# 3

# Klassieke testtheorie en generaliseerbaarheidstheorie

De klassieke testtheorie beschrijft het verschijnsel meetfout en procedures om de grootte van meetfouten te bepalen. Het uitgangspunt van de klassieke testtheorie is een meting $x_{vt}$ die verkregen is door een meetinstrument $t$ voor te leggen aan een persoon $v$. Zoals is uiteen- gezet in het vorige hoofdstuk, wordt een meting altijd gecodeerd als een getal. Zo'n gecodeerde meting noemt men een score. De klassieke testtheorie houdt zich niet bezig met de aard, het schaalniveau en de interpretatie van een score. Zij houdt zich met slechts een enkel probleem bezig, en wel met de meetfout waarmee een score $x_{vt}$ behept is. De meetfout wordt geacht op te treden doordat men bij het meten niet alle factoren in de hand heeft die op een meting van invloed zijn. Zulke factoren verstoren de meetprocedure en zorgen er voor dat men niet de meting krijgt die men graag had willen hebben maar een daar enigszins van afwijkende score. Verstorende factoren kunnen zijn gelegen in de te meten persoon, in het meetinstrument, en in de meetsituatie. Een voorbeeld van de eerste soort is de bloeddruk: deze vertoont in de loop van de dag zulke grote fluctuaties dat een enkele meting eigenlijk onvoldoende is. Een voorbeeld van de tweede soort verstorende factoren is de thermometer. Dat instrument wisselt warmte uit met het te meten voorwerp, waardoor de thermometer niet de exacte temperatuur van het voorwerp aangeeft. Een voorbeeld van een verstoring in de meetsituatie is het eindexamen dat wordt afgenomen in een schoolgebouw waarnaast een heistelling palen de grond in boort.

De belangrijkste parameters uit de klassieke testtheorie zijn correlaties en standaard-afwijkingen. Het gebruik van dergelijke parameters brengt met zich mee dat alle uitspraken van de klassieke testtheorie over personen en over meetinstrumenten gerelateerd zijn aan een bepaalde populatie. Zo kan men eigenschappen van een meetinstrument die bepaald zijn in een populatie, niet zonder meer voor geldend houden in een andere populatie. Voor een aantal meetproblemen schiet de klassieke testtheorie dan ook tekort. De wens, te kunnen beschikken over parameters van

personen en meetinstrumenten die niet aan een populatie gebonden zijn, heeft geleid tot de itemresponstheorie. Deze theorie wordt behandeld in hoofdstuk 4.

De klassieke testtheorie wordt eerst, in de paragrafen 3.1 tot en met 3.6, in abstracte termen beschreven. In de paragrafen 3.7 tot en met 3.10 worden diverse grootheden concreet geïllustreerd aan de hand van een voorbeeld. Daarbij worden ook grootheden behandeld die optreden bij het construeren van toetsen. De toets uit het voorbeeld is klein gehouden om het de lezer mogelijk te maken het rekenwerk te volgen. Een uitbreiding van de klassieke testtheorie, de generaliseerbaarheidstheorie, wordt in de paragrafen 3.11 tot en met 3.14 besproken.

## 3.1 Ware score

De waargenomen score is door de verstorende factoren niet altijd de meting die we zouden willen hebben. De klassieke testtheorie veronderstelt nu dat het effect van de verstorende factoren beschouwd kan worden als een aselecte trekking uit een kansverdeling. In feite is dit de enige veronderstelling die de klassieke testtheorie kent. De afleiding die nu volgt is gebaseerd op Novick (1966). Uit de zojuist genoemde veronderstelling kan men de gehele klassieke testtheorie opbouwen. Als de bij de meting $x_{vt}$ optredende meetfout wordt aangeduid met $\varepsilon_{vt}$, veronderstelt de klassieke testtheorie dat deze meetfout een realisatie is van een toevalsvariabele $E_{vt}$. Deze toevalsvariabele draagt twee subscripten om aan te geven dat zij varieert binnen de combinatie van de vaste persoon $v$ en het vaste meetinstrument $t$. Beschouw nu de voor de meetfout gecorrigeerde meting $\tau_{vt} = x_{vt} - \varepsilon_{vt}$. Men kan dan ook schrijven: $x_{vt} = \tau_{vt} + \varepsilon_{vt}$. Deze uitdrukking schrijft de score $x_{vt}$ als een ontbinding, een decompositie, in twee termen. De eerste term, $\tau_{vt}$, zou men kunnen opvatten als de meting die men had willen verkrijgen. Maar de gegeven ontbinding is niet uniek. Men kan namelijk bij de term $\tau_{vt}$ een willekeurige constante $c$ optellen en deze constante van de term $\varepsilon_{vt}$ aftrekken zonder dat het resultaat verandert: $x_{vt} = \tau_{vt} + \varepsilon_{vt} = (\tau_{vt} + c) + (\varepsilon_{vt} - c)$. In feite is dit een geval van een vergelijking met twee onbekenden. Om met de gegeven decompositie uit de voeten te kunnen, moet men normeren. Daaronder verstaat men het kiezen en vastleggen van een waarde voor de constante $c$. In de klassieke testtheorie heeft men voor de volgende normering gekozen. Aangezien $E_{vt}$ een toevalsvariabele is met realisaties $\varepsilon_{vt}$, en $\tau_{vt}$ een vaste waarde heeft, is $x_{vt}$ een realisatie van een toevalsvariabele $X_{vt}$. Voor de constante $c$ is in de klassieke testtheorie de verwachte waarde van de toevalsvariabele $E_{vt}$ gekozen: $c = \mathscr{E}(E_{vt})$. De verwachte waarde van een toevalsvariabele kan men in dit boek opvatten als het

gemiddelde van een hele grote steekproef van trekkingen uit de verdeling van die variabele. De verwachte waarde van een constante is gelijk aan die constante. Met de gekozen normering kan men nu de toevalsvariabele $X_{vt}$ schrijven als: $X_{vt} = \{\tau_{vt} + \mathscr{E}(E_{vt})\} + \{E_{vt} - \mathscr{E}(E_{vt})\}$. Daaruit volgt onmiddellijk dat $\mathscr{E}(X_{vt}) = \tau_{vt} + \mathscr{E}(E_{vt})$. Ook deze decompositie moet genormeerd worden. In de klassieke testtheorie stelt men daartoe $\mathscr{E}(E_{vt})$ gelijk aan 0. Het resultaat is de volgende belangrijke uitdrukking:

$$\mathscr{E}(X_{vt}) = \tau_{vt}. \tag{3.1}$$

Het rechterlid van (3.1) heet in de klassieke testtheorie de ware score van persoon *v* op meet- instrument *t.* Men dient te beseffen dat de door (3.1) gedefinieerde ware score een wis- kundige constructie is en niet noodzakelijkerwijze gelijk is aan de score die verkregen zou zijn als er geen verstorende factoren aanwezig waren. Het kan bijvoorbeeld goed zijn dat de toevalsvariabele $X_{vt}$ alleen maar gehele waarden kan aannemen; dat sluit echter niet uit dat de verwachte waarde van die variabele, de ware score, een gebroken getal is.

### 3.2 De centrale formule van de klassieke testtheorie

De ware score is, omdat hij is gedefinieerd als een verwachte waarde, een maat voor de centrale tendentie van de scores: hij geeft aan om welke waarde de verkregen metingen variëren. Het is van groot belang, te weten in welke mate de metingen rondom de ware score variëren. Bekende maten voor de variatie van een toevalsvariabele zijn de variantie en de standaardafwijking van die variabele. De variantie van een toevalsvariabele is gelijk aan de verwachte waarde van het kwadraat van het verschil tussen een score en de daarbij behorende ware score. Voor de toevalsvariabele $X_{vt}$ schrijft men de variantie als volgt: $\sigma^2_{X_{vt}} = \mathscr{E}\{(X_{vt} - \tau_{vt})^2\}$. Omdat geldt dat $X_{vt} - \tau_{vt}$ gelijk is aan $E_{vt}$ en omdat $\mathscr{E}(E_{vt})$ gelijk is aan 0, kan men de zojuist geschreven variantie ook schrijven als: $\sigma^2_{X_{vt}} = \mathscr{E}\{(E_{vt})^2\}$. De laatste uitdrukking kan men natuurlijk ook schrijven als: $\sigma^2_{E_{vt}}$.

Merk op dat de in deze paragraaf genoemde varianties alle betrekking hebben op de variatie van toevalsvariabelen die zijn gedefinieerd voor een vaste persoon *v* en een vast meetinstrument *t*. Om de varianties te kunnen schatten, zou men moeten beschikken over herhaalde metingen van *v* met *t*, verkregen onder identieke omstandigheden. Door de eerder genoemde verstorende factoren is het echter niet mogelijk, herhaalde metingen te verkrijgen onder identieke omstandigheden. In plaats

van herhaalde metingen te gebruiken, gaat de klassieke testtheorie er toe over meer personen tegelijk te beschouwen. Het is duidelijk dat nu kenmerken van een populatie $\mathcal{P}$ van personen een rol gaan spelen.

Beschouw een willekeurig uit de populatie $\mathcal{P}$ getrokken persoon. Om aan te geven dat de persoon willekeurig is getrokken, duiden we die persoon aan met een $\star$. Zodra we de persoon $\star$ hebben getrokken, geldt alles wat hierboven gezegd is. Men kan denken aan een tweestapsprocedure: eerst trekt men willekeurig een persoon $\star$ uit de populatie $\mathcal{P}$, en dan trekt men een meetfout $\varepsilon_{\star t}$ uit de verdeling van de toevalsvariabele $E_{\star t}$. Bij de persoon $\star$ behoort een ware score $\tau_{\star t}$. Men kan nu ook zeggen dat er drie nieuwe toevalsvariabelen zijn gemaakt: $T_{\star t}$, $E_{\star t}$ en $X_{\star t}$. De laatste twee variabelen variëren zowel over personen als binnen de aselect gekozen persoon; de eerste varieert alleen over personen. De betrekking tussen de drie toevalsvariabelen kan men schrijven als: $X_{\star t} = T_{\star t} + E_{\star t}$. Omdat we in het vervolg steeds een enkel meetinstrument en een enkele populatie beschouwen, laten we waar dat mogelijk is de subscripten weg. De laatst geschreven betrekking kan men dan schrijven als:

$$X = T + E \ . \tag{3.2}$$

Formule (3.2) is de centrale formule van de klassieke testtheorie. Men kan er, jammer genoeg, niet aan zien dat de toevalsvariabele $T$ alleen over personen varieert maar niet binnen een persoon, en dat de toevalsvariabelen $X$ en $E$ zowel tussen de personen als binnen elke persoon variëren. In het bovenstaande is daarom uiteengezet hoe deze formule tot stand komt.

### 3.3 Betrouwbaarheid

Uit (3.2) kan men enige interessante betrekkingen afleiden. In de eerste plaats geldt dat de verwachte waarde van de toevalsvariabele $E$ over de populatie $\mathcal{P}$ gelijk is aan 0: $\mathcal{E}_{\mathcal{P}}\mathcal{E}(E) = \mathcal{E}_{\mathcal{P}}(0) = 0$. Er zijn twee verwachtingen genomen: in de eerste plaats de verwachting over de meetfouten binnen een persoon, en in de tweede plaats de verwachting over personen van de verwachte meetfout. Dit komt overeen met het feit dat $E$ zowel binnen een persoon als over personen varieert.

In de tweede plaats kan men afleiden dat de correlatie tussen de variabelen $T$ en $E$ gelijk is aan 0. Immers, voor elke persoon $v$ in $\mathcal{P}$ geldt dat $\mathcal{E}(E_{vt}) = 0$. Dit geldt dan ook voor een willekeurig uit de populatie $\mathcal{P}$ getrokken persoon $\star$. A fortiori geldt dit voor elke persoon $\star$ uit $\mathcal{P}$ die een ware score gelijk aan $\tau_{\star t}$ heeft: $\mathcal{E}(E_{\star t}|\tau_{\star t}) = 0$. Dit geldt natuurlijk voor elke waarde van $\tau_{\star t}$. De uitdrukking $\mathcal{E}(E_{\star t}|\tau_{\star t})$ heet: de regressie

van $E$ op $T$. Aangezien de regressie van $E$ op $T$ gelijk is aan 0, is ook de correlatie tussen $E$ en $T$ gelijk aan 0.

In de derde plaats kan men uit de decompositie van $X$ die gegeven is in (3.2), de volgende decompositie afleiden van de variantie $\sigma_X^2$ van de variabele $X$:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{3.3}$$

De drie varianties zijn de varianties van respectievelijk de waargenomen toetsscores, de ware toetsscores en de meetfouten. Men noemt de drie varianties doorgaans: geobserveerde variantie, ware variantie en foutenvariantie.

Een van de voornaamste grootheden in de klassieke testtheorie is de betrouwbaarheid. Deze grootheid, die wordt voorgesteld door het symbool $\rho_{XT}^2$, is als volgt gedefinieerd:

$$\rho_{XT}^2 = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \{\sigma_T^2 + \sigma_E^2\}. \tag{3.4}$$

Zolang de geobserveerde variantie groter is dan 0, neemt de betrouwbaarheid waarden aan tussen 0 en 1. De betrouwbaarheid is gelijk aan 0 als er geen ware variantie is: men meet alleen maar meetfouten met het meetinstrument. De betrouwbaarheid is gelijk aan 1 als er geen sprake is van meetfouten: $\sigma_E^2 = 0$, wat overeenkomt met $\sigma_X^2 = \sigma_T^2$. Elke geobserveerde score van een persoon is dan gelijk aan de ware score van die persoon. In het uitzonderlijke geval dat $\sigma_X^2$ gelijk is aan 0, is de betrouwbaarheid niet gedefinieerd.

Waarom de betrouwbaarheid wordt aangeduid met het symbool $\rho_{XT}^2$, wordt duidelijk als men de correlatie beschouwt tussen de geobserveerde scores $X$ en de ware scores $T$. De teller van deze correlatie is gelijk aan de covariantie tussen $X$ en $T$:

$$Cov(X,T) = \mathcal{E}[\{X - \mathcal{E}(X)\} \times \{T - \mathcal{E}(T)\}] =$$

$$\mathcal{E}([\{T - \mathcal{E}(T)\} + \{E - \mathcal{E}(E)\}] \times \{T - \mathcal{E}(T)\}) =$$

$$\mathcal{E}\{T - \mathcal{E}(T)\}^2 + \mathcal{E}[\{T - \mathcal{E}(T)\} \times \{E - \mathcal{E}(E)\}] = \sigma_T^2 + Cov(T,E) =$$

$$\sigma_T^2 + \sigma_T \sigma_E \rho_{TE} = \sigma_T^2.$$

In deze afleiding is gebruik gemaakt van het eerder gegeven resultaat dat de correlatie tussen $T$ en $E$, hier aangeduid met $\rho_{TE}$, gelijk is aan 0. De noemer van de correlatie $X$ en $T$ is gelijk aan $\sigma_X \sigma_T$. We zien dan dat de correlatie $\rho_{XT}$ tussen de geobserveerde

scores $X$ en de ware scores $T$ gelijk is aan $\sigma_T/\sigma_X$; deze uitdrukking is gelijk aan de wortel uit de in (3.4) gegeven uitdrukking voor de betrouwbaarheid.

### 3.4 Standaardmeetfout

De wortel uit de foutenvariantie $\sigma_E^2$ heet de standaardmeetfout. Uit (3.4) kan men afleiden dat de standaardmeetfout $\sigma_E$ kan worden bepaald uit de geobserveerde variantie en de betrouwbaarheid: $\sigma_E = \sigma_X(1-\rho_{XT}^2)^{½}$. De standaardmeetfout is uitgedrukt in de schaal- eenheid van het meetinstrument. Men kan twee standaardmeetfouten van verschillende meetinstrumenten dan ook niet zomaar met elkaar vergelijken. De betrouwbaarheid daaren- tegen is louter een getal; men kan de betrouwbaarheden van twee toetsen wel onderling vergelijken. De standaardmeetfout wordt voornamelijk gebruikt om uit een geobserveerde score een intervalschatting voor de ware score te bepalen.

Men heeft het wel als een bezwaar van de klassieke testtheorie gezien dat er een enkele standaardmeetfout is die wordt toegepast bij elke score $x_{vt}$. Het wordt onrealistisch geacht aan te nemen dat een toets op elk scoreniveau even nauwkeurig meet. Aan dit bezwaar wordt tegemoet gekomen in de itemresponstheorie die in hoofdstuk 4 wordt besproken. Ook binnen de klassieke testtheorie heeft men dit bezwaar erkend. Er zijn diverse procedures ontwikkeld om voor verschillende scoreniveaus een eigen standaardmeetfout te bepalen. Een overzicht van deze procedures vindt men bij Feldt, Steffen en Gupta (1985). Een van die procedures is ontwikkeld door Thorndike (1951).

De methode van Thorndike maakt gebruik van het begrip parallelle metingen. Dit begrip wordt besproken in paragraaf 3.6.1. Een paar eigenschappen van parallelle metingen worden hier gebruikt. Veronderstel dat het mogelijk is, het meetinstrument te verdelen in twee parallelle deeltoetsen. Voor zulke parallelle deeltoetsen, met scorevariabelen $X_1$ en $X_2$, geldt dat $\mathscr{E}(X_1) = \mathscr{E}(X_2)$ en $\sigma_{X_1}^2 = \sigma_{X_2}^2$. Bovendien geldt dat de bijbehorende meetfouten $E_1$ en $E_2$ onderling onafhankelijk, en dus ongecorreleerd zijn. De standaardafwijking van de verschilscore $X_1 - X_2$ kan men nu schrijven:

$$\sigma_{(X_1-X_2)} = \sigma_{(E_1-E_2)} = (\sigma_{E_1}^2 + \sigma_{E_2}^2)^{½} = \sigma_E. \qquad (3.5)$$

In deze afleiding is gebruik gemaakt van het feit dat de correlatie tussen de meetfouten $E_1$ en $E_2$ gelijk is aan 0, van het feit dat $\sigma_{E_1}^2 = \sigma_{E_2}^2$, en van het feit dat $\sigma_{E_1}^2 = ½\sigma_E^2$. Met (3.5) kan men de standaardmeetfout van een meetinstrument schatten. Thorndike

stelt voor, (3.5) toe te passen op deelgroepen van personen die dezelfde score hebben. Zulke groepen noemt men wel scoregroepen. Het is dan mogelijk, met behulp van (3.5) standaardmeetfouten te schatten in verschillende scoregroepen afzonderlijk. In de praktijk zal het vaak nodig zijn, scoregroepen samen te nemen om te komen tot groepen met een voldoende aantal waarnemingen voor het nauwkeurig schatten van de standaardmeetfout.

