogeen is en extreem laag en hoog scorende kandidaten niet voorkomen.

Samenvattend:
$\alpha = \mu_0 = \lambda_3$
B1 $= \mu_1 = \lambda_2$
B2 $= \mu_2$
BRS $= \alpha*$

Ten Berge en Zegers raden aan om B2 te berekenen indien de toets uit enkele items bestaat. Verder adviseren zij om in het algemeen B1 te berekenen.

Bij de totale toets en de twee subtoetsen van ons voorbeeld hebben we de vier betrouwbaarheidscoëfficiënten de volgende waarden.

Tabel 1. Vier betrouwbaarheidsmaten

	gehele toets (31 vragen)	meerkeuze-deel (22 vragen)	openvraag-deel (9 vragen)
α	.77	.64	.67
B1	.78	.64	.69
B2	.78	.64	.69
BRS	.97	.97	.92

Uit bovenstaande tabel blijkt dat B1 en B2 vrijwel niet afwijken van α. De BRS mag volgens Ten Berge niet met α vergeleken worden maar moet als een grootheid naast α beschouwd worden. Benadrukt moet worden dat genoemde alternatieven voor α en KR-20 (nog) niet algemeen gebruikt worden.

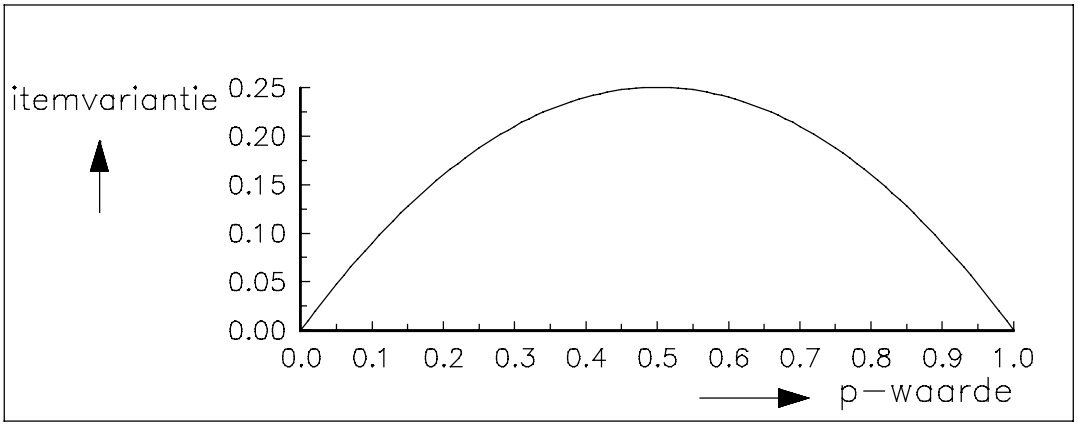
6 Normen voor toets- en itemindices

Alvorens normen te presenteren willen we erop wijzen dat deze eigenlijk alleen maar gehanteerd mogen worden bij toetsen waarmee we willen selecteren en kwalificeren. Gebruikt men de toetsen voor andere doeleinden zoals voortgangscontrole dan mag men minder hoge eisen aan de toets- en itemindices stellen (Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen, 1988; Frisbie, 1988). Verder willen we erop attenderen dat de indices bij kleine aantallen leerlingen een relatief grote steekproeffout hebben zodat voorzichtigheid geboden is bij de interpretatie ervan. In hoofdstuk 10 gaan we in op de nauwkeurigheid van diverse indices.

We zullen voor de p/p'-waarden, de r_{it}-waarden, de betrouwbaarheid en de r_{ir}/r_{ar}-waarden aangeven waaraan zij moeten voldoen wil men van psychometrisch geslaagde vragen spreken c.q. van een psychometrisch geslaagde toets.

6.1 Normen voor p/p'-waarden

Een vraag kan maximaal bijdragen aan de variantie van een toets indien de helft van de kandidaten de vraag goed heeft en de andere helft van de kandidaten de vraag fout. De p/p'-waarde is dan 0.50. Onderstaande figuur geeft voor een gesloten vraag het verband weer tussen de p-waarde en de maximale itemvariantie (s_i^2).



Figuur 3. Verband tussen de p-waarde en maximale itemvariantie.

Een p-waarde van 0.50 geeft een maximale itemvariantie van 0.25 maar alle waarden tussen 0.70 en 0.30 geven eveneens nog een bevredigende hoge itemvariantie. (Bij een gesloten vraag is

$s_i^2 = p * q$, dus bij een p-waarde van 0.70 of 0.30 bedraagt de itemvariantie 0.21).

6.2 Normen voor r_{it}-waarden

Absolute normen voor r_{it}-waarden zal men in de literatuur niet gauw aantreffen. Zoals bekend kan een r_{it}-waarde variëren tussen -1 en +1 waarbij we in de psychometrie streven naar een zo hoog mogelijke r_{it}. Ebel en Frisbie (1986) hebben wel een beoordeling opgesteld voor de r_{it}-waarden. In onderstaande tabel staat deze vermeld.

Tabel 2. Normen voor r_{it}-waarden

r_{it}-waarde	itembeoordeling
0.40 en hoger	zeer goed
0.30 - 0.39	goed
0.20 - 0.29	twijfelgeval
0.19 en lager	slecht

Omdat de grootte van de r_{it} afhankelijk is van het aantal items in een toets moet men strikt genomen bovenstaande normen alleen hanteren bij r_{it}'s die gecorrigeerd zijn voor toetslengte. De correctie kan uitgevoerd worden met de reeds eerder genoemde correctieformule van Henrysson (Thorndike, 1982).
De correctieformule heeft de volgende vorm:

$$r_{itc} = \sqrt{\frac{k}{k-1}} * \frac{r_{it}s_x - s_i}{\sqrt{s_x^2 - \sum_{i=1}^{k} s_i^2}} \qquad (17)$$

waarbij:
r_{itc} = gecorrigeerde r_{it}-waarde
r_{it} = r_{it}-waarde
k = aantal items
s_x^2 = variantie van de toetsscores
s_x = standaarddeviatie van de toetsscores ($s_x = \sqrt{s_x^2}$)
s_i^2 = variantie van de scores bij item i
s_i = standaarddeviatie van de scores bij item i ($s_i = \sqrt{s_i^2}$)

Vanwege het geringe effect kan de correctie achterwege blijven indien de items afkomstig zijn uit toetsen met 40 of meer items.

6.3 Normen voor de betrouwbaarheid

Voor de betrouwbaarheid zijn (ook) geen absolute normen te geven. Een ondergrens van 0.85 wordt wel genoemd indien de toets het enige middel is geweest is om de vaardigheid van een kandidaat te bepalen. Wanneer dat niet het geval is zijn lagere ondergrenzen acceptabel waarbij in de literatuur 0.65 wel als laagste ondergrens wordt genoemd (Frisbie, 1988).

Een ander licht op de betrouwbaarheid kunnen we laten schijnen door de betrouwbaarheid in verband te brengen met het percentage niet-consistente beslissingen bij diverse percentage gezakten/onvoldoendes. Dit verband is te zien in onderstaande tabel die afkomstig is uit Dousma en Horsten (1980).

Tabel 3. Percentages niet-consistente beslissingen als functie van het percentage gezakten en de betrouwbaarheid

percentage gezakten	Betrouwbaarheid						
	0.0	0.50	0.60	0.70	0.80	0.90	1.00
5	10	8	7	6	5	4	0
10	18	14	12	11	9	6	0
15	26	18	17	14	12	8	0
20	32	23	20	17	14	10	0
25	38	26	23	20	16	11	0
30	42	29	25	22	18	12	0
35	46	31	27	23	19	13	0
40	48	32	29	24	20	14	0
45	50	33	29	25	20	14	0
50	50	33	30	25	20	14	0

Voorbeeld:
In de tabel kan gelezen worden dat bij een betrouwbaarheid van 0.80 en bij 30% gezakten, het percentage niet consistente beslissing 18% is. Dat wil zeggen 9% van de gezakten (dus bijna 1/3 deel) zou tot de geslaagden kunnen hebben behoord en 9% van de geslaagden tot de gezakten. Dus voor 18% van alle leerlingen had de beslissing anders kunnen uitpakken.
Opgemerkt moet nog worden dat het gebruik van tabel 3 alleen zinvol is wanneer de toetsscores (bijna) normaal zijn verdeeld en wanneer de betrouwbaarheidscoëfficiënt opgevat wordt als de correlatie tussen een toets en een (denkbeeldige) paralleltoets.

6.4 Normen voor r_{ir}/r_{ar}-waarden

Bij het Cito worden in een toets- en itemanalyse van gesloten vragen de items van codes (A,B of C) voorzien indien de r_{ir}/r_{ar}-waarden niet aan bepaalde normen voldoen (zie ook blz. 28). De normen zijn aan de praktijk ontleend en voldoen goed omdat aan een 'gesignaleerd' item vaak inhoudelijke bezwaren blijken te kleven. Indien een vraag inhoudelijk niet juist is geweest kan men maatregelen treffen zoals twee of meer antwoorden goedkeuren, iedereen eenzelfde score geven of de vraag laten vervallen. De laatste twee maatregelen zijn eveneens mogelijk bij open vragen.

7 Vaststellen van de cesuur

De plaats van de cesuur, de grens tussen onvoldoende en voldoende, is voor kandidaten een kwestie van wel of geen diploma krijgen, het wel of niet toegelaten worden tot een bepaalde onderwijsinstelling, etc. Bij het bepalen van de cesuur kan men uitgaan van twee principes die in paragraaf 7.1 uitgewerkt worden. Binnen elk principe zijn er diverse methoden om de cesuur te bepalen.
Ondanks dat er in de literatuur zoveel cesuurbepalingsmethoden beschreven worden, is een waarschuwing voor een te groot optimisme op zijn plaats (zie ook Dousma en Horsten, 1980; Mills en Melican, 1988). Dè methode om de cesuur te bepalen bestaat niet behalve wanneer men een eenmaal vastgestelde cesuur via normhandhaving kan overbrengen. Zie daarvoor hoofdstuk 9.

7.1 Twee principes om de cesuur te bepalen

Bij het bepalen van de cesuur kan men twee principes hanteren: men kan of van de leerstof uitgaan of van de toetsresultaten van de kandidaten. Een cesuurbepaling met als uitgangspunt de leerstof wordt aangeduid met absolute cesuurbepaling en een cesuurbepaling met als uitgangspunt de resultaten van de groep kandidaten wordt aangeduid met relatieve cesuurbepaling.
Bij absolute cesuurbepaling wordt vóór afname van de toets vanuit de leerstof een beheersingsstandaard bepaald: een hoeveelheid kennis die vereist is om precies een voldoende te krijgen. De cesuur is dan vastgesteld los van de resultaten van de groep kandidaten die de toets heeft gemaakt. In principe is het mogelijk dat alle kandidaten slagen of dat alle kandidaten zakken.
Bij relatieve cesuurbepaling wordt ná afname van de toets vanuit de resultaten van de groep kandidaten de cesuur vastgesteld. Er is geen sprake van een vaste beheersingsstandaard. Waren de resultaten van de betreffende kandidaten anders geweest dan had dat een andere cesuur opgeleverd.