### 3.5 Schattingen van de ware score

Een voor de hand liggende schatter van de ware score $\tau$ is de waargenomen score $x$. De waargenomen score is een zuivere schatter van de ware score. Men noemt een schatter zuiver als zijn verwachte waarde gelijk is aan de te schatten parameter. De vraag rijst hoe precies de geobserveerde score als schatter van de ware score is. Onder de veronderstelling dat de meetfout binnen elke persoon een normale verdeling heeft met gemiddelde 0 en standaard- afwijking $\sigma_E$, bestaat er een intervalschatting van de ware score. Dit interval bestaat uit de getallen $\hat{\tau}$ waarvoor geldt dat de volgende nulhypothese bij een van te voren vastgesteld significantieniveau niet wordt verworpen:

$$H_0\text{: } x - z \times \sigma_E \leq \hat{\tau} \leq x + z \times \sigma_E \qquad (3.6)$$

waarin $z$ de standaardnormale afwijking is die behoort bij het gekozen significantieniveau. Als dit bijvoorbeeld vastgesteld is op de waarde 0.05, is de waarde van $z$ gelijk aan 1.96. Merk op dat (3.6) een schattingsvoorschrift is. Men kiest eerst de getallen $z$ en $\hat{\tau}$, terwijl $\sigma_E$ bekend is verondersteld. Dan neemt men de realisatie $x_{vt}$ van de toevalsvariabele $X$ waar, en vult de verkregen waarde in (3.6) in. Als de gegeven ongelijkheden worden geschonden, besluit men dat het van te voren gekozen getal $\hat{\tau}$ geen goede schatting is van de ware score. Alle getallen $\hat{\tau}$ waarvoor de ongelijkheden in (3.6) niet geschonden zijn, vormen gezamenlijk een intervalschatting voor de ware score die behoort bij de geobserveerde score $x$. In de praktijk berekent men natuurlijk, zodra de score $x$ is geobserveerd, de intervalgrenzen $x \pm z \times \sigma_E$. Het zo verkregen interval heet in de statistiek een betrouwbaarheidsinterval voor de ware score; de naam heeft niets te maken met het begrip betrouwbaarheid uit de klassieke testtheorie.

Een tweede schatter voor de ware score is de zogenoemde Kelley-schatter (Kelley, 1947; Lord & Novick, 1968). Deze schatter levert een kleinere standaardfout op, maar daarvoor betaalt men wel een prijs. Men moet namelijk veronderstellen dat de regressie

van $T$ op $X$ lineair is. Men kan afleiden dat deze regressie de volgende gedaante heeft:

$$\mathcal{E}(T \mid X = x) = (\rho^2_{XT})\, x + (1 - \rho^2_{XT})\, \bar{x} \qquad (3.7)$$

waarin $\bar{x}$ de gemiddelde geobserveerde score is van de steekproef van personen uit de populatie $\mathcal{P}$ aan wie men de toets heeft afgenomen (zie voor de afleiding Lord en Novick, 1968, p. 65). Zoals Kelley (1947, p. 409) zegt: "This is an interesting equation in that it expresses the estimate of true ability as a weighted sum of two separate estimates - one based upon the individual's observed score, [$x$], and the other based upon the mean of the group to which he belongs, ... If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice versa." De standaardfout van de Kelley-schatter is gelijk aan $\sigma_E\,(\rho^2_{XT})^{\frac{1}{2}}$, de spreiding van het verschil $T - \mathcal{E}(T \mid X = x)$. In de regressie-analyse noemt men deze spreiding wel de spreiding om de regressielijn. Als men de standaardfout van de Kelley-schatter substitueert voor $\sigma_E$ in (3.6) verkrijgt men een andere intervalschatter voor de ware score. Deze schatter leidt tot kleinere intervallen dan de schatter uit (3.6) omdat de gebruikte standaardfout kleiner is dan de in (3.6) als standaardfout gebruikte standaard-meetfout.

In de praktijk zal men niet vaak schattingen van ware scores tegenkomen. De reden daarvan is, dat toetsscores doorgaans relatief worden geïnterpreteerd. Niet de waarde van de score zelf is van belang, maar zijn rangnummer in de verdeling van scores in de populatie $\mathcal{P}$. De beschreven schatters van de ware score leiden tot dezelfde rangorde van personen als de geobserveerde scores; daarom heeft men geen geschatte ware scores nodig. Anders wordt het als een score wordt gerelateerd aan een op voorhand gegeven criterium. Zo'n criterium is bijvoorbeeld een getal waarboven een score moet liggen om als voldoende aangemerkt te worden. Dan bestaat de mogelijkheid, door het gebruik van geschatte ware scores het aantal classificatiefouten te verminderen.

In veel boeken en artikelen over de klassieke testtheorie ziet men verwarring optreden tussen de begrippen standaardfout en standaardmeetfout. De standaardfout, die eigenlijk 'standaardfout van een schatting' (standard error of estimate) heet, is een maat voor de nauwkeurigheid van een schatter. Men kan de nauwkeurigheid van een schatter opvoeren door een grotere steekproef te trekken (hoofdstuk 2). De standaardmeetfout daarentegen is een kenmerk van een toets; het groter maken van een steekproef van aan de toets onderworpen personen heeft op de standaardmeetfout geen enkele invloed. Om de standaardmeetfout kleiner te maken moet men de betrouwbaarheid van de toets groter maken. Een van de middelen daartoe is, de toets met een aantal items te verlengen. Het verlengen van een toets wordt besproken in

paragraaf 3.6.2. De verwarring tussen de begrippen standaardfout en standaardmeetfout wordt wellicht verklaard door het feit dat de standaardmeetfout de rol speelt van standaardfout in (3.6).

## 3.6 Het schatten van de betrouwbaarheid en de standaardmeetfout

Er zijn diverse procedures ontwikkeld om de betrouwbaarheid en de standaardmeetfout van een toets te schatten. Men kan die grootheden immers niet precies bepalen omdat men in de praktijk alleen maar kan beschikken over een steekproef van personen uit de populatie $\mathcal{P}$. In de volgende paragrafen bespreken we methoden om de betrouwbaarheid en de standaardmeetfout te schatten uit parallelle metingen, uit twee afnames van de toets, uit toetsverlenging, en uit coëfficiënt alpha als een ondergrens van de betrouwbaarheid. In paragraaf 3.11 zullen we zien dat men ook de betrouwbaarheid kan schatten door middel van een variantie-analyse van itemscores.

### *3.6.1 Parallelle metingen*

Een belangrijk begrip dat is toegevoegd aan de klassieke testtheorie is dat van de parallelle meting. Men beschikt niet alleen over de realisaties van de geobserveerde toetsscore $X$ maar ook over die van een toetsscore $X'$ die voldoet aan de volgende eigenschappen: $\mathcal{E}(X') = \mathcal{E}(X)$ en $\sigma^2_{X'} = \sigma^2_X$ in elke deelpopulatie van $\mathcal{P}$. Metingen die aan deze eigenschappen voldoen, noemt men parallelle metingen, of ook wel streng parallelle metingen. Beschouw nu de correlatie $\rho_{XX'}$ tussen parallelle metingen. De teller hiervan is gelijk aan:

$$\mathrm{Cov}(X, X') = \mathrm{Cov}(T + E, T + E') = \mathrm{Cov}(T, T) + \mathrm{Cov}(E, E') = \sigma^2_T + \mathrm{Cov}(E, E').$$

Nu wordt er verondersteld dat de bij beide metingen optredende meetfouten $E$ en $E'$ onder- ling onafhankelijk zijn; de meetfouten zijn niet gecorreleerd. Een correlatie ongelijk aan nul zou duiden op de aanwezigheid van een factor die beide metingen systematisch beïnvloedt. Bij parallelle metingen veronderstelt men dat zo'n factor er niet is. De meetfouten worden geacht experimenteel onafhankelijk te zijn. Experimentele onafhankelijkheid brengt met zich mee dat de meetouten niet gecorreleerd zijn. Er geldt dus: $\mathrm{Cov}(E, E') = 0$, en dus $\mathrm{Cov}(X, X') = \sigma^2_T$. De noemer van de correlatie tussen $X$ en $X'$ is gelijk aan: $\sigma_X \sigma_{X'} = \sigma_X \sigma_X = \sigma^2_X$. We zien hieruit dat de correlatie tussen parallelle metingen, $\rho_{XX'}$, gelijk is aan de betrouwbaarheid van

de meting $X$ en ook aan die van de meting $X'$. Dit verklaart het gebruik van het symbool $\rho_{XX'}$ voor de betrouwbaarheid in veel boeken en artikelen over de klassieke testtheorie.

In de praktijk is het niet eenvoudig, parallelle metingen te construeren. Soms slaagt men er in metingen te maken die wel een paar, maar niet alle eigenschappen van parallelle metingen hebben. In tabel 3.1 zijn enige vormen van parallelliteit opgesomd, die afnemen in de strengheid van de eisen.

Tabel 3.1
Enige vormen van parallelliteit

| Soort parallelliteit | Eigenschappen |
|---|---|
| Parallelliteit | $\mathscr{E}(X) = \mathscr{E}(X'),\ \sigma_X^2 = \sigma_{X'}^2$ |
| Tau-equivalentie | $\mathscr{E}(X) = \mathscr{E}(X')$ |
| Essentiële tau-equivalentie | $\mathscr{E}(X) = \mathscr{E}(X') + \kappa\ (\kappa \neq 0)$ |
| Congenerieke parallelliteit | $T = \lambda T' + \kappa,\ (\lambda \neq 0)$ |

In deze tabel zijn $\kappa$ en $\lambda$ constanten die van de meetinstrumenten afhangen. De genoemde eigenschappen gelden in elke deelpopulatie van $\mathscr{P}$. Dat betekent onder meer dat voor elke persoon de ware scores op de parallelle toetsen aan elkaar gelijk zijn, en dus dat $\sigma^2(T) = \sigma^2(T')$. Uit tabel 3.1 ziet men dat men als eerste de veronderstelling laat vallen dat parallelle toetsen dezelfde geobserveerde variantie hebben en dus dezelfde foutenvariantie. Daarna verruimt men de relatie die tussen de ware scores van de beide toetsen bestaat: voor essentieel tau-equivalente metingen verschillen de ware scores een constante, terwijl voor congenerieke metingen de ware scores lineaire transformaties zijn van elkaar. Of aan de diverse vormen van parallelliteit is voldaan, kan men onderzoeken met methoden voor lineaire-structuurmodellen. Zulke methoden zijn beschreven in Bollen (1989).

In de praktijk zal men vaak moeite hebben, meetinstrumenten te maken die aan een van de genoemde definities van parallelliteit voldoen. Daarom heeft men, om de betrouwbaarheid en de standaardmeetfout van een meting $X$ te schatten, methoden bedacht die geen gebruik maken van parallelle metingen. Een van die methoden bestaat eruit, de toets tweemaal af te nemen bij dezelfde personen. Andere methoden vereisen wel dat het mogelijk is het meetinstrument in stukken te verdelen. Bij toetsen die items bevatten, en ook als er diverse beoordelaars zijn, kan men spreken over onderdelen of deeltoetsen.

### *3.6.2 Test-hertestmethode*

Als men niet kan beschikken over parallelvormen van een toets, kan men onder bepaalde omstandigheden dezelfde toets twee keer afnemen bij dezelfde personen. In feite beschouwt men de toets als parallel aan zichzelf. De procedure veronderstelt dat er geen leereffecten kunnen optreden tussen de twee toetsmomenten, en dat tussen die momenten in de populatie niet wezenlijk van karakter verandert. De betrouwbaarheid van de toets kan men dan eenvoudig schatten uit de correlatie tussen de twee verkregen toetsscores.

### *3.6.3 Toetsverlenging*

Een van de methoden om de betrouwbaarheid te schatten, bestaat er uit het meetinstrument op de een of andere wijze in $k$ parallelle delen te verdelen. Elk paar deeltoetsen heeft dezelfde correlatie $\rho$; deze correlatie is dan ook per definitie de betrouwbaarheid van elk der deeltoetsen. Deze betrouwbaarheid $\rho$ wordt bekend verondersteld. In de praktijk kan dit het geval zijn als men een nieuwe toets wil samenstellen uit bestaande toetsen; een dergelijke samengestelde toets noemt men wel een verlengde toets. Als toetsscore op de verlengde toets kiest men de som van de scores op de deeltoetsen. Men kan dan het volgende afleiden. De geobserveerde variantie kan men als volgt schrijven:

$$\sigma_X^2 = \sigma^2\left(\sum_i^k X_i\right) = \sum_i^k \sigma_{X_i}^2 + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j) = k\sigma_{X_i}^2 + \sum_{i\neq j} \sigma_{X_i}\sigma_{X_j}\rho =$$

$$= k\sigma_{X_i}^2 + k(k-1)\sigma_{X_i}^2\rho = k\sigma_{X_i}^2\left[1 + (k-1)\rho\right].$$

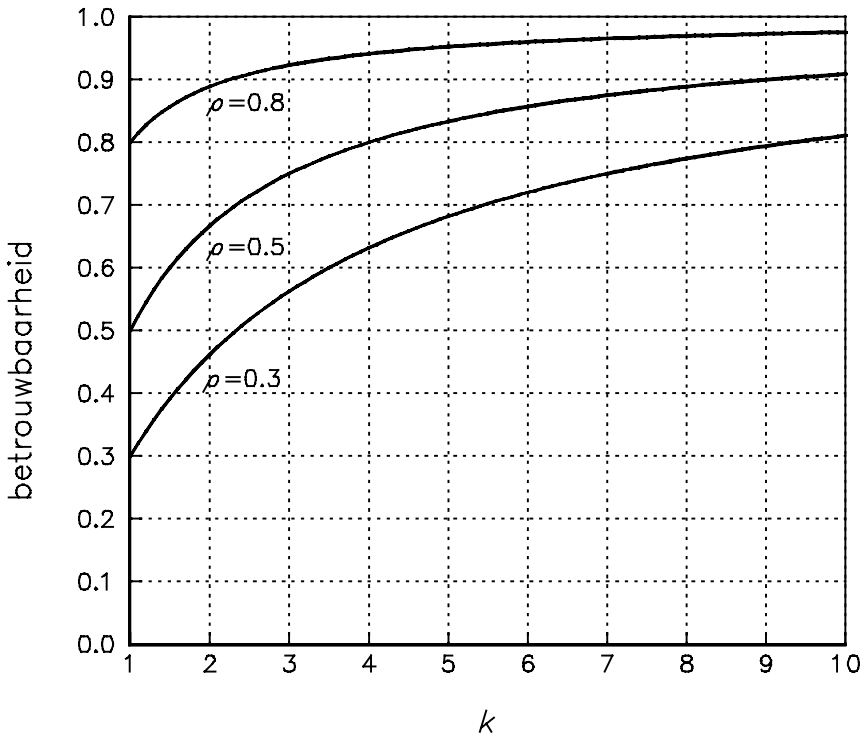
Evenzo kan men de ware variantie schrijven als:

$$\sigma_T^2 = \sigma^2\left(\sum_i^k T_i\right) = \sum_i^k \sigma_{T_i}^2 + \sum_{i\neq j} \mathrm{Cov}(T_i, T_j) = k\sigma_{T_i}^2 + k(k-1)\sigma_{T_i}^2 = k^2\sigma_{T_i}^2.$$

Als men deze twee uitdrukkingen substitueert in formule (3.4), verkrijgt men het volgende resultaat:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{k^2\sigma_{T_i}^2}{k\sigma_{X_i}^2\left[1 + (k-1)\rho\right]} = \frac{k\rho}{1 + (k-1)\rho}. \qquad (3.8)$$

Formule (3.8) is de Spearman-Brown-formule voor toetsverlenging (Brown, 1910; Spearman, 1910). Zij speelt een rol bij het samenstellen van toetsen uit gegeven deeltoetsen of items, vooral om te bepalen of men aan een toets in wording nog delen moet toevoegen om een bepaalde betrouwbaarheid te kunnen bewerkstelligen. In figuur 3.1 is voor een aantal waarden van $\rho$ de betrouwbaarheid uitgezet tegen het aantal deeltoetsen $k$.




Figuur 3.1
Het verband tussen de lengte en de betrouwbaarheid van een toets

In de praktijk wordt de Spearman-Brown-formule voornamelijk gebruikt bij het construeren van toetsen. Een toets met $k$ items blijkt een betrouwbaarheid $\rho$ te hebben. Met behulp van de Spearman-Brown-formule kan men dan uitrekenen hoeveel maal men $k$ items aan de toets moet toevoegen om een gewenste betrouwbaarheid $\rho' > \rho$ te bereiken.

### *3.6.4 Coëfficiënt alpha*

De Spearman-Brown-formule veronderstelt dat men de betrouwbaarheid van de deeltoetsen kent. Aangezien dat in de praktijk dikwijls niet het geval is, kan men gebruik maken van de volgende ongelijkheid:

$$\rho_{XT}^2 \geq \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \sigma_{X_i}^2}{\sigma_X^2}\right]. \qquad (3.9)$$

Het rechterlid van ongelijkheid (3.9) heet coëfficiënt alpha, of ook wel Cronbachs alpha (Cronbach, 1951). Merk op dat coëfficiënt alpha louter te schatten grootheden bevat. Met deze coëfficiënt is dus een ondergrens voor de betrouwbaarheid van een meetinstrument gegeven. De afleiding van coëfficiënt alpha bestaat uit een aantal stappen. In de eerste stap vormen we alle paren deeltoetsen, berekenen in elk paar de som van de ware varianties, en leiden voor de som van deze sommen een ongelijkheid af:

$$\sigma_{(T_i - T_j)}^2 = \sigma_{T_i}^2 + \sigma_{T_j}^2 - 2\,\mathrm{Cov}(T_i, T_j) \geq 0 \Rightarrow \sum_{i \neq j}\left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] \geq 2\sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

De eerste ongelijkheid geldt omdat het linkerlid een variantie is, en dus nooit negatief kan zijn. In de tweede stap berekenen we opnieuw de som van sommen van ware varianties, maar nu met inbegrip van de oneigenlijke paren waarin elke deeltoets met zichzelf wordt gecombineerd. Voor de zo verkregen som leiden we weer een ongelijkheid af, waarbij de in de eerste stap afgeleide ongelijkheid wordt gebruikt:

$$\sum_i \sum_j \left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] = 2k\sum_i \sigma_{T_i}^2 = 2\sum_i \sigma_{T_i}^2 + \sum_{i \neq j}\left[\sigma_{T_i}^2 + \sigma_{T_j}^2\right] \geq$$

$$2\sum_i \sigma_{T_i}^2 + 2\sum_{i \neq j}\mathrm{Cov}(T_i, T_j) \Rightarrow (k-1)\sum_i \sigma_{T_i}^2 \geq \sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

In de derde stap leiden we een eenvoudige ongelijkheid af voor de ware variantie:

$$\sigma_T^2 = \sigma^2\left(\sum_i T_i\right) = \sum_i \sigma_{T_i}^2 + \sum_{i \neq j}\mathrm{Cov}(T_i, T_j) \geq$$

$$\geq \frac{k}{k-1}\sum_{i \neq j}\mathrm{Cov}(T_i, T_j).$$

De som in het rechterlid van deze ongelijkheid kan als volgt worden herschreven:

$$\sum_{i \neq j}\mathrm{Cov}(T_i, T_j) = \sum_{i \neq j}\mathrm{Cov}(X_i, X_j) = \sigma_X^2 - \sum_i \sigma_{X_i}^2.$$

Als we alle ongelijkheden substitueren in formule (3.4), is het resultaat de volgende ongelijkheid:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} \geq \frac{k}{k-1}\left[1 - \frac{\sum_i \sigma^2_{X_i}}{\sigma^2_X}\right]. \qquad (3.10)$$

Als men coëfficiënt alpha beschouwt als een schatter van de betrouwbaarheid, kan men de standaardmeetfout schatten met: $\hat{\sigma}_E = \hat{\sigma}_X\sqrt{(1-\alpha)}$.

In het rechterlid van (3.10), dat gelijk is aan coëfficiënt alpha, ziet men de varianties optreden van de verschillende deeltoetsen. Er is niet verondersteld dat deze varianties aan elkaar gelijk zijn. In feite is het voldoende dat de deeltoetsen essentieel tau-equivalent zijn, als gedefinieerd in tabel 3.1.

Coëfficiënt alpha wordt wel een maat voor de interne consistentie van een toets genoemd. Men noemt een toets intern consistent als de items in de toets niet alle een correlatie van 0 met elkaar hebben. Men kan laten zien dat coëfficiënt alpha op de volgende manier kan worden geschreven:

$$\alpha = \frac{\bar{c}\langle X_i, X_j\rangle}{\sigma^2_{\bar{X}}}. \qquad (3.11)$$

In (3.11) is de teller, $\bar{c}\langle X_i, X_j\rangle$, gelijk aan het gemiddelde van de covarianties tussen alle paren itemscores: $\bar{c}(X_i, X_j) = [k(k-1)]^{-1}\sum_{i\neq j}\mathrm{Cov}(X_i, X_j)$. De noemer is gelijk aan de variantie van het gemiddelde van de itemscores: $\bar{X} = k^{-1}\sum_{i=1}^{k} X_i$. Als alle items onderling perfect correleren, zijn alle varianties van de itemscores aan elkaar gelijk, zijn de covarianties tussen de items gelijk aan deze varianties, en is de gemiddelde itemscore gelijk aan elk der itemscores. Uit (3.11) blijkt dat coëfficiënt alpha in dat geval gelijk is aan 1. Een enkele keer komt men in de literatuur de opvatting tegen dat een toets met een hoge interne consistentie, dus met een hoge waarde van coëfficiënt alpha, een enkele factor in de zin van de factoranalyse meet. Dat deze opvatting op een misverstand berust, is overtuigend aangetoond door Green en Lissitz (1977).

### 3.7 Toets- en itemanalyse

De toets- en itemanalyse is de praktische uitvoering van het schatten van de in de voorafgaande paragrafen beschreven grootheden. Aangezien in de praktijk toetsen

bestaan uit opgaven of items, worden ook kengetallen voor items berekend. Deze laatste grootheden spelen een belangrijke rol in het proces van toetsconstructie. Zij vormen niet alleen de bouwstenen van schattingsformules voor de betrouwbaarheid en de standaardmeetfout, maar zijn ook op zichzelf beschouwd van belang om eigenschappen van items te beschrijven. Doorgaans bepaalt men de kengetallen van items en toetsen in een proefafname: een concepttoets wordt aan een groep personen afgenomen, en op basis van de verkregen gegevens worden de grootheden van de items en de toets geschat. Zonodig worden er items herzien of wordt de samenstelling van de toets veranderd.

In deze paragraaf worden eerst de toets- en itemindices van een toets met meerkeuzevragen besproken. Daarna komen de indices van een toets met open vragen aan de orde voor zover deze niet besproken zijn bij de toets met meerkeuzevragen. In paragraaf 3.8 worden de betrouwbaarheid en de standaardmeetfout apart besproken. Omdat de toets- en itemindices veelal gebaseerd zijn op steekproeven, is paragraaf 3.9 gewijd aan standaardfouten van de geschatte toets- en itemindices. In paragraaf 3.10 tenslotte schenken we aandacht aan normen en richtlijnen voor diverse toets- en itemindices.

Aangezien er in een toets- en itemanalyse voortdurend sprake is van schattingen van grootheden op basis van de gegevens van een steekproef van personen, zal dikwijls de conventie worden gevolgd, de schatters aan te duiden met gewone letters. Zo zal een (schatter van de) variantie worden geschreven als $s^2$ en niet als $\hat{\sigma}^2$.

### *3.7.1 Toets- en itemindices bij toetsen met meerkeuzevragen*

Toetsen met meerkeuzevragen bestaan uit vragen of items waarbij een persoon het goede antwoord moet kiezen uit verschillende alternatieven. We gaan er van uit dat elk goed beantwoord item 1 scorepunt oplevert en elk fout beantwoord item 0 scorepunten. De som van de itemscores vormt de toetsscore van een persoon. De toets- en itemindices worden besproken aan de hand van een toets die een tweekeuze-item en twee driekeuze-items bevat. De toets is door vier personen gemaakt. Dit is weliswaar geen realistische situatie maar het stelt de lezer in staat de indices na te rekenen. De itemantwoorden staan in tabel 3.2. In de kop van deze tabel zijn de goede antwoorden, samen wel de sleutel genoemd, vermeld. De itemantwoorden zijn met behulp van de sleutel omgezet in itemscores. Deze staan samen met de toetsscores in tabel 3.3.

Tabel 3.2

Antwoorden per persoon en per item (tussen haakjes de sleutel)

Tabel 3.3

Itemscores en toetsscores

| persoon | item | | |
|---|---|---|---|
| | 1(B) | 2(A) | 3(C) |
| 1 | B | A | C |
| 2 | B | A | A |
| 3 | B | B | B |
| 4 | A | C | A |

| persoon | item | | | toetsscore |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 1 | 1 | 1 | 1 | 3 |
| 2 | 1 | 1 | 0 | 2 |
| 3 | 1 | 0 | 0 | 1 |
| 4 | 0 | 0 | 0 | 0 |
| som | 3 | 2 | 1 | 6 |

De resultaten van de toets- en itemanalyse van de gegevens uit tabel 3.3 staan in tabel 3.4. De indices uit deze tabel worden in de volgende deelparagraaf besproken.

Tabel 3.4

Resultaten toets- en itemanalyse van de toets met meerkeuzevragen

| | $p$- en $a$-waarden | | | discriminatie-indices | | | | $r_{ir}$- en $r_{ar}$-waarden | | |
|---|---|---|---|---|---|---|---|---|---|---|
| item | A | B | C | $s_i$ | $r_{it}$ | $r_{ir}$ | eff | A | B | C |
| 1 | 0.25 | 0.75* | | 0.43 | 0.77 | 0.52 | 0.30 | –0.52 | 0.52* | |
| 2 | 0.50* | 0.25 | 0.25 | 0.50 | 0.89 | 0.71 | 0.40 | 0.71* | 0.00 | –0.82 |
| 3 | 0.50 | 0.25 | 0.25* | 0.43 | 0.77 | 0.52 | 0.30 | –0.30 | –0.17 | 0.52* |

aantal personen : 4 gemiddelde p-waarde : 0.50
gemiddelde toetsscore : 1.50 betrouwbaarheid (KR-20) : 0.75
standaardafwijking : 1.12 standaardmeetfout : 0.56

### *3.7.2 Itemindices bij toetsen met meerkeuzevragen*

In tabel 3.4 staan de waarden voor de moeilijkheid van een item en de aantrekkelijkheid van de afleiders onder de kop '$p$- en $a$-waarden'. Bij elk alternatief is de fractie personen vermeld die het alternatief heeft gekozen. De fractie waarbij een ster (*) staat, hoort bij het goede antwoord en wordt de $p$-waarde van het item genoemd. De $p$-waarde wordt berekend door het aantal personen dat het item goed heeft, te delen door het aantal personen dat het item heeft gemaakt. De bij de afleiders of foute antwoorden vermelde fracties worden de $a$-waarden van het item genoemd en worden berekend door het aantal personen dat een afleider heeft gekozen te delen door het aantal personen dat het item heeft gemaakt. Bij item 2 in ons voorbeeld, een driekeuze-item, zien we bij de alternatieven A, B en C respectievelijk de waarden

0.50*, 0.25 en 0.25 staan. Dit betekent dat alternatief A het goede antwoord is met een $p$-waarde van 0.50. De $a$-waarden van de alternatieven B en C zijn beide gelijk aan 0.25.

Een $p$-waarde ligt per definitie tussen 0 en 1. Bij een $p$-waarde gelijk aan 0 hebben alle personen het item fout; bij een $p$-waarde gelijk aan 1 hebben alle personen het item goed. Het kan voorkomen dat een item een afleider heeft met een $a$-waarde die groter is dan de $p$-waarde. Dit kan er op wijzen dat een afleider niet fout is of dat het als goed bestempelde alternatief wellicht niet goed is. In het algemeen geeft een hoge $a$-waarde ons informatie over het item die in combinatie met andere informatie tot een definitief oordeel over de kwaliteit van het item moet leiden.

Onder het kopje '$s_i$' is de standaardafwijking van de items vermeld. De standaardafwijking van een item, $s_i$, wordt bij dichotome scores berekend als: $s_i = \sqrt{p\,q} = \sqrt{p(1-p)}$, waarin $p$ de $p$-waarde van het item is en $q$ gelijk is aan 1 - $p$. Wanneer alle personen een item goed dan wel fout hebben, is de standaardafwijking gelijk aan 0. De standaardafwijking is maximaal als $p$ = 0.50, dus als de ene helft van de personen het item fout heeft en de andere helft het item goed. In dat geval is $s_i = \sqrt{0.5(1-0.5)} = 0.5$.

Omdat een item een onderdeel van een toets is, zijn er diverse indices ontwikkeld om de samenhang tussen een itemscore en de toetsscore weer te geven. Een index die veel gebruikt wordt is de $r_{it}$. De $r_{it}$ is de produkt-moment-correlatie tussen de itemscore en de toetsscore. Deze correlatie wordt bij dichotoom gescoorde items wel puntbiseriële correlatie genoemd: het is de correlatie tussen een dichotome en een continu geachte variabele. Een produkt- moment-correlatie neemt waarden aan tussen +1 en -1. Een correlatie van +1 betekent dat er een perfect positief lineair verband bestaat tussen twee variabelen, in ons geval tussen de itemscore en de toetsscore. Dat de $r_{it}$-waarden in tabel 3.4 zo hoog zijn, heeft te maken met het feit dat de toets uit slechts drie items bestaat. Bij toetsen van veertig of meer items is een $r_{it}$ van 0.50 al hoog (zie tabel 3.12).

De $r_{it}$ wordt een discriminatie-index genoemd omdat zij aangeeft in hoeverre een item onderscheid maakt tussen personen met hoge toetsscores en personen met lage toetsscores. Een hoge $r_{it}$ betekent dat veel personen met een hoge toetsscore het item goed hebben beantwoord en veel personen met een lage toetsscore het item fout hebben beantwoord. Later zullen we zien dat een hoge $r_{it}$ ook betekent dat het item relatief veel bijdraagt aan de betrouwbaarheid van de toets (zie paragraaf 3.8.1).

Hiervoor zagen we dat de $r_{it}$ een produkt-moment-correlatie is. Die kan met een van de algemene formules voor een correlatie berekend worden. Afgeleid kan worden dat voor dichotome scores de $r_{it}$ van een item ook geschreven kan worden als:

$$r_{it} = \frac{\bar{x}_g - \bar{x}_f}{s_x} \sqrt{p(1-p)}\ , \tag{3.12}$$

waarin:
$\bar{x}_g$ = gemiddelde toetsscore van de personen die het item goed hebben,
$\bar{x}_f$ = gemiddelde toetsscore van de personen die het item fout hebben,
$s_x$ = standaardafwijking van de toetsscores.

De teller in het deel voor het wortelteken in (3.12) maakt duidelijk waarom we de $r_{it}$ een discriminatie-index noemen: hoe groter het verschil tussen $\bar{x}_g$ en $\bar{x}_f$, des te groter de $r_{it}$.

Naast de $r_{it}$ is de $r_{ir}$ een veel gebruikte discriminatie-index. De $r_{ir}$ is een soortgelijke index als de $r_{it}$. Gaat het bij de $r_{it}$ om de correlatie tussen itemscores en toetsscores, bij de $r_{ir}$ gaat het om de correlatie tussen itemscores en restscores. De restscore van een persoon is gelijk aan zijn toetsscore minus de score op het desbetreffende item. Een persoon heeft dus evenzoveel restscores als er items zijn in de toets.

Zowel aan de $r_{it}$ als aan de $r_{ir}$ kleven bezwaren. De $r_{it}$ geeft een geflatteerd beeld van de samenhang tussen de score op een item en de toetsscore, omdat de itemscore onderdeel is van de toetsscore. We correleren dus het item voor een deel met zichzelf. De $r_{ir}$ ondervangt dit bezwaar, maar heeft als bezwaar dat de restscore waarmee een item gecorreleerd wordt, met het item varieert. De $r_{ir}$-waarden van eenzelfde toets zijn daardoor onderling niet te vergelijken. Als echter het aantal items in een toets veertig of meer is, zijn beide bezwaren van geen belang meer.

Nog een andere maat om het discriminerend vermogen van een item te karakteriseren is het effectieve gewicht dat te vinden is onder het kopje 'eff'. Onder het effectieve gewicht verstaan we de bijdrage van een item aan de spreiding van toetsscores. Hoe hoger het effectieve gewicht van een item is, des meer spreiding in de toetsscores toegeschreven kan worden aan het item. Het volgende kan worden afgeleid (Gulliksen, 1950; Ferguson & Takane, 1989):

$$\sum_{i=1}^{k} r_{it} s_i = s_x, \tag{3.13}$$

waarin $k$ het aantal items is.

Het effectieve gewicht van item $i$ is gedefinieerd als:

$$\frac{r_{it} \times s_i}{s_x}. \tag{3.14}$$

De teller in (3.14) wordt de itembetrouwbaarheidsindex genoemd en is een onderdeel van de formule om de betrouwbaarheid van de toets te schatten (zie paragraaf 3.8.1). Uit (3.14) volgt dat de som van de effectieve gewichten gelijk is aan 1. In ons voorbeeld van tabel 3.4 heeft item 2 een effectief gewicht van 0.40; dat betekent dat het item voor 40% bijdraagt aan de standaardafwijking van de toetsscores. Een andere interpretatie van het effectieve gewicht wordt gegeven door regressie-analyse. Als men de lineaire regressievergelijking van de itemscore op de toetsscore opstelt, blijkt de regressiecoëfficiënt gelijk te zijn aan het effectieve gewicht van het item.

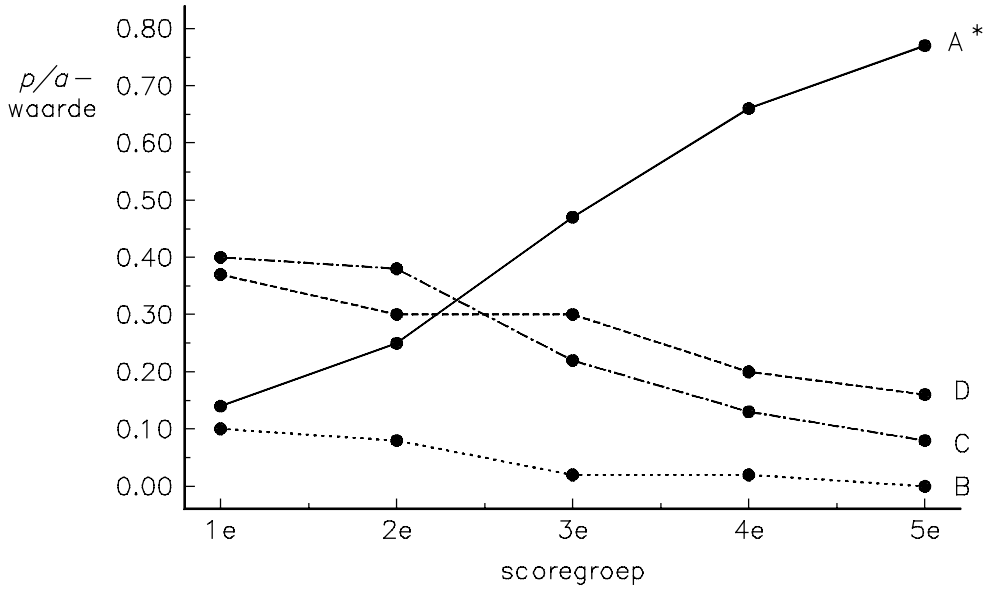
Bij een toets met meerkeuzevragen is het mogelijk, naast een discriminatie-index voor het goede antwoord discriminatie-indices voor de afleiders (foute antwoorden) te berekenen. In tabel 3.4 kunnen we zien dat er bij elk item $r_{ar}$-waarden zijn vermeld naast de $r_{ir}$-waarde. Per item zijn er uiteraard evenveel $r_{ar}$-waarden als er afleiders zijn. De $r_{ar}$ wordt berekend door personen die het desbetreffende foute antwoord hebben gekozen een itemscore 1 en de anderen een itemscore 0 te geven. Vervolgens wordt de correlatie tussen het foute antwoord en de restscore berekend, waarbij de restscore per definitie dezelfde waarde heeft als bij de berekening van de $r_{ir}$. Omdat we toetsen met een hoge betrouwbaarheid nastreven, zijn items met positieve $r_{ir}$- en negatieve $r_{ar}$-waarden gewenst. Zulke waarden impliceren dat relatief veel personen met een hoge toetsscore het item goed hebben beantwoord en relatief veel personen met een lage toetsscore het item fout hebben beantwoord. Een positieve $r_{ar}$ geeft aan dat relatief veel goede personen de desbetreffende afleider als het goede antwoord hebben aangemerkt. Soms kan dit een sleutelfout zijn: de verkeerde sleutel is per ongeluk opgegeven of bij nader inzien blijkt dat de afleider met de positieve $r_{ar}$ het goede antwoord is.