7.2 Cesuurbepaling bij de VWO-, HAVO- en LBO/MAVO-examens

Bij de centrale examens van het VWO, HAVO en LBO/MAVO wordt de cesuur bij de openvraag-toetsen en de gemengde-toetsen vooraf vastgesteld. Het betreft vaak toetsen waarop 100 punten te verdienen zijn en de cesuur is dan altijd 54/55, d.w.z. 54 punten correspondeert met het cijfer 5.4 en 55 punten met het cijfer

5.5. Na afname van de toetsen kan echter de cesuur bijgesteld worden. Het bijstellen bij deze toetsen betekent altijd dat de cesuur verlaagd wordt. Bij de toets die we als voorbeeld gebruiken is de cesuur verlaagd van 54/55 naar 48/49.

Bij de gesloten-toetsen van het VWO, HAVO en LBO/MAVO wordt de cesuur vooraf ook vastgesteld en deze kan in principe variëren tussen de grenzen 24/25 en 31/32 indien het een toets betreft met 50 gesloten vragen. Bij deze toetsen wordt eveneens pas na afname de definitieve cesuur vastgesteld en deze kan zowel omhoog als omlaag mits men binnen de grenzen 24/25 en 31/32 blijft.

De conclusie die men uit het voorgaande kan trekken is dat bij de VWO, HAVO en LBO/MAVO-toetsen de cesuur vooraf wel vastgelegd wordt maar dat men achteraf de cesuur kan bijstellen. Deze methode is dus een combinatie van absolute en relatieve cesuurbepaling en is wel eens aangeduid met 'thermostaat-methode'.

7.3 Samenvatting

Bij het samenstellen van een toets zou men idealiter vooraf moeten aangeven waar de cesuur komt te liggen. Dit vereist in de eerste plaats dat men weet hoe moeilijk de vragen zijn in de tweede plaats dat men een idee heeft hoe de vaardigheid van de groep kandidaten is die de toets moet maken. Zonder deze kennis lijkt de plaatsbepaling van de cesuur natte vingerwerk.
In hoofdstuk 9 besteden we aandacht aan methoden waarmee we een steviger fundament kunnen leggen onder de cesuurbepaling.

8 Cijfers geven

Is de cesuur vastgelegd dan is de volgende stap meestal dat men aan de toetsscores cijfers wil toekennen. Als we er vanuit gaan dat gelijke scoreverschillen moeten resulteren in gelijke cijferverschillen dan krijgen we een lineaire (rechtlijnige) omzetting van scores naar cijfers. Gaan we van andere principes uit dan zijn er net zoveel omzettingsformules als principes. Twee omzettingsmethoden, de lineaire en de lineaire-met-knik zullen besproken worden omdat deze regelmatig gebruikt worden.

8.1 Lineaire omzetting

Bij de lineaire omzettingsmethode worden twee punten van een rechte lijn vastgelegd. Aan de maximaal haalbare toetsscore kennen we het cijfer 10.0 toe en aan de cesuur het cijfer 5.45, indien we althans cijfers willen toekennen met één decimaal achter de komma.
De formule ziet er als volgt uit:

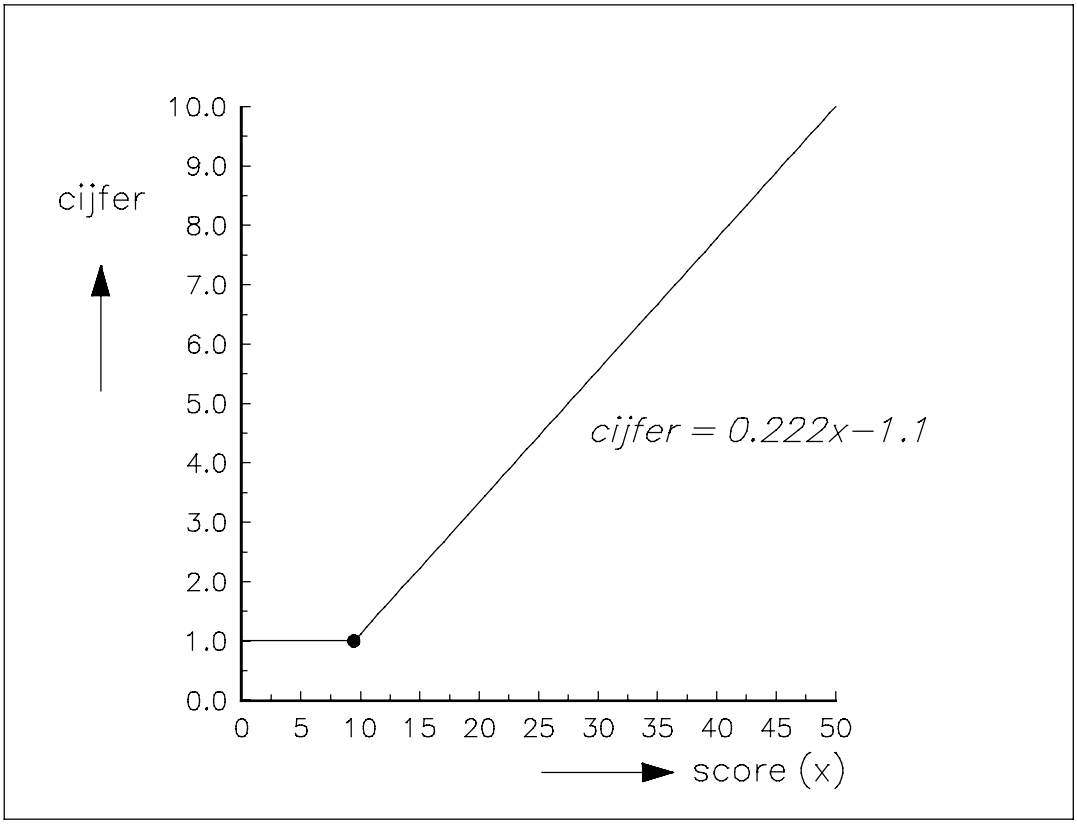
$$\text{cijfer} = 5.45 + \left(\frac{X - C}{H - C}\right) * 4.55 \qquad (18)$$