Tabel 3.5

Per scoregroep de $p$- en $a$-waarden van een item

| score | n | A* | B | C | D |
|---|---|---|---|---|---|
| 0 - 18 | 123 | 0.14 | 0.10 | 0.40 | 0.37 |
| 19 - 22 | 124 | 0.25 | 0.08 | 0.38 | 0.30 |
| 23 - 29 | 124 | 0.47 | 0.02 | 0.22 | 0.30 |
| 30 - 35 | 124 | 0.66 | 0.02 | 0.13 | 0.20 |
| 36 - 47 | 124 | 0.77 | 0.00 | 0.08 | 0.16 |
| 0 - 47 | 619 | 0.46* | 0.04 | 0.24 | 0.26 |
| gem. score | 26.0 | 30.8 | 18.8 | 21.0 | 23.5 |

Het discriminerend vermogen van een item kunnen we ook weergeven door de personen in een aantal scoregroepen op te delen en vervolgens per scoregroep de *p*- en *a*-waarden te berekenen. Als voorbeeld presenteren we in tabel 3.5 van een item de *p*- en *a*-waarden per scoregroep. In die tabel lezen we dat alternatief A het goede antwoord is met een *p*-waarde van 0.46. Van de afleiders is D het meest aantrekkelijk met een *a*-waarde van 0.26. Verder zien we dat de totale groep van 619 personen is opgesplitst in vijf bijna even grote scoregroepen. Bekijken we nu van het item de *p*-waarde per scoregroep, dan heeft het item in de minst vaardige groep, met scores tussen 0 en 18, een *p*-waarde van 0.14. De *p*-waarde van het item wordt groter met het vaardiger worden van de groep, en in de meest vaardige groep heeft het item een *p*-waarde van 0.77. Bij de afleiders is de tendens andersom; hoe vaardiger de groep, des te lager de *a*-waarde. Het item is dus een voorbeeld van een goed discriminerend item: de *p*-waarde van het item is in de groep van de beste personen veel hoger dan in de groep van de slechtste personen, en de *a*-waarden van het item zijn voor de slechtste personen hoger dan de *a*-waarden voor de beste personen. De *p*- en *a*-waarden uit tabel 3.5 zijn grafisch weergegeven in figuur 3.2. De keuze van het aantal scoregroepen is arbitrair. Om er echter voor te zorgen dat de standaardfout van een fractie niet te groot wordt, moet het aantal personen per scoregroep niet te klein zijn (zie tabel 3.8).




Figuur 3.2
Per scoregroep *p*- en *a*-waarden van het item uit tabel 3.5

### *3.7.3 Toetsindices bij toetsen met meerkeuzevragen*

Behalve informatie over de drie afzonderlijke items uit de toets, bevat tabel 3.4 ook informatie die betrekking heeft op de toets als geheel. We kunnen in de tabel lezen dat vier personen, $n$ = 4, de toets gemaakt hebben. Een maat voor de moeilijkheidsgraad van een toets is de gemiddelde toetsscore $\bar{x}$, die bij deze toets gelijk is aan 6/4=1.50. De standaardafwijking van de toetsscores, $s_x$, is een maat voor de spreiding van de toetsscores en kan als volgt berekend worden:

$$s_x = \left( \frac{\sum_{v=1}^{n} (x_v - \bar{x})^2}{n} \right)^{1/2}, \qquad (3.15)$$

waarin $x_v$ de toetsscore is van persoon $v$.

De standaardafwijking kan volgens (3.13) ook verkregen worden door de itembetrouwbaar-
heidsindices te sommeren. Wanneer de standaardafwijking gelijk is aan 0, hebben alle personen dezelfde toetsscore. De standaardafwijking is maximaal wanneer de ene helft van de personen alle items goed heeft en de andere helft alle items fout.

De gemiddelde $p$-waarde, $\bar{p}$, is het gemiddelde van de $p$-waarden van de afzonderlijke items. Bij toetsen met meerkeuzevragen kan de gemiddelde $p$-waarde berekend worden hetzij door alle $p$-waarden op te tellen en de som te delen door het aantal items $k$, hetzij door de gemiddelde toetsscore te delen door het aantal items in de toets. In formulevorm:

$$\bar{p} = \frac{\sum_{i=1}^{k} p_i}{k} \text{ of } \bar{p} = \frac{\bar{x}}{k}. \qquad (3.16)$$

De toetsindices betrouwbaarheid en standaardmeetfout worden in paragraaf 3.8 besproken.

### *3.7.4 Toets- en itemindices bij toetsen met open vragen*

Bij toetsen met open vragen moeten personen zelf het antwoord formuleren op de vragen die voorgelegd worden. Het is gebruikelijk dat er per vraag meer dan een

scorepunt behaald kan worden en dat de antwoorden door beoordelaars met behulp van een correctievoorschrift gescoord worden. In deze paragraaf gaan we er van uit dat beoordelaars geen factor zijn die de meetprocedure verstoren. In dat geval is er ook geen wezenlijk verschil tussen de analyse van een toets met open vragen en de analyse van een toets met meerkeuzevragen. Het enige verschil is dat er bij open vragen andere itemscores dan alleen maar 0 en 1 mogelijk zijn. Indien beoordelaars wel een storende factor zijn, dient er een analyse als beschreven in paragraaf 3.13 plaats te vinden.

In het voorbeeld in tabel 3.6 gaan we uit van vier open vragen die door zes personen beantwoord zijn. Op elke vraag kunnen maximaal twintig punten behaald worden.

Tabel 3.6

Itemscores en toetsscores

| persoon | item | | | | toetsscore |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 1 | 17 | 8 | 14 | 3 | 42 |
| 2 | 16 | 10 | 13 | 5 | 44 |
| 3 | 18 | 15 | 14 | 18 | 65 |
| 4 | 16 | 14 | 14 | 8 | 52 |
| 5 | 14 | 7 | 7 | 4 | 32 |
| 6 | 17 | 15 | 17 | 16 | 65 |
| som | 98 | 69 | 79 | 54 | 300 |

De resultaten van de toets- en itemanalyse staan in tabel 3.7. Aangezien de toets- en itemanalyse van open vragen voor een deel dezelfde indices bevat als de toets- en itemanalyse van meerkeuzevragen, komen hierna niet meer alle toets- en itemindices aan de orde. Alleen de voor open vragen specifieke indices worden besproken.

Tabel 3.7

Resultaten van de toets- en itemanalyse van de toets met open vragen

| item | max. score | gem. score | $p'$ | $s_i$ | $r_{it}$ | $r_{ir}$ | eff |
|---|---|---|---|---|---|---|---|
| 1 | 20.00 | 16.33 | 0.82 | 1.25 | 0.81 | 0.77 | 0.08 |
| 2 | 20.00 | 11.50 | 0.58 | 3.30 | 0.95 | 0.91 | 0.26 |
| 3 | 20.00 | 13.17 | 0.66 | 3.02 | 0.81 | 0.69 | 0.20 |
| 4 | 20.00 | 9.00 | 0.45 | 5.89 | 0.94 | 0.79 | 0.46 |

aantal personen : 6 gemiddelde $p'$-waarde : 0.63

| | | | |
|---|---|---|---|
| gemiddelde toetsscore | : 50.00 | betrouwbaarheid (alpha) | : 0.82 |
| standaardafwijking | : 12.10 | standaardmeetfout | : 5.12 |

### *3.7.5 Itemindices bij toetsen met open vragen*

Bij een toets met open vragen kan het aantal te behalen scorepunten van vraag tot vraag variëren. Daarom is in tabel 3.7 een kolom met het opschrift 'max. score' opgenomen. In deze kolom staat het aantal punten dat op een item behaald kan worden. In het voorbeeld zijn bij alle items de maxima gelijk.

Een andere voor open vragen specifieke index staat in de kolom met opschrift 'gem. score'. In deze kolom staat de gemiddelde score die op elk van de items behaald is. Bij ongelijke maximale scores zijn de gemiddelde itemscores niet vergelijkbaar. Daarom wordt de $p'$-waarde berekend; deze staat in de kolom met het opschrift '$p'$'. De $p'$-waarde duidt de moeilijkheidsgraad van een item aan, en wordt berekend door de gemiddelde itemscore te delen door de maximale itemscore. Merk op dat we bij open vragen over de $p'$-waarde spreken en bij meerkeuzevragen over de $p$-waarde. De definitie van de twee grootheden is gelijk; het verschil in notatie heeft geen andere functie dan aan te geven om welke soort vraag het gaat.

### *3.7.6 Toetsindices bij toetsen met open vragen*

Bij toetsen met open vragen worden dezelfde toetsindices berekend als bij toetsen met meerkeuzevragen. Om misverstanden te voorkomen, verdient de berekening van de gemiddelde $p'$-waarde enige toelichting. De gemiddelde $p'$-waarde wordt berekend door de gemiddelde toetsscore te delen door de maximaal te behalen toetsscore. In tegenstelling tot bij een toets met meerkeuzevragen mag de gemiddelde $p'$-waarde bij een toets met open vragen alleen maar op deze manier berekend worden en niet via de $p'$-waarden van de individuele vragen. Als men dat wel zou doen, zou men verschillen in maximaal te behalen itemscores veronachtzamen.

## 3.8 Betrouwbaarheid en standaardmeetfout

Bij de toets- en itemanalyse van de meerkeuzevragen is de KR-20 als betrouwbaarheidsmaat berekend en bij de toets- en itemanalyse van de open vragen coëfficiënt alpha. Hierna laten we zien dat de KR-20 een speciaal geval is van coëfficiënt alpha. In paragraaf 3.5 zijn twee manieren besproken om met behulp van de standaardmeetfout een intervalschatting voor de ware score te bepalen. Deze twee manieren worden in paragraaf 3.8.3 gebruikt om intervalschattingen te verkrijgen voor ware verschilscores.

### *3.8.1 Coëfficiënt alpha en de KR-20*

Het is gebruikelijk, de betrouwbaarheid van een toets met coëfficiënt alpha te schatten. De formule voor coëfficiënt alpha is gegeven in het rechterlid van (3.9). Omdat bij dichotoom gescoorde vragen geldt dat $s_i^2 = p_i q_i$, kan coëfficiënt alpha voor dichotoom gescoorde items geschreven worden als:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} p_i q_i}{s_x^2}\right]. \qquad (3.17)$$

Formule (3.17) staat bekend als de KR-20 en is onafhankelijk van Cronbachs coëfficiënt alpha door Kuder en Richardson (1937) ontwikkeld. Vanwege (3.12) kan coëfficiënt alpha ook geformuleerd worden als:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{\left(\sum_{i=1}^{k} r_{it} s_i\right)^2}\right]. \qquad (3.18)$$

Uit (3.18) laat zich het verband tussen de $r_{it}$ en de betrouwbaarheid nog niet eenvoudig aflezen. Bij dichotoom gescoorde items liggen de itemvarianties in de praktijk tussen 0.21 en 0.25 ($0.3 < p < 0.7$). Indien we de itemvarianties nu als constant beschouwen voor alle items, kunnen we afleiden (Thorndike, 1982):

$$\alpha \approx \frac{k}{k-1}\left(1 - \frac{1}{k(\bar{r}_{it})^2}\right), \qquad (3.19)$$

waarin $\bar{r}_{it}$ het gemiddelde van de $r_{it}$-waarden is.

### *3.8.2 Verschilscores*

In paragraaf 3.5 zijn schattingen van de ware score aan de orde geweest. Er is op gewezen dat het schatten van ware scores niet altijd nodig is. In de praktijk zou men willen weten of een toetsscore van 30 voor Kay en een toetsscore van 33 voor Wilko betekent dat de laatstgenoemde meer weet dan Kay. Daar kan men niet achter komen, omdat men de ware scores van Kay en Wilko niet kent. Wel kan men iets zeggen over het volgende probleem. Als men aselect twee personen uit de populatie trekt waarvan de waargenomen scores drie punten verschillen, kan men dan zeggen of dit verschil substantieel is? Statistisch gezien betekent dit dat we de nulhypothese willen toetsen dat de ware toetsscores van de twee aselect getrokken personen gelijk zijn. Noem deze ware scores $\tau_1$ en $\tau_2$, en de geobserveerde scores $x_1$ en $x_2$. Veronderstel dat de geobserveerde scores $x_1$ en $x_2$ normaal verdeeld zijn met verwachte waarden $\tau_1$ respectievelijk $\tau_2$, en beide met standaardafwijking $\sigma_E$. Dan is de verschilscore $x_1 - x_2$ normaal verdeeld met gemiddelde $\tau_1 - \tau_2$ en standaardafwijking $\sigma_E\sqrt{2}$. Naar analogie van (3.6) kunnen we een intervalschatting maken van het verschil $\delta = \tau_1 - \tau_2$. Dit interval bestaat uit alle waarden $\hat{\delta}$ waarvoor de volgende nulhypothese niet wordt verworpen:

$$H_0\text{: } (x_1 - x_2) - z \times \sigma_E\sqrt{2} \leq \hat{\delta} \leq (x_1 - x_2) + z \times \sigma_E\sqrt{2}\,.$$

Veronderstel dat de toets een standaardmeetfout $\sigma_E$ heeft van 1, dan vindt men, bij een verschil van drie punten in geobserveerde scores, het 95%-betrouwbaarheidsinterval: $0.23 \leq \tau_1 - \tau_2 \leq 5.77$. Aangezien dit interval niet de waarde 0 bevat, zal men bij een waargenomen verschil van drie punten, de hypothese verwerpen dat de bijbehorende ware scores aan elkaar gelijk zijn.

Men kan ook een intervalschatting voor verschilscores bepalen op basis van de in paragraaf 3.5 genoemde Kelley-schatter. Men kan afleiden dat de verschilscore $\delta = \tau_1 - \tau_2$ een verwachte waarde heeft gelijk aan $\rho^2_{XT}(x_1 - x_2)$ en een standaardafwijking gelijk aan $(2\rho^2_{XT}\sigma^2_E)^{½}$. Voor een toets met een betrouwbaarheid van 0.80 en een standaardmeetfout van 1 is, bij een verschil in waargenomen scores van 3 punten, het 95%-betrouwbaarheidsinterval gelijk aan: $-0.08 \leq \tau_1 - \tau_2 \leq 4.88$. Nu zal men de nulhypothese van gelijke ware scores niet verwerpen. Merk op dat het laatst

gegeven betrouwbaarheidsinterval iets kleiner is dan het eerst gegeven interval: 4.96 tegenover 5.54.

## 3.9 Nauwkeurigheid van toets- en itemindices

Bij het berekenen van toets- en itemindices is het buitengewoon belangrijk dat men er zich rekenschap van geeft hoe nauwkeurig die indices geschat zijn. De statistiek geeft ons op deze vraag een antwoord omdat het mogelijk is betrouwbaarheidsintervallen te construeren. Zoals reeds eerder is aangegeven, is een betrouwbaarheidsinterval een stochastisch interval om een steekproefwaarde dat met een gegeven kans de te schatten populatiewaarde bevat. De $p$-waarde, de gemiddelde score, de $r_{it}$-waarde, de KR-20 en coëfficiënt alpha zijn allemaal voorbeelden van grootheden die gebaseerd zijn op steekproeven en daardoor behept met steekproeffouten. In de volgende paragrafen zullen we op deze steekproeffouten en op de constructie van betrouwbaarheidsintervallen ingaan.

### *3.9.1 Standaardfout van een p-waarde*

De standaardfout $s_p$ van een $p$-waarde wordt met de volgende formule berekend:

$$s_p = \left(\frac{p(1-p)}{n}\right)^{½}. \qquad (3.20)$$

In (3.20) staat $n$ voor het aantal personen in de aselect getrokken steekproef. Nu zegt een vuistregel in de statistiek dat, indien $n > \{9 \times (1-p)/p\}$ bij $p \leq 0.50$ en $n > \{9 \times p/(1-p)\}$ bij $p \geq 0.50$, een $p$-waarde bij benadering normaal verdeeld is. Hiervan uitgaande, kunnen we een betrouwbaarheidsinterval construeren voor de werkelijke $p$-waarde. Veronderstel dat de geschatte $p$-waarde van een item 0.20 is en dat het item door 100 personen is gemaakt, dan is de bijbehorende standaardfout $\sqrt{0.2 \times 0.8/100} = 0.04$. We kunnen dan bijvoorbeeld de grenzen van het 95%-betrouwbaarheidsinterval berekenen. Uit de berekening volgt dat in 95% van de gevallen bij items met een geschatte $p$-waarde van 0.20 de werkelijke $p$-waarde tussen 0.12 en 0.28 zal liggen (0.12 = 0.20 - 1.96 × 0.04 en 0.28 = 0.20 +1.96 × 0.04). In tabel 3.8, die gebaseerd is op exacte berekeningen (De Jonge, 1963), kan men bij $p = 0.20$ en $n = 100$ aflezen dat de grenzen 0.13 en 0.29 zijn. De afwijkingen zijn minimaal.