waarbij:
X = toetsscore
C = cesuur
H = maximaal te behalen toetsscore

Is bijvoorbeeld bij een toets waar maximaal 50 punten behaald kan worden, de cesuur 29/30 dan luidt de formule:

$$\text{cijfer} = 5.45 + \left(\frac{X - 29.5}{50 - 29.5}\right) * 4.55$$

Een score van 0 zou dan een cijfer van -1.1 opleveren maar omdat cijfers in de regel variëren van 1.0 t/m 10.0 zal een score van 0 met het cijfer 1.0 corresponderen. De score 29 correspondeert met 5.3 (de hoogste onvoldoende) en de score 30 met 5.6 (de laagste voldoende) en de score 50 tenslotte levert het cijfer 10.0 op. Verder komt een verschil van 1 scorepunt overeen met $\frac{4.55}{20.5} = .222$ cijferpunt. Figuur 4 (blz. 54) laat de lineaire omzetting van score naar cijfer zien.



Figuur 4. Lineaire omzetting van score naar cijfer (cesuur 29/30).

Indien we gehele cijfers willen toekennen dan kan men het cijfer 10 aan de maximale toetsscore toekennen en het cijfer 5.5 aan de cesuur.
De omzettingsformule heeft dan de volgende gedaante:

$$\text{cijfer} = 5.5 + \left(\frac{X - C}{H - C}\right) * 4.5 \qquad (19)$$

8.2 Lineaire omzetting met knik

Een andere omzettingsformule wordt sinds enkele jaren bij de meerkeuze-examens van het voortgezet onderwijs gebruikt. Deze omzettingsformule heeft namelijk een knik wanneer men de vooraf vastgestelde cesuur verhoogt. De omzettingformule wordt daarom ook wel de lineaire omzetting met knik of de 'hondepoot-formule' genoemd. Kenmerkend voor deze hondepoot-formule is dat onder de verhoogde cesuur een scoreverschil van 1 punt een ander cijferverschil oplevert dan boven de verhoogde cesuur.

Voorbeeld:
Bij de meerkeuze-examens in het voortgezet onderwijs wordt bij een toets van 50 items de cesuur voordat het examen is afgenomen op 24/25 gelegd. Wordt na afloop van het examen de cesuur op

29/30 gelegd, dan vindt de omzetting van score naar cijfer onder de nieuwe cesuur plaats met de formule:

$$C = \frac{9(Y+V)}{N} + 1.0 \qquad (20)$$

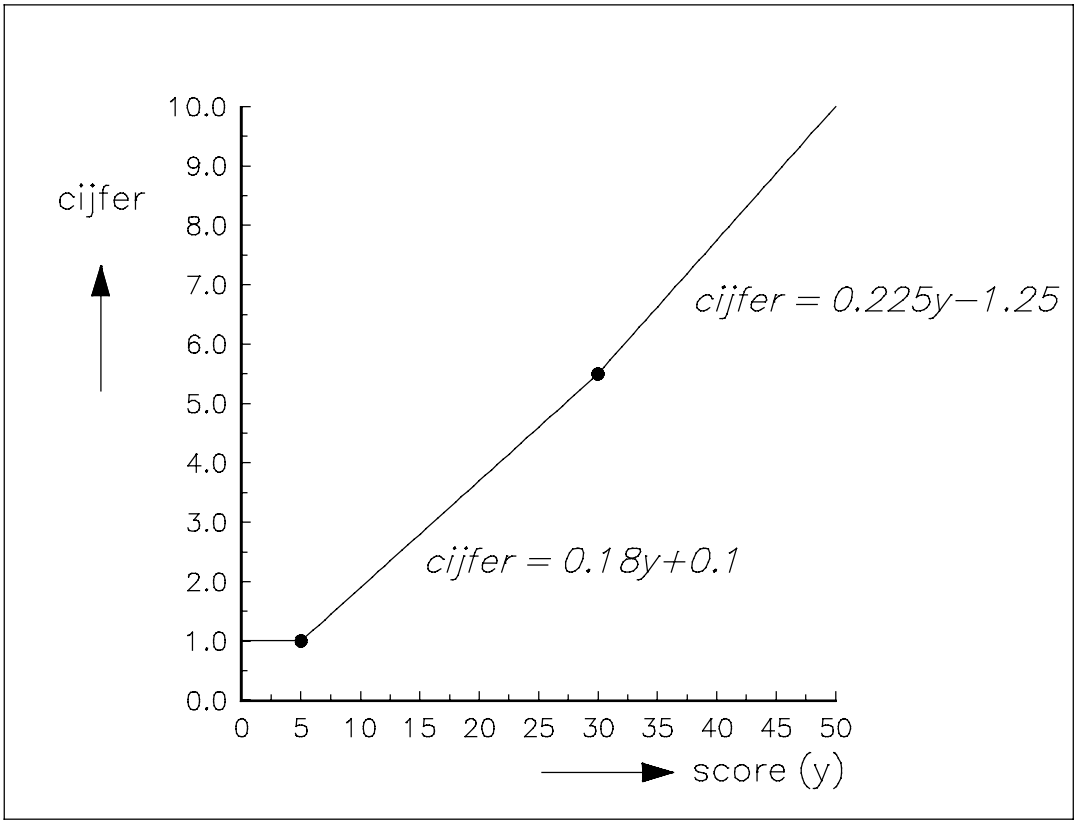
en boven de nieuwe cesuur vindt de omzetting plaats met de formule:

$$C = \frac{9Y+N+20V}{N+2V} \qquad (21)$$

waarbij:
C = cijfer
Y = aantal juist beantwoorde items
N = totale aantal items in de toets
V = cesuurverschuiving ($V = ces_o - ces_n$, waarbij ces_o de oude cesuur is en ces_n de nieuwe)

Figuur 5 laat de lineair-met-knik omzetting zien. Bij de score 30 ziet men de knik.

Onder de cesuur komt nu een verschil van 1 scorepunt overeen met een verschil van $\frac{9}{N} = \frac{9}{50} = 0.18$ cijferpunt en boven de cesuur komt een verschil van 1 scorepunt overeen met $\frac{9}{N+2V} = 0.225$



Figuur 5. Lineaire omzetting met knik (cesuur 29/30).

cijferpunt.

We zullen het bij deze twee methoden laten wat de toelichting betreft. Dit betekent niet dat we andere methoden verwerpen maar een toelichting daarvan is weinig zinvol omdat deze methoden niet, althans niet op het Cito, in praktijk worden gebracht.

Opgemerkt dient nog te worden dat in deze paragraaf de lineaire omzetting met knik niet in extenso is behandeld, daarvoor wordt verwezen naar Goldebeld (1989).

8.3 Samenvatting

Ook bij cijfergeving zijn diverse methoden voorhanden. Uitgaande van het principe dat gelijke scoreverschillen moeten leiden tot gelijke cijferverschillen is de lineaire omzettingsmethode de meest voor de hand liggende. Wil men daarvan afwijken dan is de lineaire omzettingsmethode met knik één van de alternatieven.

9 Normhandhaving

Uit het oogpunt van billijkheid is het gewenst dat toetsen uit opeenvolgende jaren/tijdvakken even moeilijk zijn. Dat wil zeggen: het moet voor een kandidaat niet uitmaken uit welk jaar hij de toets maakt. De kandidaat zou altijd hetzelfde cijfer moeten krijgen voor dezelfde prestatie.

Het is niet altijd mogelijk gelijke moeilijkheid van opeenvolgende toetsen te garanderen. In dit hoofdstuk wordt aandacht geschonken aan normhandhavingsmethoden waarmee we gelijke moeilijkheid van toetsen kunnen realiseren. We zullen vier normhandhavingsmethoden bespreken, namelijk normhandhaving met gepreteste items, normhandhaving met een ankertoets, normhandhaving via pre-equivaleren en normhandhaving via de regressiemethode. In paragraaf 9.5 zullen we enkele rekentechnieken noemen waarmee we de scores van toetsen die in moeilijkheid verschillen op eenzelfde schaal kunnen brengen (ook wel equivaleren genoemd).

9.1 Normhandhaving met gepreteste vragen

Indien men over vragen beschikt waarvan men via pretesten de p- en r_{it}-waarden heeft bepaald, dan is dat verreweg de beste manier om parallelliteit (gelijkwaardigheid) van toetsen te bewerkstelligen. We gaan dan uit van een grote verzameling gekalibreerde vragen, dat wil zeggen vragen waarvan de p- en r_{it}-waarden onderling vergelijkbaar zijn gemaakt. Uit deze verzameling vragen kunnen we vervolgens toetsen samenstellen die even moeilijk en even betrouwbaar zijn.

9.2 Normhandhaving met een ankertoets

Bij deze methode wordt naast de eigenlijke toets een ankertoets afgenomen. De scores behaald op de ankertoets kan men meetellen, men spreekt dan van een intern anker of men kan de scores buiten beschouwing laten en men spreekt dan van een extern anker. De ankertoets bestaat bij een toets van 50 items uit 10 à 20 items die ongeveer even moeilijk zijn als de andere items en de bedoeling is dat deze ankertoets telkens weer bij een volgende toets wordt afgenomen. Terwijl de eigenlijke toets openbaar is wordt de ankertoets onder geheimhouding afgenomen.

Qua proefopzet zijn er een aantal varianten denkbaar. Indien bijvoorbeeld de ankertoets te groot is om in zijn geheel af te

nemen dan kan men de ankertoets in stukken knippen en elk stukje aan een andere groep kandidaten voorleggen.
Een andere variant is om het anker niet aan alle kandidaten voor te leggen maar alleen aan een representatieve steekproef van kandidaten opdat niet alle kandidaten extra belast worden. Een voorbeeld van deze variant wordt bij de Eindtoets Basisonderwijs toegepast. Bij de Eindtoets die uit 180 items bestaat, wordt een anker afgenomen van 45 items. Dit anker wordt een paar dagen voor de werkelijke afname van de Eindtoets onder geheimhouding afgenomen bij een steekproef van meer dan 1000 leerlingen. Hetzelfde anker wordt een paar jaar achter elkaar gebruikt om vervolgens 'ververst' te worden.

9.3 Normhandhaving via pre-equivaleren

Holland en Wightmann (1982) hebben normhandhavingsmethoden ontwikkeld die zij aanduiden als pre-equivaleren. Kenmerkend voor deze methode is dat de toets vóór de eigenlijke afnamedatum op een populatie wordt afgenomen die vergelijkbaar is met de examenpopulatie. Naast de te equivaleren toets wordt een standaardtoets afgenomen. Deze standaardtoets waarvan de cesuur bekend is, is het anker en wordt in de opeenvolgende jaren steeds weer afgenomen samen met de in een bepaald jaar af te nemen toets. In Sanders en Goldebeld (1986) wordt een voorbeeld gepresenteerd van normhandhaving via pre-equivaleren. In het voorbeeld worden de examens tekstbegrip Duits MAVO-C van het 1e en het en 2e tijdvak 1984 gepreëquivaleerd met een standaardexamen (het examen MAVO-3 Duits van het 1e tijdvak 1982).