Tabel 3.8

95%-betrouwbaarheidsintervallen voor fracties

| steekproef -fractie $p$ | aantal personen in de steekproef ($n$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 50 | | 100 | | 200 | | 500 | | 1000 | |
| 0.00 | 0.00 | 0.07 | 0.00 | 0.04 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.00 |
| 0.10 | 0.03 | 0.22 | 0.05 | 0.18 | 0.06 | 0.15 | 0.08 | 0.13 | 0.08 | 0.12 |
| 0.20 | 0.10 | 0.34 | 0.13 | 0.29 | 0.15 | 0.26 | 0.17 | 0.24 | 0.18 | 0.23 |
| 0.30 | 0.18 | 0.45 | 0.21 | 0.40 | 0.24 | 0.37 | 0.26 | 0.34 | 0.27 | 0.33 |
| 0.40 | 0.26 | 0.55 | 0.30 | 0.50 | 0.33 | 0.47 | 0.36 | 0.45 | 0.37 | 0.43 |
| 0.50 | 0.35 | 0.65 | 0.40 | 0.60 | 0.43 | 0.57 | 0.46 | 0.55 | 0.47 | 0.53 |
| 0.60 | 0.45 | 0.74 | 0.50 | 0.70 | 0.53 | 0.67 | 0.55 | 0.64 | 0.57 | 0.63 |
| 0.70 | 0.55 | 0.82 | 0.60 | 0.79 | 0.63 | 0.76 | 0.66 | 0.74 | 0.67 | 0.73 |
| 0.80 | 0.66 | 0.90 | 0.71 | 0.87 | 0.74 | 0.85 | 0.76 | 0.83 | 0.77 | 0.82 |
| 0.90 | 0.78 | 0.97 | 0.82 | 0.95 | 0.85 | 0.94 | 0.87 | 0.92 | 0.88 | 0.92 |
| 1.00 | 0.93 | 1.00 | 0.96 | 1.00 | 0.98 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 |

### *3.9.2 Standaardfout van een gemiddelde toetsscore en van een p'-waarde*

De standaardfout $s_{\bar{x}}$ van de gemiddelde toetsscore $\bar{x}$ is gelijk aan:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}. \tag{3.21}$$

Neem als voorbeeld een toets die door 429 personen gemaakt is, en waarvan de gemiddelde toetsscore gelijk is aan 32.24 en de standaardafwijking van de toetsscores 6.29 is. De standaardfout bedraagt dan 0.30 en het 95%-betrouwbaarheidsinterval heeft de grenzen 31.64 en 32.84.

De standaardfout $s_p$ van een $p'$-waarde is gelijk aan:

$$s_{p'} = \frac{s_i}{m\sqrt{n}}. \tag{3.22}$$

In (3.22) staat $m$ voor de maximaal te behalen score op de vraag. Bij de toets met open vragen in tabel 3.7 heeft item 4 een $p'$-waarde van 0.45. We kunnen daarvan de standaardfout berekenen; deze bedraagt 0.12. Het 95%-betrouwbaarheidsinterval voor de werkelijke $p'$-waarde heeft de grenzen 0.14 en 0.76. Dit interval is groot omdat zo weinig personen het item gemaakt hebben.

### *3.9.3 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 3.9

95%-betrouwbaarheidsintervallen voor $r_{it}$-waarden

| $r_{it}$-waarde (steekproef) | aantal personen 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 |

Tabel 3.9 is gebaseerd op Iker en Perry, en is van toepassing op $p$-waarden die tussen 0.20 en 0.80 liggen. In tabel 3.9 staan voor diverse waarden van de $r_{it}$ en $n$ de 95%-betrouwbaarheids- intervallen voor de werkelijke waarden van de $r_{it}$ vermeld. Indien bijvoorbeeld bij een toets- en itemanalyse die gebaseerd is op 1000 personen, de $r_{it}$-waarde van een item 0.20 is, dan zijn de 95%-betrouwbaarheidsgrenzen van de werkelijke $r_{it}$-waarde 0.14 en 0.26.

### *3.9.4 Standaardfout van coëfficiënt alpha*

Voor coëfficiënt alpha heeft Feldt (1965) de steekproefverdeling afgeleid waarop tabel 3.10 gebaseerd is. In deze tabel zijn bij diverse steekproefwaarden van coëfficiënt alpha de onder- en bovengrenzen vermeld van het 95%-betrouwbaarheidsinterval voor de werkelijke waarde van coëfficiënt alpha. De tabel mag alleen gebruikt worden indien een toets tien of meer vragen bevat. Als bijvoor- beeld de betrouwbaarheid van een toets die is afgenomen bij 500 personen gelijk is aan 0.70, dan loopt het 95%-betrouwbaarheidsinterval van 0.66 tot 0.74.

Tabel 3.10

95%-betrouwbaarheidsintervallen voor coëfficiënt alpha

| α (steekproef) | aantal personen 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 |

## 3.10 Normen voor toets- en itemindices

In de volgende paragrafen worden normen en richtlijnen voor toets- en itemindices geformuleerd. We moeten bedenken dat deze normen en richtlijnen opgesteld zijn met de gedachte dat we er naar moeten streven een toets met een zo hoog mogelijke betrouwbaarheid te construeren. Nogmaals dient er op gewezen te worden dat de indices bij kleine aantallen personen een relatief kleine precisie hebben, zodat voorzichtigheid geboden is bij de interpretatie van zulke indices.

### *3.10.1 Normen voor p- en p'-waarden*

In de literatuur vinden we verschillende opvattingen over de optimale $p$-waarde van een item. Crocker en Algina (1986) stellen dat de optimale $p$-waarde halverwege de raadkans en 1.0 moet liggen. De veronderstelling hierbij is dat er geraden wordt als men niet weet wat het goede antwoord op een meerkeuze-item is. In formulevorm uitgedrukt: $p = 0.5 + 0.5/m$, waarin $m$ het aantal alternatieven is en $p$ de gewenste $p$-waarde. Naar aanleiding van een simulatie-onderzoek komt Lord (1952) tot een andere conclusie. De aanbevelingen van voornoemde auteurs over de optimale $p$-waarde van items met verschillende aantallen alternatieven staan in tabel 3.11.

De conclusie van een onderzoek van Feldt (1993) is, dat de optimale $p$-waarde tussen 0.57 en 0.67 moet liggen wanneer er geraden kan worden. Indien er geen reden is om aan te

Tabel 3.11

Optimale $p$-waarde bij items met 2-5 alternatieven

| aantal alternatieven | optimale $p$-waarde ($p$=0.5+0.5/$m$) | optimale $p$-waarde (Lord) |
|---|---|---|
| 2 | 0.75 | 0.85 |
| 3 | 0.67 | 0.77 |
| 4 | 0.63 | 0.74 |
| 5 | 0.60 | 0.70 |

nemen dat er geraden wordt, of als er niet geraden kan worden zoals bij open vragen, is de
optimale $p$-waarde gelijk aan 0.50. Het effect van de moeilijkheid van een item op de betrouwbaarheid blijkt echter verbazingwekkend klein te zijn, zelfs als de $p$-waarden variëren van 0.27 tot 0.79.

### *3.10.2 Normen voor $r_{it}$-waarden*

Ook voor $r_{it}$-waarden vindt men in de literatuur geen absolute normen. Zoals bekend kan een produkt-moment-correlatie, dus ook een $r_{it}$-waarde, variëren tussen -1 en +1. Een $r_{it}$-waarde van 0.50 en hoger is echter in de praktijk bij toetsen met meer dan veertig items al erg hoog. Ebel en Frisbie (1986) komen tot de in tabel 3.12 vermelde normen voor de $r_{it}$-waarden.

Tabel 3.12

Normen voor $r_{it}$-waarden

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

Omdat de grootte van de $r_{it}$ onder andere afhankelijk is van het aantal items in een toets, moet men strikt genomen bovenstaande normen alleen hanteren bij $r_{it}$-waarden die gecorrigeerd zijn voor toetslengte. De correctie kan uitgevoerd worden met een correctie-formule van Henrysson (1963). Vanwege het geringe effect kan de correctie achterwege blijven indien de items afkomstig zijn uit toetsen met veertig of meer items.

### *3.10.3 Normen voor de betrouwbaarheid*

In de literatuur wordt 0.85 als vereiste ondergrens voor de betrouwbaarheid van een toets genoemd wanneer de vaardigheid van een groep personen op basis van slechts een enkele toets wordt bepaald. Wanneer de vaardigheid met meer toetsen of op verschillende momenten wordt getoetst zijn lagere ondergrenzen acceptabel, waarbij in de literatuur 0.65 wel als gewenste ondergrens wordt genoemd (Frisbie, 1988).

Een mogelijke norm voor de betrouwbaarheid zouden we kunnen ontlenen aan het percentage ten onrechte gezakte en ten onrechte geslaagde personen, ofwel het percentage niet-consistente beslissingen, bij een selectietoets (Dousma & Horsten, 1989). Met de ten onrechte gezakte en de ten onrechte geslaagde personen bedoelen we de personen waarvoor, indien ze een parallelle toets hadden afgelegd, de beslissing anders geweest had kunnen zijn. Het percentage niet-consistente beslissingen neemt toe als de betrouwbaarheid lager wordt en ook als het percentage gezakten stijgt, waarbij het percentage gezakten afhangt van de cesuur of grensscore. Tabel 3.13 laat de percentages niet-consistente beslissingen zien als functie van het percentage gezakten en van de betrouwbaarheid. Daarbij moet opgemerkt worden dat het gebruik van de tabel alleen zinvol is wanneer de toetsscores ongeveer normaal verdeeld zijn.

Tabel 3.13

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 |

In tabel 3.13 kunnen we zien dat bij een toets met een betrouwbaarheid van 0.80 en met een percentage gezakten van 30, het percentage niet-consistente beslissingen gelijk aan 18 is. Dat wil dan zeggen dat 9% van de gezakten tot de geslaagden zou kunnen hebben behoord en 9% van de geslaagden tot de gezakten. Dus voor 18% van alle personen had de beslissing anders kunnen zijn.

## 3.11 Generaliseerbaarheidstheorie

De bespreking van de generaliseerbaarheidstheorie, (Cronbach, Gleser, Nanda & Rajaratnam, 1972), in dit hoofdstuk bestaat uit vier paragrafen. Het begrippenkader dat in de generaliseerbaarheidstheorie gehanteerd wordt en dat in belangrijke mate ontleend is aan de variantie-analytische literatuur, wordt in deze paragraaf besproken. In paragraaf 3.12 wordt de generaliseerbaarheidstheorie behandeld aan de hand van de analyse van de toets met meerkeuzevragen die in paragraaf 3.7 met de klassieke testtheorie geanalyseerd is. In paragraaf 3.13 wordt de generaliseerbaarheidstheorie verder toegelicht aan de hand van een analyse van een toets waarbij beoordelaars de antwoorden van personen op vragen beoordelen. In beide paragrafen wordt aandacht besteed aan verschillen tussen de klassieke testtheorie en generaliseerbaarheidstheorie. In paragraaf 3.14 komen kort een aantal andere aspecten van de generaliseerbaarheidstheorie aan de orde. Merk op dat de notatie die in de paragrafen 3.11 tot en met 3.14 gehanteerd wordt afwijkt van die uit voorgaande paragrafen. De reden hiervoor, is de notatie aan te laten sluiten bij de in de literatuur gebruikelijke notatie.

In de generaliseerbaarheidstheorie worden observaties of metingen beschreven in termen van de condities waaronder zij geobserveerd worden. Condities van een bepaalde soort worden aangeduid als 'facet'. De dertig meerkeuzevragen van een toets zijn volgens deze terminologie de dertig condities van het facet 'vragen'. En bij een toets bestaande uit tien open vragen waarbij de antwoorden door twee beoordelaars beoordeeld worden, spreken we over de tien condities van het facet 'vragen' en de twee condities van het facet 'beoordelaars'. Het door personen laten beantwoorden van vragen, kunnen we opvatten als een gestandaardiseerd experiment (Meerling, 1981). Een proefopzet waarin responsen of antwoorden van personen op (condities van het facet) vragen worden geobserveerd, wordt een een-facet-design genoemd. Een proefopzet waarin de observaties beoordelingen zijn van responsen van personen op (condities van het facet) vragen die beoordeeld worden door (condities van het facet) beoordelaars, wordt een twee-facet-design genoemd. Het aantal observaties dat per

persoon verkregen wordt, is afhankelijk van het design dat gebruikt wordt. Wanneer we aan tien personen een toets van dertig vragen voorleggen, een zogenaamd gekruist een-facet-design (personen × vragen), hebben we per persoon dertig observaties. Zouden we echter aan elke persoon drie andere vragen voorleggen, dan hebben we per persoon slechts drie observaties. Wanneer we aan tien personen een toets van tien vragen voorleggen en de responsen op de tien vragen laten beoordelen door twee beoordelaars, een zogenaamd gekruist twee-facet-design (personen × vragen × beoordelaars), krijgen we twintig observaties per persoon. Zouden we echter vijf vragen door de eerste beoordelaar en vijf andere vragen door de tweede beoordelaar laten beoordelen, dan krijgen we tien observaties per persoon.

Voor het bepalen van de rekenvaardigheid van personen, kunnen we antwoorden van personen op meerkeuzevragen observeren. De verzameling van alle denkbare observaties die naar onze mening acceptabel of geschikt zijn voor het geven van een oordeel over personen, wordt in de generaliseerbaarheidstheorie het universum genoemd. Uiteraard zouden we het bepalen van de rekenvaardigheid van personen willen baseren op de observaties of scores verkregen op alle vragen uit het universum, de universumscores. Om praktische redenen kunnen we de personen echter niet meer dan een steekproef van bijvoorbeeld dertig vragen uit het universum voorleggen. Het bepalen van de rekenvaardigheid baseren we op de scores die op de dertig vragen behaald worden, de geobserveerde scores. De nauwkeurigheid waarmee we menen te kunnen generaliseren van geobserveerde scores naar universumscores, dat wil zeggen de geobserveerde scores kunnen opvatten als universumscores, wordt 'generaliseerbaarheid' genoemd. Als maat voor de generaliseerbaarheid wordt de generaliseerbaarheidscoëfficiënt gebruikt. Deze coëfficiënt heeft een benedengrens van 0 en een bovengrens van 1.

In het geval van de meerkeuzevragen bestaat het universum alleen uit het facet vragen. Bestaat het universum niet uit meerkeuzevragen maar uit open vragen waarvan de antwoorden door beoordelaars beoordeeld moeten worden, dan kunnen we de beoordeling door alle in aanmerking komende beoordelaars laten verrichten. In dit geval bestaat het universum uit twee facetten: het facet 'open vragen' en het facet 'beoordelaars.' De universumscores zijn gelijk aan de scores die verkregen zouden zijn na het beoordelen van alle antwoorden op alle open vragen door alle beoordelaars. Aangezien we in de praktijk de beoordeling zullen moeten beperken tot een klein aantal beoordelaars, zijn de geobserveerde scores van de personen de scores verkregen na het beoordelen van de open vragen door dit kleine aantal beoordelaars.

De voorbeelden laten zien dat voor het generaliseren naar een universum een duidelijke beschrijving van het universum een voorwaarde is. Deze beschrijving bevat

in de eerste plaats de facetten waaruit het universum bestaat. In het eerste voorbeeld bestaat het universum alleen uit het facet 'vragen'. In het tweede voorbeeld bestaat het universum uit de facetten 'vragen' en 'beoordelaars'. In de tweede plaats moet een beschrijving van het universum uitsluitsel geven over de condities die binnen het universum vallen. Dit heeft te maken met het belangrijke onderscheid dat in de variantie-analyse aangeduid wordt met de termen 'random' en 'fixed'. In het eerste voorbeeld zijn de vragen uit de toets opgevat als een aselecte of random steekproef uit een zeer grote verzameling of 'oneindig universum' van vragen. In het tweede voorbeeld zijn vragen en beoordelaars opgevat als een random steekproef uit een oneindig universum van vragen en beoordelaars. In het voorbeeld van de meerkeuzevragen impliceert een random facet dat we vinden dat ook dertig andere vragen in aanmerking hadden kunnen komen om de rekenvaardigheid van personen te bepalen. Deze twee (of meer) toetsen van dertig vragen worden in de generaliseerbaarheidstheorie random parallelle toetsen genoemd. Voor het voorbeeld van de open vragen betekent een random facet 'open vragen' en een random facet 'beoordelaars' dat we vinden dat ook tien andere open vragen en twee andere beoordelaars in aanmerking hadden kunnen komen om de vaardigheid te bepalen. Zouden we in het tweede voorbeeld vinden dat slechts twee bepaalde beoordelaars in aanmerking komen, dan spreken we van een fixed facet 'beoordelaars'. Bij een fixed facet hebben we alle condities van een facet in ons design opgenomen en hoeven dan ook niet te generaliseren naar het universum. Later zullen we zien dat het onderscheid tussen random en fixed facetten consequenties voor de generaliseerbaarheid heeft.

### 3.12 Design met een facet

In een gekruist een-facet-design wordt de geobserveerde score van een persoon op een item, $X_{pv}$, uitgedrukt als een decompositie in vier componenten:

$$\begin{aligned} X_{pv} &= \mu && = \text{algemeen gemiddelde} \\ &+ \mu_p - \mu && = \text{persoonseffect} \\ &+ \mu_v - \mu && = \text{itemeffect} \\ &+ X_{pv} - \mu_p - \mu_v + \mu && = \text{residu} \end{aligned} \tag{3.23}$$

In (3.23) is de eerste component, het algemene gemiddelde, gedefinieerd als $\mu \equiv \mathcal{E}_p \mathcal{E}_v X_{pv}$, de gemiddelde score (= verwachting over personen en items) verkregen na het beantwoorden van alle items uit het universum door alle personen uit de

populatie. Het algemene gemiddelde geeft dezelfde constante bijdrage aan de geobserveerde score van alle personen.