9.4 Normhandhaving via de regressiemethode

De regressiemethode kan toegepast worden op de meerkeuze-examens van het 2e tijdvak bij het VWO, HAVO en LBO/MAVO (Goldebeld, 1982 en 1984). Bij de regressiemethode wordt het rechtlijnige verband bepaald tussen de scores van het 1e en 2e tijdvak van die kandidaten die in het eerste tijdvak een onvoldoende hebben en in het 2e tijdvak weer examen doen. Vervolgens wordt het gemiddelde van het 2e tijdvak geschat door het gemiddelde van alle kandidaten van het 1e tijdvak in de regressievergelijking te substitueren. Wijken de gemiddelden van het 1e en 2e tijdvak niet van elkaar af dan is de cesuur van het 2e tijdvak gelijk aan de

cesuur van het 1e tijdvak. Anders moet de cesuur van het 2e tijdvak bijgesteld worden.

9.5 Equivaleringstechnieken

Equivaleren is een procedure om scores van toetsen die in moeilijkheid verschillen op dezelfde schaal te brengen zodanig dat aangegeven wordt welke scores van de verschillende toetsen gelijkwaardig (equivalent) zijn. We kunnen equivaleren via klassieke methoden of via methoden die gebruik maken van de item-response-theorie.
Kolen (1988) geeft een heldere uiteenzetting van de klassieke equivaleringstechnieken zoals lineair equivaleren en equipercentiel equivaleren. Crocker en Algina (1986) behandelen naast het klassieke equivaleren ook het equivaleren met item-response-technieken. Goldebeld en De Jong (1988) tenslotte hebben vier equivaleringsmethoden met elkaar vergeleken waarvan er drie klassiek waren en een gebaseerd op de item-response-theorie. De laatste bleek hierbij het meest nauwkeurig te zijn.

10 Nauwkeurigheid van schattingen zoals p-waarde, r_{it}-waarde en KR-20

De p-waarde, het percentage (on)voldoendes, de p'-waarde, de gemiddelde score, de r_{it}-waarde, de KR-20 en α zijn allemaal voorbeelden van grootheden die vaak gebaseerd zijn op steekproeven. Indien dit het geval is dan zijn deze grootheden behept met steekproeffouten. In de volgende hoofdstukken wordt op deze steekproeffouten, standaardfouten genoemd, ingegaan.

10.1 Standaardfout van een p-waarde en van een percentage onvoldoendes

De standaardfout (s_p) van een p-waarde en van een percentage onvoldoendes wordt als volgt berekend:

$$s_p = \sqrt{\frac{p(100-p)}{n}}$$

waarbij:
s_p = standaardfout p-waarde (uitgedrukt in %)
p = p-waarde (uitgedrukt in %)
n = aantal kandidaten (steekproefgrootte)

Bovenstaande formule is een benadering en mag alleen gebruikt worden indien:

$$n > (9 * \frac{100-p}{p}) \text{ bij } p < 50 \qquad \text{of}$$

$$n > (9 * \frac{p}{100-p}) \text{ bij } p > 50$$

In tabel 4 (blz.61) die gebaseerd is op exacte berekeningen kan men voor diverse waarden van p en n aflezen tussen welke grenzen de werkelijke p-waarde ligt (zie ook H. de Jonge, 1963).

Tabel 4. 95%-betrouwbaarheidsintervallen voor percentages

steekproef-percentage p	aantal kandidaten in de steekproef (n)											
	50		100		200		400		600		1000	
0	0	7	0	4	0	2	0	1	0	1	0	0
5	1	17	2	11	2	9	3	8	3	7	4	7
10	3	22	5	18	6	15	7	13	8	13	8	12
15	7	29	9	24	10	21	12	19	12	18	13	17
20	10	34	13	29	15	26	16	24	17	23	18	23
25	15	40	17	35	19	32	21	30	22	29	22	28
30	18	45	21	40	24	37	26	35	26	34	27	33
35	23	51	26	45	28	42	30	40	31	39	32	38
40	26	55	30	50	33	47	35	45	36	44	37	43
45	32	61	35	55	38	52	40	50	41	49	42	48
50	35	65	40	60	43	57	45	55	46	54	47	53
55	39	68	45	65	48	62	50	60	51	59	52	58
60	45	74	50	70	53	67	55	65	56	64	57	63
65	49	77	55	74	58	72	60	70	61	69	62	68
70	55	82	60	79	63	76	65	74	66	74	67	73
75	60	85	65	83	68	81	70	79	71	78	72	78
80	66	90	71	87	74	85	76	84	77	83	77	82
85	71	93	76	91	79	90	81	88	82	88	83	87
90	78	97	82	95	85	94	87	93	87	92	88	92
95	83	99	89	98	91	98	92	97	93	97	93	96
100	93	100	96	100	98	100	99	100	99	100	100	100

Indien bijvoorbeeld een item bij een steekproef van 200 kandidaten een p-waarde heeft van 40 dan ligt de werkelijke p-waarde met betrouwbaarheid van 95% tussen de grenzen 33 en 47. Ander voorbeeld: een percentage onvoldoendes van 25% bij een steekproef van 1000 kandidaten betekent dat het werkelijke percentage onvoldoendes in de totale populatie kan variëren van 22 tot 28%.

10.2 Standaardfout van een gemiddelde en een p'-waarde

De standaardfout ($s_{\overline{x}}$) van een gemiddelde toetsscore ($\overline{X}$) is gelijk aan:

$$s_{\overline{x}} = \frac{S_x}{\sqrt{n}}$$

waarbij:

$s_{\bar{X}}$ = standaardfout van $\bar{X}$

s_X = standaarddeviatie van toetsscores

n = aantal kandidaten

Het gemiddelde van de toetsscores gebaseerd op de steekproef van 2595 kandidaten is 60.51 (bijlage V). De standaardfout van dit gemiddelde bedraagt $\frac{12.83}{\sqrt{2595}} = 0.25$.

Dit betekent dat het werkelijke gemiddelde (het gemiddelde van alle examenkandidaten) tussen 60.02 en 61.00 ligt met een betrouwbaarheid van 95%. (60.02 = 60.51 - 1.96 * 0.25 en 61.00 = 60.51 + 1.96 * 0.25)

De standaardfout van een p'-waarde is gelijk aan

$$s_{p'} = \frac{SD}{\sqrt{n} * M} * 100$$

waarbij:

$s_{p'}$ = standaardfout p'-waarde (in %)

SD = standaarddeviatie van itemscores

n = aantal kandidaten

M = maximale itemscore (MAX.SCORE)

Vraag 28 heeft bijvoorbeeld een p'-waarde van 60.4 (bijlage VII). De standaardfout van deze p'-waarde is gelijk aan $\frac{1.78}{\sqrt{2595} * 5} * 100 = 0.70$.

Dit betekent dat de werkelijke p'-waarde tussen 59.0 en 61.8 ligt.

10.3 Standaardfout van een r_{it}-waarde

De berekening van de standaardfout van een r_{it}-waarde is nogal gecompliceerd. In Iker en Perry (1960) staan benaderingsformules en tabellen voor de standaardfout. Tabel 5 is gebaseerd op Iker en Perry. In deze tabel zijn voor diverse waarden van de r_{it} en n (steekproefgrootte) de grenzen aangegeven waartussen de werkelijke waarde van de r_{it} ligt met een betrouwbaarheid van 95%.

Tabel 5. 95% betrouwbaarheidsintervallen voor r_{it}-waarden

r_{it}-waarde (steekproef)	aantal kandidaten in de steekproef (n)			
	100	200	500	1000

0.00	-0.20	0.20	-0.14	0.14	-0.08	0.08	-0.06	0.06
0.10	-0.10	0.30	-0.04	0.24	0.02	0.18	0.04	0.16
0.20	0.00	0.40	0.06	0.34	0.12	0.28	0.14	0.26
0.30	0.12	0.48	0.18	0.42	0.22	0.38	0.24	0.36
0.40	0.24	0.56	0.28	0.52	0.32	0.48	0.34	0.46
0.50	0.36	0.64	0.40	0.60	0.44	0.56	0.46	0.54
0.60	0.48	0.72	0.51	0.69	0.54	0.66	0.56	0.64

Indien bijvoorbeeld bij een toets- en itemanalyse gebaseerd op 1000 kandidaten de r_{it}-waarde van een item 0.20 is, dan ligt de werkelijke waarde van de r_{it} tussen 0.14 en 0.26.
Opgemerkt dient te worden dat tabel 5 van toepassing is op p-waarden die globaal tussen de 20 en 80 liggen. (Voor andere p-waarden zijn de intervallen ruimer en kunnen bij OPD berekend worden via een door T. Heuvelmans ontwikkeld computerprogramma geheten 'Iker'.)

10.4 Standaardfout van de KR-20 en α

Feldt (1965) heeft een steekproefverdeling voor de KR-20 en **α** afgeleid. Hiervan uitgaande is tabel 6 geconstrueerd. In deze tabel zijn bij diverse steekproefwaarden van de KR-20 en n (steekproefgrootte) onder- en bovengrenzen vermeld waarbinnen de werkelijke waarde van de KR-20 met een betrouwbaarheid van 95% ligt.

Tabel 6. 95% betrouwbaarheidsintervallen voor KR-20 en **α**

KR-20/**α** (steekproef)	aantal kandidaten in de steekproef (n)							
	100		200		500		1000	
0.10	-0.17	0.33	-0.09	0.27	-0.02	0.21	0.02	0.18
0.20	-0.04	0.41	0.03	0.35	0.10	0.30	0.13	0.27
0.30	0.09	0.48	0.25	0.43	0.21	0.38	0.24	0.30
0.40	0.22	0.55	0.27	0.51	0.32	0.47	0.35	0.45
0.50	0.35	0.63	0.40	0.59	0.44	0.56	0.45	0.54
0.60	0.48	0.70	0.52	0.67	0.55	0.65	0.56	0.63
0.70	0.61	0.78	0.64	0.76	0.66	0.74	0.67	0.73
0.80	0.74	0.85	0.76	0.84	0.77	0.82	0.78	0.82
0.90	0.87	0.93	0.88	0.92	0.89	0.91	0.89	0.91

Indien bijvoorbeeld bij een toets- en itemanalyse van 500 kandidaten de betrouwbaarheid 0.70 is dan mogen we met 95% zekerheid zeggen dat de werkelijke betrouwbaarheid, dus de betrouwbaarheid gebaseerd op oneindig veel kandidaten, tussen de waarden 0.74 en 0.66 ligt.
(Opgemerkt dient te worden dat tabel 6 alleen gebruikt mag worden bij 10 of meer vragen in een (sub)toets. Zijn er minder vragen dan is er bij OPD een door T. Heuvelmans ontwikkeld computerprogramma beschikbaar om de grenzen te berekenen. Dit programma heet 'Alpharel'.)