De universumscore van een persoon is hier gedefinieerd als $\mu_p \equiv \mathscr{E}_v X_{pv}$, de gemiddelde score (= verwachting over items) van een persoon verkregen na het beantwoorden van alle items uit het universum van items. De tweede component, het persoonseffect $\mu_p - \mu$, is gelijk aan het verschil tussen de universumscore van een persoon en het algemene gemiddelde. Personen met een positief persoonseffect hebben een score die hoger is dan het algemene gemiddelde terwijl personen met een negatief persoonseffect een score hebben die lager is dan het algemene gemiddelde. Verschillen in vaardigheid tussen personen kunnen we weergeven als verschillen tussen hun persoonseffecten.

De moeilijkheidsgraad van een item is gedefinieerd als $\mu_v \equiv \mathscr{E}_p X_{pv}$, de gemiddelde score (= verwachting over personen) van een item na het beantwoorden van het item door alle personen uit de populatie. De derde component, het itemeffect $\mu_v - \mu$, is gelijk aan het verschil tussen de moeilijkheidsgraad van een item en het algemene gemiddelde. Een item met een positief itemeffect is gemakkelijker dan een item met een negatief itemeffect. Verschillen in moeilijkheidsgraad tussen items kunnen we weergeven als verschillen tussen hun itemeffecten.

De vierde component, de foutencomponent of het residu, is het verschil tussen $X_{pv}$ en de eerste drie componenten. Zoals we in het voorbeeld van tabel 3.15 zullen zien, beschikken we bij het gekruiste een-facet-design maar over een enkele observatie voor elke combinatie van persoon en vraag. Dit betekent dat we het persoons- × itemeffect niet kunnen onderscheiden van andere foutenbronnen. Behalve het persoons- × itemeffect bevat het residu alle foutencomponenten die de geobserveerde score doen afwijken van de som van de eerste drie componenten.

Met uitzondering van het algemene gemiddelde, hebben de componenten in (3.23) een verdeling. Uit de wijze waarop de effecten in (3.23) gedefinieerd zijn, volgt dat hun gemiddelden gelijk zijn aan nul. De definitie van het gemiddelde van het persoonseffect bijvoorbeeld luidt $\mathscr{E}_p(\mu_p - \mu) = \mathscr{E}_p(\mu_p) - \mathscr{E}_p(\mu) = \mu - \mu = 0$. De drie componenten hebben ook elk een eigen variantie die we aanduiden met variantiecomponent. De variantiecomponenten voor respectievelijk personen, items en het residu zijn gedefinieerd als:

$$\sigma_p^2 = \mathscr{E}_p(\mu_p - \mu)^2, \tag{3.24}$$

$$\sigma_v^2 = \mathscr{E}_v(\mu_v - \mu)^2, \text{ en} \tag{3.25}$$

$$\sigma^2_{pv,e} = \mathcal{E}_p \mathcal{E}_v (X_{pv} - \mu_p - \mu_v + \mu)^2. \quad (3.26)$$

De notatie van de variantiecomponent voor het residu laat zien dat de component uit een variantiecomponent personen × vragen en een variantiecomponent voor de fouten (error) bestaat.

De variantie van de geobserveerde scores is gedefinieerd als

$$\sigma^2_X = \sigma^2_{(X_{pv})} = \mathcal{E}_p \mathcal{E}_v (X_{pv} - \mu)^2,$$

en deze totale variantie is gelijk aan de som van de drie variantiecomponenten, ofwel

$$\sigma^2_X = \sigma^2_p + \sigma^2_v + \sigma^2_{pv,e}. \quad (3.27)$$

### *3.12.1 Generaliseerbaarheidsstudie*

Om schattingen van de variantiecomponenten van effecten te verkrijgen, dienen we een onderzoek, of wat wel genoemd wordt een generaliseerbaarheidsstudie of G-studie, uit te voeren. Het schatten gebeurt met behulp van procedures uit de variantie-analyse. We bespreken hieronder een gekruist design waarbij $n_p$ personen en $n_v$ items of vragen aselecte steekproeven zijn uit respectievelijk een populatie van personen en een universum van items. Tabel 3.14 bevat de variantie-analysetabel van dit gekruist random-effecten-design.

Tabel 3.14
Variantie-analysetabel van een gekruist design met twee random effecten

| Effecten | Kwadraten-sommen | Vrijheids-graden | Gemiddelde kwadratensommen | Verwachte gemiddelde kwadratensommen |
|---|---|---|---|---|
| Personen *(p)* | $SS_p$ | $df_p = n_p - 1$ | $MS_p = SS_p/df_p$ | $\mathcal{E}(MS_p) = \sigma^2_{pv,e} + n_v\sigma^2_p$ |
| Items *(v)* | $SS_v$ | $df_v = n_v - 1$ | $MS_v = SS_v/df_v$ | $\mathcal{E}(MS_v) = \sigma^2_{pv,e} + n_p\sigma^2_v$ |
| Residu *(pv,e)* | $SS_{pv,e}$ | $df_{pv,e} = (n_p - 1) \times (n_v - 1)$ | $MS_{pv,e} = SS_{pv,e}/df_{pv,e}$ | $\mathcal{E}(MS_{pv,e}) = \sigma^2_{pv,e}$ |

Schattingen van de variantiecomponenten krijgen we door het oplossen van vergelijkingen voor de verwachte gemiddelde kwadratensommen (expected mean squares). Daartoe worden de verwachte gemiddelde kwadratensommen gelijkgesteld aan de geobserveerde gemiddelde kwadratensommen (mean squares) en de exacte waarden van de variantiecomponenten vervangen door de geschatte waarden. Dit resulteert in de volgende vergelijkingen:

$$MS_{pv,e} = \hat{\sigma}^2_{pv,e},$$

$$MS_v = \hat{\sigma}^2_{pv,e} + n_p\hat{\sigma}^2_v, \text{ ofwel } \hat{\sigma}^2_v = (MS_v - MS_{pv,e})/n_p,$$

$$MS_p = \hat{\sigma}^2_{pv,e} + n_v\hat{\sigma}^2_p, \text{ ofwel } \hat{\sigma}^2_p = (MS_p - MS_{pv,e})/n_v.$$

Omdat de gemiddelde kwadratensom voor het residu gelijk is aan de schatting van de variantiecomponent voor het residu, $\hat{\sigma}^2_{pv,e} = MS_{pv,e}$, kunnen we de vergelijking voor de gemiddelde kwadratensom voor de items schrijven als $\hat{\sigma}^2_v = (MS_v - \hat{\sigma}^2_{pv,e})/n_p$. Door in deze vergelijking de gemiddelde kwadratensom van de items, berekend door het uitvoeren van een variantie-analyse, en de geschatte waarde voor de variantiecomponent van het residu in te vullen, verkrijgen we een schatting van de variantiecomponent voor items. Door herschrijven van de vergelijking voor de gemiddelde kwadratensom van de personen als $\hat{\sigma}^2_p = (MS_p - \hat{\sigma}^2_{pv,e})/n_v$, verkrijgen we op analoge wijze een schatting van de variantiecomponent voor personen.

In tabel 3.14 zien we, dat we om de drie variantiecomponenten te kunnen schatten, over de kwadratensommen (sums of squares) dienen te beschikken. Daartoe vervangen we de drie parameters $\mu, \mu_p$ en $\mu_v$ in (3.14) door hun geobserveerde equivalenten, wat resulteert in de volgende decompositie:

$$X_{pv} = \bar{X} + (\bar{X}_p - \bar{X}) + (\bar{X}_v - \bar{X}) + (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X}). \qquad (3.28)$$

We illustreren de berekening van de kwadratensommen aan de hand van het voorbeeld in tabel 3.15. Deze tabel bevat de itemscores die vier personen op drie items behaald hebben. Daarnaast bevat de tabel de volgende statistische grootheden: de toetsgemiddelden, $\bar{X}_p$, van de vier personen, de itemgemiddelden, $\bar{X}_v$, van de drie items en het algemene gemiddelde, $\bar{X}$. Merk op dat het voorbeeld gelijk aan is aan het voorbeeld dat in paragraaf 3.7 bij de behandeling van de klassieke testtheorie besproken is. Voor de observaties en grootheden in deze tabel hebben we vergelijking (3.24) uitgeschreven in tabel 3.16.

De kwadratensom voor personen berekenen we door de getallen uit de kolom $(\bar{X}_p - \bar{X})$ van tabel 3.16 te kwadrateren en dan te sommeren.

Tabel 3.15

De itemscores van vier personen op drie items, de gemiddelde score per persoon en per item en het algemene gemiddelde

| | Item | | | |
|---|---|---|---|---|
| Persoon | 1 | 2 | 3 | $\bar{X}_p$ |
| 1 | 1 | 1 | 1 | 1.00 |
| 2 | 1 | 1 | 0 | .67 |
| 3 | 1 | 0 | 0 | .33 |
| 4 | 0 | 0 | 0 | .00 |
| $\bar{X}_v$ | .75 | .50 | .25 | 0.50 = $\bar{X}$ |

Op analoge wijze verkrijgen we de kwadratensom voor de items uit de kolom $(\bar{X}_v - \bar{X})$, en die voor het residu uit de kolom $(X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})$.

Tabel 3.16

Vergelijking (3.28) uitgeschreven voor de observaties en grootheden uit tabel 3.15

| $X_{pv} =$ | $\bar{X}$ | $+(\bar{X}_p - \bar{X})$ | $+ (\bar{X}_v - \bar{X})$ | $+ (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})$ |
|---|---|---|---|---|
| $X_{11} = 1 =$ | .500 | + .500 | + .250 | — .250 |
| $X_{12} = 1 =$ | .500 | + .500 | + .000 | + .000 |
| $X_{13} = 1 =$ | .500 | + .500 | — .250 | + .250 |
| $X_{21} = 1 =$ | .500 | + .167 | + .250 | + .083 |
| $X_{22} = 1 =$ | .500 | + .167 | + .000 | + .333 |
| $X_{23} = 0 =$ | .500 | + .167 | — .250 | — .417 |
| $X_{31} = 1 =$ | .500 | — .167 | + .250 | + .417 |
| $X_{32} = 0 =$ | .500 | — .167 | + .000 | — .333 |
| $X_{33} = 0 =$ | .500 | — .167 | — .250 | — .083 |
| $X_{41} = 0 =$ | .500 | — .500 | + .250 | — .250 |
| $X_{42} = 0 =$ | .500 | — .500 | + .000 | + .000 |
| $X_{43} = 0 =$ | .500 | — .500 | — .250 | + .250 |

Voor de berekening van de totale kwadratensom brengen we in vergelijking (3.28) het algemene gemiddelde naar het linkerlid waardoor we in tabel 3.16 een nieuwe kolom, $(X_{pv} - \bar{X})$, krijgen. De getallen in deze kolom worden gekwadrateerd en daarna gesommeerd. De totale kwadratensom, $SS_{tot}$, is gelijk aan de som van de drie andere kwadratensommen en wordt geschreven als:

$$\Sigma_p \Sigma_v (X_{pv} - \bar{X})^2 = n_v \Sigma_p (\bar{X}_p - \bar{X})^2 + n_p \Sigma_v (\bar{X}_v - \bar{X})^2 + \Sigma_p \Sigma_v (X_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})^2,$$

of:

$$\Sigma_p \Sigma_v (X_{pv} - \bar{X})^2 = SS_p + SS_v + SS_{pv,e}.$$

Tabel 3.17 bevat de resultaten van de generaliseerbaarheidsstudie voor de data uit tabel 3.15.
We laten het aan de lezer over de resultaten in tabel 3.17 na te rekenen. In de laatste kolom van de tabel staan de schattingen van de variantiecomponenten voor de drie effecten. Aangezien de grootte van de componenten afhangt van de scoreschaal die gebruikt wordt, geeft de absolute grootte van de variantiecomponenten ons geen bruikbare informatie.

Tabel 3.17
Resultaten generaliseerbaarheidsstudie voor data uit tabel 3.15

| Effecten | Kwadraten-sommen | Vrijheids-graden | Gemiddelde kwadratensommen | Schattingen van variantiecomponenten |
|---|---|---|---|---|
| Personen *(p)* | 1.667 | 3 | 0.555 | $\hat{\sigma}^2_p$ = 0.139 (45.5%) |
| Items *(v)* | 0.500 | 2 | 0.250 | $\hat{\sigma}^2_v$ = 0.028 (9%) |
| Residu *(pv,e)* | 0.833 | 6 | 0.139 | $\hat{\sigma}^2_{pv,e}$ = 0.139 (45.5%) |

Vandaar dat we voor elke component de procentuele bijdrage aan de totale variantie vermelden. In verband met de interpretatie van de variantiecomponenten willen we er met verwijzing naar de definities (3.24)-(3.27) nog eens benadrukken dat de variantie-componenten het resultaat zijn van de decompositie van de geschatte totale variantie van scores van afzonderlijke personen op afzonderlijke items. Dit betekent dus dat $\hat{\sigma}^2_v$ en $\hat{\sigma}^2_{pv,e}$ geen variantiecomponenten van gemiddelde of totaalscores zijn. Merk op dat we de items dichotoom gescoord hebben, zodat de variantiecomponenten in de tabel nooit groter kunnen zijn dan 0.25. De variantiecomponent voor de personen, de

geschatte universumscore-variantie, bedraagt bijna de helft van de totale variantie. De geschatte variantiecomponent voor de items is relatief klein. De geschatte variantiecomponent voor het residu is ook relatief groot. Deze variantiecomponent bestaat uit de interactiecomponent personen × vragen en andere foutenvariantie. Wanneer het residu louter uit de interactiecomponent zou bestaan, zou dit betekenen dat de rangorde van de personen niet voor alle items gelijk is. Dit zou in het voorbeeld het geval geweest zijn wanneer de eerste persoon het derde item fout en de vierde persoon het derde item goed beantwoord zou hebben.

### *3.12.2 Decisiestudie*

Tot nu toe had de bespreking uitsluitend betrekking op de decompositie van een score van een persoon op een item uit het universum van items. Een persoon krijgt echter altijd een toets voorgelegd die uit een aantal items bestaat. Decisies of beslissingen over een persoon zijn dan ook altijd gebaseerd op de gemiddelde score of de totaalscore die behaald is op dat aantal items. In ons voorbeeld bestaat de toets uit drie random getrokken rekenitems uit het universum van rekenitems. Een andere toets met ook drie random getrokken items uit hetzelfde universum zouden we ook geschikt gevonden hebben voor het meten van de rekenvaardigheid. Dit betekent dat het universum waar in dit geval naar gegeneraliseerd wordt, het universum van random parallelle toetsen met drie items is.

Het lineaire model voor de decompositie van de gemiddelde score van een persoon op een toets met $n_v$ items, aangeduid met $X_{pV}$, luidt:

$$X_{pV} = \mu + (\mu_p - \mu) + (\mu_V - \mu) + (X_{pV} - \mu_p - \mu_V + \mu). \qquad (3.29)$$

Vergelijking (3.29) is gelijk aan vergelijking (3.23) met dit verschil dat we in (3.29) de score, behaald op een enkel item, vervangen hebben door de gemiddelde score behaald op $n_v$ items. In de notatie van (3.29) wordt een hoofdletter *V* gebruikt om aan te geven dat het de gemiddelde score van $n_v$ items betreft. In (3.29) wordt de universumscore gedefinieerd als $\mu_p = \mathcal{E}_V X_{pV}$, de verwachte waarde van $X_{pV}$ over random parallelle toetsen. De definities van de variantiecomponenten zijn gelijk aan die van (3.24), (3.25) en (3.26) met dien verstande dat *v* vervangen is door *V*. Het spreekt vanzelf dat door bij (3.24) de verwachting over *V* te nemen, de universumscorevariantie $\sigma_p^2$ niet verandert. De twee andere variantiecomponenten zijn: $\sigma_V^2 = \sigma_v^2 / n_v$ en $\sigma_{pV,e}^2 = \sigma_{pV,e}^2 / n_v$. Deze twee variantiecomponenten hebben betrekking op de populatie van personen en

het universum van random parallelle toetsen. De variantiecomponent $\sigma^2_V = \sigma^2_v / n_v$ moet geïnterpreteerd worden als de variantie van de verdeling van gemiddelde scores van random parallelle toetsen. De totale variantie, $\sigma^2_X = \sigma^2_{(XpV)}$ is gelijk aan $\sigma^2_X = \sigma^2_p + \sigma^2_V + \sigma^2_{pV,e}$. Wat het voorgaande betekent voor ons voorbeeld, hebben we samengevat in tabel 3.18.

In tabel 3.18 zien we hoe groot de variantiecomponenten die we in de generaliseerbaarheids-studie (G-studie) geschat hebben, in een zogenaamde decisiestudie (D-studie) worden wanneer de toets uit $n_v$ items bestaat. Voor een gekruist een-facet-random-effect design zijn twee decisies of beslissingen van belang: de beslissing of we de toets voor het nemen van relatieve of absolute beslissingen zullen gebruiken en de beslissing uit hoeveel items we onze toets moeten laten bestaan.

Tabel 3.18

Resultaten decisiestudie voor data uit tabel 3.15

| Effecten | Variantiecomponenten G-studie | Variantiecomponenten D-studie |
|---|---|---|
| Personen *(p)* | $\hat{\sigma}^2_p = 0.139$ | $\hat{\sigma}^2_p = 0.139$ |
| Items *(v)* | $\hat{\sigma}^2_v = 0.028$ | $\hat{\sigma}^2_V = 0.028/3 = .009$ |
| Residu *(pv,e)* | $\hat{\sigma}^2_{pv,e} = 0.139$ | $\hat{\sigma}^2_{pV,e} = 0.139/3 = .046$ |

Het doel van een toets kan zijn, vast te stellen hoe de prestatie van een persoon zich verhoudt tot de prestaties van andere personen. Wanneer beslissingen over personen gebaseerd zijn op wat personen presteren in relatie tot andere personen, spreken we van relatieve beslissingen. De mate waarin we er met de toets in slagen personen van elkaar te onderscheiden, drukken we uit in een generaliseerbaarheidscoëfficiënt voor relatieve beslissingen. Voor het gekruiste één-facet-random-effect-design is de schatting van deze generaliseerbaarheidscoëfficiënt, een ratio van variantiecomponenten, gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pv,e}}{n_v}}. \qquad (3.30)$$

De noemer van (3.30) bevat de universumscorevariantie $\hat{\sigma}^2_p$ en de foutenvariantie $\hat{\sigma}^2_{pv,e}/n_v$. Merk op dat de variantiecomponent $\hat{\sigma}^2_v/n_v$ niet als foutenvariantie in de noemer van (3.30) voorkomt. De reden hiervoor is dat verschillen in gemiddelde scores van random parallelle toetsen geen rol spelen wanneer we personen met elkaar willen

vergelijken. Wanneer we willen beslissen of Jan beter kan rekenen dan Piet, dan maakt het niet uit of we ze een toets met makkelijke of een toets met moeilijke items voorleggen. Brennan (1992, p. 16) laat formeel zien dat verschillen tussen scores van personen de voor beiden gelijke itemcomponent doet wegvallen.