11 Antwoordbladen

Op het Cito zijn vijf standaard-antwoordbladen (optisch leesbare formulieren) beschikbaar. In onderstaande tabel is schematisch weergegeven wat voor mogelijkheden de antwoordbladen bieden.

Tabel 7. Overzicht van de mogelijkheden van de 5 standaard-antwoordbladen

Formuliernr.	C5-210-1	C250-1	C250-2	C250-3	C250-4
Zie bijlage	XIX	XX	XXI	XXII	XXIII
Plano of ketting	plano	plano	ketting	plano	ketting
Aantal vragen	80	80	80	60	60
Aantal alternatieven/max.score per vraag	4 (A-D)	6 (A-F)	6 (A-F)	10	10

De plano-formulieren zijn 'losse' formulieren waarop alle relevante gegevens door de kandidaten moeten worden ingevuld. Op de kettingformulieren kunnen de identificatiegegevens door de afdeling automatisering worden voorbedrukt (in tekst en in optisch leesbare codes).
De formulieren met de nummers C5-210-1, C250-1 en C250-2 zijn antwoordbladen voor gesloten vragen en de formulieren met de nummers C250-3 en C250-4 zijn antwoordbladen voor open vragen. In de bijlagen XIX t/m XXIV zijn de 5 standaardantwoordbladen afgedrukt.
Een standaardantwoordblad voor toetsen bestaande uit gesloten en open vragen (de zgn. gemengde-toetsen) is nog niet ontwikkeld. Specifiek voor de gemengde-toetsen van de eindexamens LBO/MAVO is wel zo'n antwoordblad ontwikkeld en deze is in bijlage XXIV afgedrukt.

12 Bibliografie

Berge ten, J.F.M. (1984). Een definitie van betrouwbaarheid in termen van ruwe scores. *Kwantitatieve Methoden, 16*, 63-72.

Berge ten, J.M.F., & Zegers, F.E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43,* 575-579.

Brink van den, W. (1979). Het optimale aantal alternatieven per item. *Tijdschrift voor Onderwijsresearch, 4*, 151-158.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Daal van, W.J. (1992). *Handleiding standaardpakket TIA*. Arnhem: Cito.

Dousma, T., & Horsten, A. (1980). Hoger Onderwijs reeks. *Tentamineren*. Groningen: Wolters-Noordhoff.

Drenth, P.J.D., & Sijtsma K. (1990). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten/Antwerpen: Bohn Stafleu Van Loghum.

Ebel, L., & Frisbie, D.A. (1986). *Essentials of educational measurement*. Prentice Hall, Englewood Cliffs: New York.

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*, 357-370.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and practice, 7,* 53-63.

Goldebeld, P. (1982) *Cesuurbepaling 2e tijdvak.* Interne Documentatie nr. 78. Arnhem: Cito.

Goldebeld, P. (1983). *Boordelaarsovereenstemming bij briefopdrachten Duits in het LBO*. Interne Documentatie nr. 106. Arnhem: Cito.

Goldebeld, P. (1984). *Evaluatie van de regressiemethode*. Interne Documentatie nr. 142, Arnhem: Cito.

Goldebeld, P. (1989). *Omzetting van score naar cijfer vanaf 1990 bij de LBO/MAVO/HAVO/VWO-examens.* Notitie nr. 89.850.188. Arnhem: Cito.

Goldebeld, P., & de Jong, J.H.A.L. (1988). *Een vergelijking van vier equivaleringsmethoden om de moeilijkheid van een toets te schatten.* Interne Documentatie nr. 286. Arnhem: Cito.

Gruijter de, D.N.M. (1982). *Tentamineren en beslissen.* 's Gravenhage: Stichting voor Onderzoek van het Onderwijs SVO, SVO reeks 63.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Harvill, L.M. (1991). Standard error of measurement. *Educational Measurement: Issues and practice, 10,* 33-41.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28,* 211-218.

Holland, P.W., & Wightmann, L.E. (1982). *Section pre-equating: a preliminary investigation*. In: P.W. Holland & D.B. Rubin (Eds.). *Test Equating*. New York: Academic Press, 271-297.

Iker, H.P., & Perry, N.C.A. (1960). A further note concerning the reliability of the point-biserial correlation. *Educational and Psychological Measurement, 20*, 505-507.

Jonge de, H. (1963). *Inleiding tot de medische statistiek deel I*. Groningen: Wolters-Noordhoff.

Kolen, M.J. (1988). Traditional equating methodology. *Educational Measurement: Issues and practice, 7,* 29-36.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Melse, L., & Mets, J. (1984). *Items met 3 of 4 alternatieven in tekstbegrip-examens?* Interne Documentatie nr. 128. Arnhem: Cito.

Mills, C.N., & Melican, G.J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education, 1*, 261-275.

Nuttall, D.L., & Willmott, A.S. (1972). *British examinations: Techniques of analysis.* National Foundation for Educational Research in England and Wales.

Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen (1988). Nederlands Instituut van Psychologen.

Sanders, P.F., & Goldebeld, P. (1985). Het pre-equivaleren van examens. In: W.J. van der Linden (Red.) *Moderne methoden voor toetsconstructie en -gebruik*. Lisse: Swets & Zeitlinger.

Thorndike, R.L.T.(Ed.) (1971). *Educational measurement.* Washington, D.C.: American Council on Education.

Thorndike, R.L.T. (1982). *Applied psychometrics*. Boston: Houghton Mifflin Company.

Toetstechnische Begrippenlijst (1988). Arnhem: Cito.

Traub, R.E., & Rowley, G.L. (1991). Understanding reliability. *Educational Measurement: Issues and practice, 10,* 37-45.