We kunnen aan (3.30) zien dat we de coëfficiënt kunnen verhogen door de toets uit meer items laten bestaan waardoor de foutenvariantie kleiner zal worden. Omdat (3.30) een schatting van de generaliseerbaarheidscoëfficiënt na toetsverlenging geeft, wordt de formule ook wel de 'stepped-up generalizability coëfficiënt' genoemd. In hoofdstuk 11 laten we zien hoe (3.30) herschreven en gebruikt kan worden als de Spearman-Brown-formule voor toetsverlenging uit de klassieke testtheorie.

In tabel 3.18 zien we dat voor de toets met drie items de universumscorevariantie gelijk is aan .139, en de foutenvariantie aan .139/3 = .046. De generaliseerbaarheidscoëfficiënt is gelijk aan .139/{.139 + .046} = 0.75. De generaliseerbaarheidscoëfficiënt kan op twee manieren geïnterpreteerd worden. De eerste interpretatie is dat de coëfficiënt bij benadering gelijk is aan de verwachte waarde van de gekwadrateerde correlatie tussen geobserveerde en universumscores. Daarnaast kan de coëfficiënt geïnterpreteerd worden als de correlatie tussen de scores van twee random parallelle toetsen, elk bestaande uit $n_v$ items.

Met behulp van de gemiddelde kwadratensommen kunnen we (3.30) ook uitdrukken als:

$$\hat{\rho}^2 = \frac{MS_p - MS_{pv,e}}{MS_p}. \qquad (3.31)$$

Bewezen kan worden dat in het geval van dichotome scores (3.31) gelijk is aan de KR-20 en in het geval van polytome scores aan Cronbachs coëfficiënt alpha (Sirotnik, 1970).

Het doel van de toets kan ook zijn, vast te stellen of personen in staat zijn een bepaalde prestatie te leveren, bijvoorbeeld tachtig procent van de items uit het universum goed te beantwoorden. In deze situatie zijn we niet geïnteresseerd in wat een persoon presteert in vergelijking met andere personen, maar in het absolute prestatieniveau van de persoon. Beslissingen die gebaseerd zijn op het absolute prestatieniveau van een persoon worden absolute beslissingen genoemd. In dit geval spelen verschillen in toetsen wel degelijk een rol bij de beslissing of personen aan het gewenste prestatieniveau voldoen. Wanneer een toets namelijk uit makkelijke items bestaat, kan eerder aan het prestatieniveau voldaan worden dan wanneer de toets uit moeilijke items bestaat. Dit betekent dat wanneer met een toets absolute beslissingen over personen genomen worden, $\hat{\sigma}_v^2/n_v$ bijdraagt aan de foutenvariantie.

De schatting van de generaliseerbaarheidscoëfficiënt voor absolute beslissingen is gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_v^2}{n_v} + \frac{\hat{\sigma}_{pv,e}^2}{n_v}}. \qquad (3.32)$$

Door de variantiecomponenten uit tabel 3.18 in (3.32) in te vullen, verkrijgen we de generaliseerbaarheidscoëfficiënt voor de toets uit ons voorbeeld. De coëfficiënt is gelijk aan .139/{.139 + .028/3 + .139/3} = 0.72. Merk op dat de coëfficiënten voor relatieve en absolute beslissingen slechts weinig verschillen. Dit verschil wordt uiteraard nog kleiner als we de toets verlengen.

Het onderscheid tussen relatieve en absolute beslissingen wijst op een belangrijk verschil tussen de generaliseerbaarheidstheorie en de klassieke testtheorie. De assumptie van parallelle toetsen in de klassieke testtheorie impliceert namelijk dat de gemiddelde toetsscores gelijk zijn wat betekent dat $\hat{\sigma}_v^2/n_v$ per definitie gelijk is aan nul. Dit sluit aan op de praktijk dat met de klassieke testtheorie doorgaans alleen relatieve beslissingen maar geen absolute beslissingen over personen genomen worden.

### 3.13 Design met twee facetten

Hiervoor hebben we de verschillende fasen van de analyse van een-facet-design besproken. Aangezien de analyse van een twee-facet-design op vergelijkbare wijze verloopt, kan de bespreking van de diverse fasen relatief kort zijn. Een voorbeeld van een gekruist twee-facet- design is een design waarbij de antwoorden op vragen van personen beoordeeld worden door beoordelaars. In een gekruist twee-facet-design wordt de geobserveerde score van een persoon *p* op een item *v*, toegekend door een beoordelaar *b*, $X_{pvb}$, uitgedrukt als een decompositie van de score in zeven componenten:

$$\begin{aligned}
X_{pvb} &= \mu && \text{(algemene gemiddelde)}\\
&+ \mu_p - \mu && \text{(persoonseffect)}\\
&+ \mu_v - \mu && \text{(itemeffect)}\\
&+ \mu_b - \mu && \text{(beoordelaarseffect)}\\
&+ \mu_{pv} - \mu_p - \mu_v + \mu && \text{(persoons- × itemeffect)}\\
&+ \mu_{pb} - \mu_p - \mu_b + \mu && \text{(persoons- × beoordelaarseffect)}\\
&+ \mu_{vb} - \mu_v - \mu_b + \mu && \text{(item- × beoordelaarseffect)}\\
&+ X_{pvb} - \mu_{pv} - \mu_{pb} - \mu_{vb} + \mu_p + \mu_v + \mu_b - \mu. && \text{(residu)}
\end{aligned} \qquad (3.33)$$

In (3.33) is het algemene gemiddelde gedefinieerd als $\mu = \mathcal{E}_p \mathcal{E}_v \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over personen, vragen en beoordelaars) na beoordeling van alle antwoorden van alle personen uit de populatie op alle vragen uit het universum door alle beoordelaars uit het universum van beoordelaars. De universumscore van een persoon is gedefinieerd als $\mu_p = \mathcal{E}_v \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over items en beoordelaars) van een persoon na beoordeling van de antwoorden op alle vragen uit het universum door alle beoordelaars uit het universum. De strengheid van een beoordelaar is gedefinieerd als $\mu_b = \mathcal{E}_p \mathcal{E}_v X_{pvb}$, de gemiddelde score (= verwachting over personen en items) van een beoordelaar na beoordeling van de antwoorden op alle vragen uit het universum door alle personen uit de populatie. De parameter $\mu_{pv}$ is gedefinieerd als $\mu_{pv} = \mathcal{E}_b X_{pvb}$, de gemiddelde score (= verwachting over beoordelaars) van een persoon op een vraag na beoordeling van het antwoord door alle beoordelaars uit het universum. De definities van de parameters $\mu_v$, $\mu_{pb}$ en $\mu_{vb}$ zijn respectievelijk $\mu_v = \mathcal{E}_p \mathcal{E}_b X_{pvb}$, $\mu_{pb} = \mathcal{E}_v X_{pvb}$ en $\mu_{vb} = \mathcal{E}_p X_{pvb}$. De definities van de variantiecomponenten voor personen, vragen en beoordelaars zijn respectievelijk $\sigma_p^2 = \mathcal{E}_p(\mu_p - \mu)^2$, $\sigma_b^2 = \mathcal{E}_b(\mu_b - \mu)^2$ en $\sigma_v^2 = \mathcal{E}_v(\mu_v - \mu)^2$. Voor wat betreft de overige variantiecomponenten volstaan we met het geven van de definitie voor het persoons-× itemeffect: $\sigma_{pv}^2 = \mathcal{E}_p \mathcal{E}_v(\mu_{pv} - \mu_p - \mu_v + \mu)^2$.

De totale variantie is gelijk aan:

$$\sigma_X^2 = \sigma_p^2 + \sigma_v^2 + \sigma_b^2 + \sigma_{pv}^2 + \sigma_{pb}^2 + \sigma_{vb}^2 + \sigma_{pvb,e}^2. \qquad (3.34)$$

In het twee-facet-design met slechts een observatie voor elke combinatie van persoon, vraag en beoordelaar, bestaat de variantiecomponent voor het residu, $\sigma_{pvb,e}^2$, uit de niet te scheiden variantiecomponenten voor de interactie personen × vragen × beoordelaars en voor de fouten. Daarnaast worden er in (3.34) nog vijf andere variantiecomponenten voor mogelijke foutenbronnen onderscheiden: de twee variantiecomponenten voor de twee hoofdeffecten en de drie variantiecomponenten voor de drie eerste-orde-interactie-effecten.

De mogelijkheid om door toepassing van designs met meer facetten verschillende foutenbronnen te onderscheiden, is het belangrijkste verschil tussen de generaliseerbaarheids-theorie en de klassieke testtheorie. In voorgaande paragrafen zagen we dat in de klassieke testtheorie geen onderscheid gemaakt wordt tussen de verschillende storende factoren die de toetsscore van een persoon beïnvloeden en dat alle foutenbronnen door een enkele variantie-component gerepresenteerd worden.

### *3.13.1 Generaliseerbaarheidsstudie*

De tabellen 3.19 en 3.20 bevatten alle informatie die nodig is om een generaliseerbaarheidsstudie uit te voeren. Tabel 3.19 geeft de variantie-analysetabel van een gekruist twee-facet-design met drie random effecten. In tabel 3.20 staat hoe men de kwadratensommen kan berekenen en hoe de zeven variantiecomponenten geschat kunnen worden.

Aan de hand van het voorbeeld, ontleend aan Thorndike (1982, p. 161), in tabel 3.21 laten we zien hoe de berekening van de kwadratensommen verloopt. Daartoe dienen we de zeven parameters in (3.33) te vervangen door hun geobserveerde equivalenten. Dit resulteert in de volgende decompositie:

$$X_{pvb} = \bar{X} + (\bar{X}_p - \bar{X}) + (\bar{X}_v - \bar{X}) + (\bar{X}_b - \bar{X}) + \bar{X}_{pv\sim} + \bar{X}_{pb\sim} + \bar{X}_{vb\sim} + X_{pvb\sim} \qquad (3.35)$$

In (3.35) staat $\bar{X}_{pv\sim}$ als afkorting voor $\bar{X}_{pv} - \bar{X}_p - \bar{X}_v + X$. De betekenis van afkortingen voor de andere interactietermen staat in tabel 3.20.

Tabel 3.19

Variantie-analysetabel van een gekruist design met drie random effecten en schattingen van variantiecomponenten

| Effecten | Kwadraten-sommen | Vrijheidsgraden | Gemiddelde kwadratensommen | |
|---|---|---|---|---|
| Personen (*p*) | $SS_p$ | $df_p = n_p - 1$ | $MS_p = SS_p/df_p$ | $\hat{\sigma}^2_p = (MS_p$ |
| Items (*v*) | $SS_v$ | $df_v = n_v - 1$ | $MS_v = SS_v/df_v$ | $\hat{\sigma}^2_v = (MS_v$ |
| Beoordelaars (*b*) | $SS_b$ | $df_b = n_b - 1$ | $MS_b = SS_b/df_b$ | $\hat{\sigma}^2_b = (MS_b$ |
| Personen x items (*pv*) | $SS_{pv}$ | $df_{pv} = (n_p - 1)(n_v - 1)$ | $MS_{pv} = SS_{pv}/df_{pv}$ | $\hat{\sigma}^2_{pv} = (MS_{pv}$ |
| Personen x beoordelaars (*pb*) | $SS_{pb}$ | $df_{pb} = (n_p - 1)(n_b - 1)$ | $MS_{pb} = SS_{pb}/df_{pb}$ | $\hat{\sigma}^2_{pb} = (MS_{pb}$ |
| Items x beoordelaars (*vb*) | $SS_{vb}$ | $df_{vb} = (n_v - 1)(n_b - 1)$ | $MS_{vb} = SS_{vb}/df_{vb}$ | $\hat{\sigma}^2_{vb} = (MS_{vb}$ |
| Residu (*pvb,e*) | $SS_{pvb,e}$ | $df_{pvb,e} = (n_p - 1)(n_v - 1)(n_b - 1)$ | $MS_{pvb,e} = SS_{pvb,e}/df_{pvb,e}$ | $\hat{\sigma}^2_{pvb,e} = MS_{pvb,e}$ |

Tabel 3.20

Definities van kwadratensommen en schattingen van variantiecomponenten

| | | |
|---|---|---|
| $SS_p$ | $= n_v n_b \sum_p (\bar{X}_p - \bar{X})^2$ | $= MS_{pvb,e}$ |
| $SS_v$ | $= n_p n_b \sum_v (\bar{X}_v - \bar{X})^2$ | $(MS_{vb} - MS_{pvb,e})/n_p$ |
| $SS_b$ | $= n_p n_v \sum_b (\bar{X}_b - \bar{X})^2$ | $(MS_{pb} - MS_{pvb,e})/n_v$ |

$$SS_{pv} = n_b \sum_p \sum_v (\bar{X}_{pv\sim})^2 = n_b \sum_p \sum_v (\bar{X}_{pv} - \bar{X}_p - \bar{X}_v + \bar{X})^2 \qquad \hat{\sigma}^2_{pv} = (MS_{pv} - MS_{pvb,e})/n_b$$

$$SS_{pb} = n_v \sum_p \sum_b (\bar{X}_{pb\sim})^2 = n_v \sum_p \sum_b (\bar{X}_{pb} - \bar{X}_p - \bar{X}_b + \bar{X})^2 \qquad \hat{\sigma}^2_b = (MS_b - MS_{vb} - MS_{pb} + MS_{pvb,e})/(n_p n_v)$$

$$SS_{vb} = n_p \sum_v \sum_b (\bar{X}_{vb\sim})^2 = n_p \sum_v \sum_b (\bar{X}_{vb} - \bar{X}_v - \bar{X}_b + \bar{X})^2 \qquad \hat{\sigma}^2_v = (MS_v - MS_{vb} - MS_{pv} + MS_{pvb,e})/(n_p n_b)$$

$$SS_{pvb,e} = \sum_p \sum_v \sum_b (X_{pvb\sim})^2 = \sum_p \sum_v \sum_b (X_{pvb} - \bar{X}_{pv} - \bar{X}_{pb} - \bar{X}_{vb} + \bar{X}_p + \bar{X}_v + \bar{X}_b - \bar{X})^2 \qquad \hat{\sigma}^2_p = (MS_p - MS_{pb} - MS_{pv} + MS_{pvb,e})/(n_v n_b)$$

$$SS_{tot} = \sum_p \sum_v \sum_b (X_{pvb} - \bar{X})^2$$

Tabel 3.21

De itemscores van zes personen op vier items en twee beoordelaars, per beoordelaar de gemiddelde score per item en per persoon, de gemiddelde score per beoordelaar, de gemiddelde score van elke persoon en het algemene gemiddelde

| | Beoordelaar 1 | | | | | Beoordelaar 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pers. | Item: 1 | 2 | 3 | 4 | Gem. | Item: 1 | 2 | 3 | 4 | Gem. | $\bar{X}_p$ |
| 1 | 9 | 6 | 6 | 2 | 5.75 | 8 | 2 | 8 | 1 | 4.75 | 5.25 |
| 2 | 9 | 5 | 4 | 0 | 4.50 | 7 | 5 | 9 | 5 | 6.50 | 5.50 |
| 3 | 8 | 9 | 5 | 8 | 7.50 | 10 | 6 | 9 | 10 | 8.75 | 8.13 |
| 4 | 7 | 6 | 5 | 4 | 5.40 | 9 | 8 | 9 | 4 | 7.70 | 6.50 |
| 5 | 7 | 3 | 2 | 3 | 3.75 | 7 | 4 | 5 | 1 | 4.25 | 4.00 |
| 6 | 10 | 8 | 7 | 7 | 8.00 | 7 | 7 | 10 | 9 | 8.25 | 8.13 |
| Gem. | 8.33 | 6.17 | 4.83 | 4.00 | 5.83 | 8.00 | 5.33 | 8.33 | 5.00 | 6.67 | $\bar{X}$ = 6.25 |

Tabel 3.21 bevat de itemscores die twee beoordelaars aan de antwoorden op vier items aan zes personen toegekend hebben. Voor persoon 1 uit deze tabel hebben we (3.35) uitgeschreven in tabel 3.22.

Tabel 3.22

Vergelijking (3.35) uitgeschreven voor persoon 1 uit tabel 3.21

| $X_{pvb}$ = | $\bar{X}$ | + $(\bar{X}_p - \bar{X})$ | + $(\bar{X}_v - \bar{X})$ | + $(\bar{X}_b - \bar{X})$ | + $\bar{X}_{pv}$~ | + $\bar{X}_{pb}$~ | + $\bar{X}_{vb}$~ | + $X_{pvb}$~ |
|---|---|---|---|---|---|---|---|---|
| $X_{111}$ = 9 = | 6.25 | — 1.00 | + 1.92 | — 0.42 | + 1.33 | + 0.92 | + 0.58 | — 0.58 |
| $X_{112}$ = 8 = | 6.25 | — 1.00 | + 1.92 | + 0.42 | + 1.33 | — 0.92 | — 0.58 | + 0.58 |
| $X_{121}$ = 6 = | 6.25 | — 1.00 | — 0.50 | — 0.42 | — 0.75 | + 0.92 | + 0.83 | + 0.67 |
| $X_{122}$ = 2 = | 6.25 | — 1.00 | — 0.50 | + 0.42 | — 0.75 | — 0.92 | — 0.83 | — 0.67 |
| $X_{131}$ = 6 = | 6.25 | — 1.00 | + 0.33 | — 0.42 | + 1.42 | + 0.92 | — 1.33 | — 0.17 |
| $X_{132}$ = 8 = | 6.25 | — 1.00 | + 0.33 | + 0.42 | + 1.42 | — 0.92 | + 1.33 | + 0.17 |
| $X_{141}$ = 2 = | 6.25 | — 1.00 | — 1.75 | — 0.42 | — 2.00 | + 0.92 | — 0.08 | + 0.08 |
| $X_{142}$ = 1 = | 6.25 | — 1.00 | — 1.75 | + 0.42 | — 2.00 | — 0.92 | + 0.08 | — 0.08 |

Voor het berekenen van de kwadratensommen moeten we vergelijking (3.35) ook nog uitschrijven voor de vijf andere personen, wat een uitbreiding betekent van tabel 3.22 met de decomposities van veertig itemscores. De zeven kwadratensommen worden verkregen door de getallen in de desbetreffende kolommen van tabel 3.22 te kwadrateren en te sommeren. Beschikken we over de kwadratensommen, dan kunnen we schattingen van de variantie-componenten eenvoudig berekenen met behulp van tabel 3.20. Wellicht ten overvloede merken we op dat de standaardfouten van variantiecomponenten bij kleine aantallen personen en condities zeer groot zijn (Brennan, 1992, p. 104). De steekproef uit de populatie moet uit minstens honderd personen bestaan teneinde acceptabele standaardfouten te verkrijgen (Smith, 1978). De resultaten van de generaliseerbaarheidsstudie voor het voorbeeld staan vermeld in tabel 3.23.

Tabel 3.23

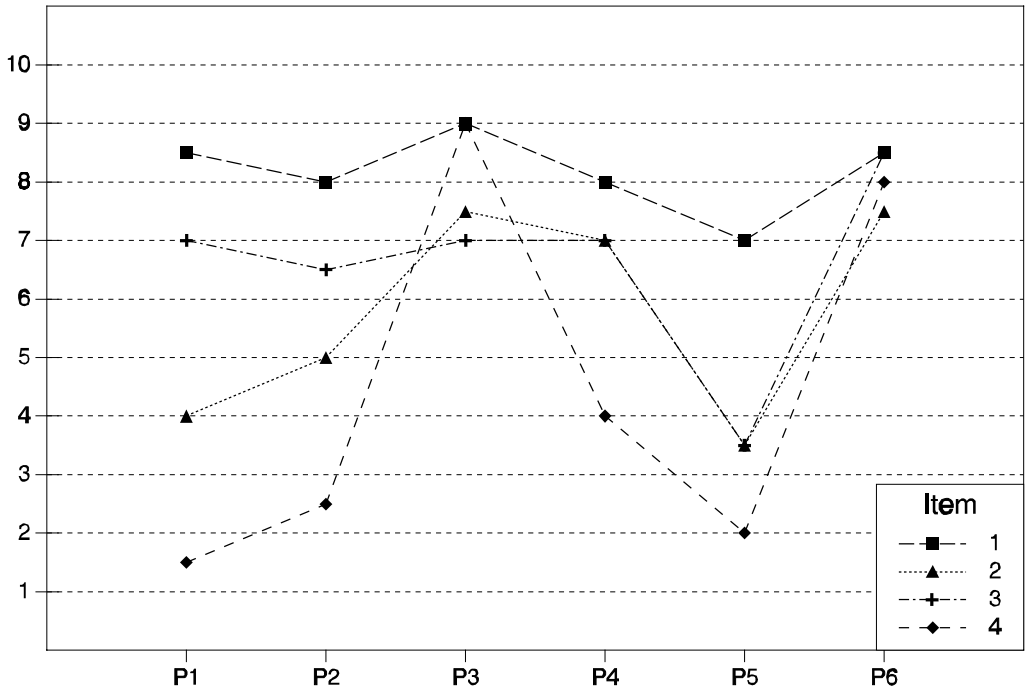
Resultaten generaliseerbaarheidsstudie voor data uit tabel 3.21

| Effecten | Kwadraten-sommen | Vrijheids-graden | Gemiddelde kwadraten-sommen | Schattingen van variantie-componenten | |
|---|---|---|---|---|---|
| Personen *(p)* | 109.75 | 5 | 21.95 | $\hat{\sigma}^2_p$ | = 2.16 (28%) |
| Items *(v)* | 85.17 | 3 | 28.39 | $\hat{\sigma}^2_v$ | = 1.26 (15%) |
| Beoordelaars *(b)* | 8.33 | 1 | 8.33 | $\hat{\sigma}^2_b$ | = –0.15 (0%) |

| | | | | |
|---|---|---|---|---|
| Personen × items *(pv)* | 59.08 | 15 | 3.94 | $\hat{\sigma}^2_{pv}$ = 0.98 (12%) |
| Personen × beoordelaars *(pb)* | 13.42 | 5 | 2.68 | $\hat{\sigma}^2_{pb}$ = 0.18 (2%) |
| Items × beoordelaars *(vb)* | 33.83 | 3 | 11.28 | $\hat{\sigma}^2_{vb}$ = 1.55 (19%) |
| Residu *(pvb,e)* | 29.42 | 15 | 1.96 | $\hat{\sigma}^2_{pvb,e}$ = 1.96 (24%) |

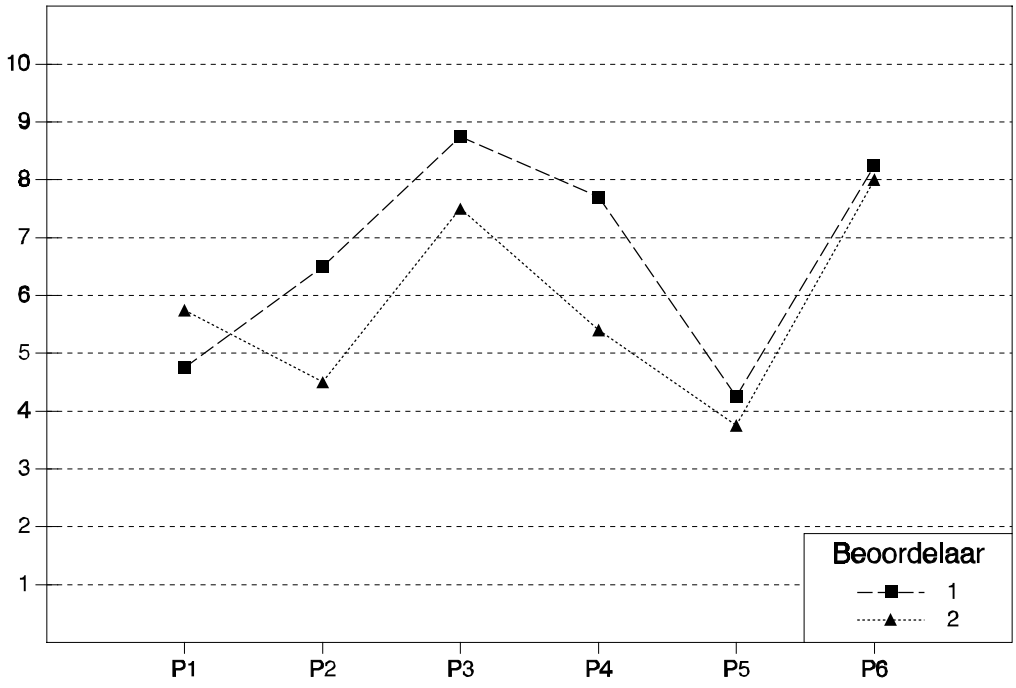
De laatste kolom van tabel 3.23 bevat de schattingen van de variantiecomponenten en hun procentuele bijdrage aan de totale variantie. We zien dat de variantiecomponent van de beoordelaars negatief is. Hoewel in theorie variantiecomponenten niet negatief kunnen zijn, kunnen schattingen van variantiecomponenten wel negatief zijn. Negatieve schattingen hebben veelal twee mogelijke oorzaken. Relatief grote negatieve componenten zijn meestal het gevolg van het gebruik van het verkeerde model. Een relatief grote negatieve component van beoordelaars had er in ons voorbeeld op kunnen wijzen dat het lineaire model in (3.33) niet het juiste model was om de data te analyseren. Relatief kleine negatieve componenten zijn meestal het gevolg van het gebruik van een te kleine steekproef. Dit laatste is waarschijnlijk de oorzaak van de negatieve component in ons voorbeeld. Aangezien negatieve componenten niet mogelijk zijn, worden negatieve schattingen vervangen door nul. Merk op dat er andere schattingsmethoden voor variantiecomponenten zijn die niet leiden tot negatieve schattingen. Een daarvan is de restrictieve grootste-aannemelijkheidschattingsmethode. De relatief grote bijdrage van de variantiecomponent voor de items is met name het gevolg van het grote verschil in moeilijkheidsgraad tussen item 1 en item 4. De gemiddelde itemscore van item 1 is 8.17, terwijl die van item 4 gelijk is aan 4.50.

De bijdrage van de interactiecomponent personen × items is veel groter dan die van de interactiecomponent personen × beoordelaars. Interactie tussen personen en items betekent dat personen niet consistent antwoorden op de verschillende items. Interactie tussen personen en beoordelaars houdt in dat personen niet consistent beoordeeld worden door verschillende beoordelaars. In figuur 3.3. hebben we de interactie personen × items grafisch gepresenteerd.




Figuur 3.3 Interactie personen × items

In figuur 3.3 is voor elk item een lijn getrokken die de gemiddelde itemscores, $\bar{X}_{pv}$, van personen, P1-P6, met elkaar verbindt. We zien dat de vier lijnen elkaar bij verschillende personen kruisen, wat betekent dat het niet dezelfde persoon is die de hoogste of laagste score op elk item behaalt. Lijnen die elkaar kruisen wijzen er op dat er sprake is van interactie. Merk op dat in tabel 3.22 de berekening van de variantiecomponent voor de interactie tussen personen en items gebaseerd is op $\bar{X}_{pv}\sim = \bar{X}_{pv} - \bar{X}_{p} - \bar{X}_{v} + \bar{X}$. We hadden de interactie tussen personen en items ook met behulp van $X_{pv}\sim$ in plaats van $\bar{X}_{pv}$ kunnen afbeelden. Wanneer de vier lijnen parallel lopen is, de kwadratensom personen × items, en dus ook de variantiecomponent, gelijk aan nul.




Figuur 3.4 Interactie personen × beoordelaars

Om mogelijke interactie tussen personen en beoordelaars te onderzoeken, is in figuur 3.4 voor elk item een lijn getrokken die de gemiddelde beoordelaarsscores, $\bar{X}_{pb}$, van personen met elkaar verbindt. We zien dat de twee lijnen elkaar bij de eerste persoon kruisen maar bij de andere vijf personen nagenoeg parallel lopen. Dit betekent dat de twee beoordelaars de eerste persoon niet, maar de vijf andere personen wel op dezelfde wijze onderscheiden. De variantiecomponent voor de interactie tussen personen en beoordelaars blijkt dan ook gering te zijn.

De interactie items × beoordelaars is de grootste eerste-orde-interactie, met name veroorzaakt door de derde vraag. Die vraag heeft van de eerste beoordelaar een lage beoordeling, gemiddelde score 4.83, en van de tweede beoordelaar een hoge beoordeling, gemiddelde score 8.33, ontvangen.

### *3.13.2 Decisiestudie*

In ons voorbeeld bestaat de toets uit vier random getrokken items uit het universum van items en twee random getrokken beoordelaars uit het universum van beoordelaars die de antwoorden op de items beoordelen. Een andere toets met vier random getrokken items en twee random getrokken beoordelaars zou ook acceptabel geweest

zijn. Het universum waar in dit geval naar gegeneraliseerd wordt, is het universum van random parallelle toetsen met vier items en twee beoordelaars.

De schatting van de generaliseerbaarheidscoëfficiënt voor relatieve beslissingen is voor het gekruiste twee-facet-random-effect-design gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pv}}{n_v} + \frac{\hat{\sigma}^2_{pb}}{n_b} + \frac{\hat{\sigma}^2_{pvb,e}}{n_v n_b}}. \qquad (3.36)$$

Naast de universumscorevariantie, bevat de noemer van (3.36) drie variantiecomponenten die interacties met personen betreffen. Hiervoor zagen we dat een relatief grote variantie- component voor de interactie tussen personen en items inhoudt dat bijvoorbeeld Jan niet op ieder item meer presteert dan Piet. Het maakt voor het nemen van relatieve beslissingen dan ook wel degelijk uit welke items aan welke personen voorgelegd worden. Een bepaald item wordt namelijk door Jan als gemakkelijk en door Piet als moeilijk opgevat, terwijl bij een ander item het omgekeerde het geval is. De variantiecomponent voor de interactie tussen personen en items dient dan ook beschouwd te worden als foutenvariantie. Ook de variantiecomponent voor de interactie tussen personen en beoordelaars, dat wil zeggen dat het van de beoordelaar afhangt of Jan beter is dan Piet, dient als foutenvariantie beschouwd te worden. De variantiecomponent voor het residu is per definitie foutenvariantie. Voor de toets uit ons voorbeeld is de generaliseerbaarheidscoëfficiënt gelijk aan: 2.16/{2.16 + 0.99/4 + 0.18/2 + 1.96/8} = .79.

De schatting van de generaliseerbaarheidscoëfficiënt voor absolute beslissingen is voor het gekruiste twee-facet-random-effect design gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_v}{n_v} + \frac{\hat{\sigma}^2_b}{n_b} + \frac{\hat{\sigma}^2_{pv}}{n_v} + \frac{\hat{\sigma}^2_{pb}}{n_b} + \frac{\hat{\sigma}^2_{pvb,e}}{n_v n_b}}. \qquad (3.37)$$

Bij het nemen van absolute beslissingen maakt het niet alleen uit of er makkelijke of moeilijke vragen aan de personen voorgelegd worden, maar ook of die vragen door milde of strenge beoordelaars beoordeeld worden. Vandaar dat in (3.37) naast de variantiecomponenten voor de drie interacties ook de variantiecomponenten voor de items en voor de beoordelaars beschouwd worden als foutenvariantie. De generaliseerbaarheidscoëfficiënt voor absolute beslissingen is gelijk aan 2.16/{2.16 + 1.26/4 + 0.0/2 + 0.99/4 + 0.18/2 + 1.96/8} = .71 voor de toets uit ons voorbeeld.

## 3.14 Andere aspecten van de generaliseerbaarheidstheorie

Formule (3.36) laat zien dat we de generaliseerbaarheidscoëfficiënt kunnen verhogen door de toets te verlengen, wat neerkomt op het vergroten van het aantal items of het aantal beoordelaars. Voor het realiseren van dezelfde generaliseerbaarheidscoëfficiënt hebben we meer condities nodig van een facet met een relatief grote variantiecomponent die bijdraagt aan de foutenvariantie, dan condities van een facet met een relatief kleine variantiecomponent. We verwijzen naar hoofdstuk 11 voor een bespreking van toetsverlenging bij designs met meer facetten.

De generaliseerbaarheidscoëfficiënt kan ook verhoogd worden door een random facet op te vatten als een fixed facet. Dat een facet fixed is, wil zeggen dat een toets alle condities van een facet bevat. Beschouwen we in ons voorbeeld de items als fixed facet, dan generaliseren we niet meer naar het universum van random parallelle toetsen met vier items en twee beoordelaars, maar naar het universum van random parallelle toetsen met twee beoordelaars. Het spreekt vanzelf dat door het beperken van het universum waar naar gegeneraliseerd wordt, de beslissingen over personen nauwkeuriger kunnen zijn. Voor een bespreking van designs met fixed facets verwijzen we naar Shavelson en Webb (1991, pp. 65-82).

De bespreking in voorgaande paragrafen heeft zich beperkt tot gekruiste designs met een enkel facet en met twee facetten. Binnen de generaliseerbaarheidstheorie kunnen echter ook designs met meer dan twee facetten geanalyseerd worden. Daarnaast kunnen ook zogenaamde genestelde designs geanalyseerd worden. Ons voorbeeld met twee facetten zou een genesteld design zijn wanneer de eerste en de tweede vraag door de eerste beoordelaar beoordeeld worden en de derde en vierde vraag door de tweede beoordelaar. In dat geval zeggen we dat de vragen genesteld zijn binnen de beoordelaars. Genestelde designs komen vooral voor bij niet-experimenteel onderzoek (Feldt & Brennan, 1989). In het algemeen heeft het gebruik van gekruiste designs de voorkeur, omdat het met de resultaten van de generaliseerbaarheidsstudie van gekruiste designs mogelijk is na te gaan hoe de resultaten voor een genesteld design geweest zouden zijn. Het omgekeerde is niet het geval.

In de voorbeelden die tot nu toe besproken zijn, hadden de beslissingen steeds betrekking op personen. In veel onderzoek, met name onderzoek op het gebied van het onderwijs, zijn we echter niet of niet uitsluitend geïnteresseerd in (verschillen tussen) personen maar ook in klassen, leerdoelen of andere meetobjecten. Om aan te geven dat elk facet uit een design het meetobject kan zijn, introduceerden Cardinet, Tourneur en Allal (1981) het zogenaamde symmetrieprincipe. Uitgaande van dat principe laten zij

zien hoe binnen het kader van de generaliseerbaarheidstheorie een grote verscheidenheid aan onderzoeksvragen beantwoord kan worden.

De meest gebruikte schatting van de universumscore van een persoon is de geobserveerde gemiddelde score van een persoon. In Cronbach e.a. (1972) worden echter ook varianten van Kelley's formule (zie paragraaf 3.5) voor schattingen van universumscores besproken. Hoe schattingen van universumscores verkregen kunnen worden met behulp van lineaire predictiefuncties wordt beschreven door Jarjoura (1983).

Tenslotte dient opgemerkt te worden dat met de generaliseerbaarheidstheorie niet alleen univariate maar ook multivariate modellen, dat wil zeggen modellen waarbij de personen een aantal universumscores hebben, geanalyseerd kunnen worden. Voor een bespreking van modellen uit de multivariate generaliseerbaarheidstheorie verwijzen we naar Cronbach e.a. (1972), Shavelson en Webb (1981) en Brennan (1992).