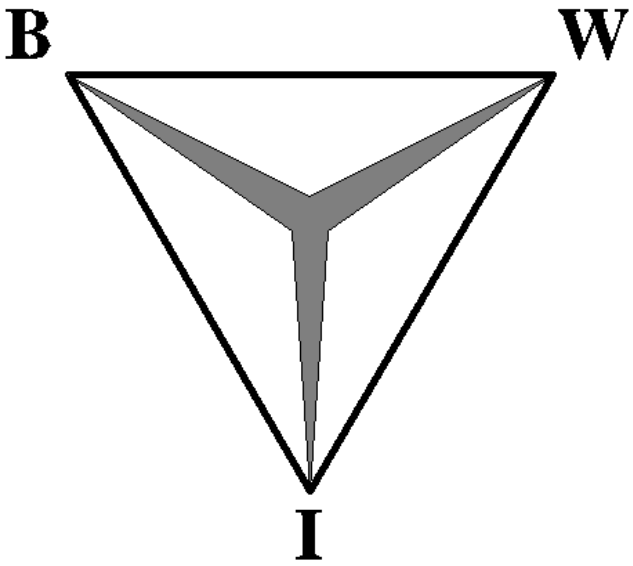
Het Rasch-model in de itemresponstheorie



Denise B. Kamerbeek
februari 2003

BWI-werkstuk

Het Rasch-model in de itemresponstheorie

Denise B. Kamerbeek
BWI-werkstuk

vrije Universiteit
Faculteit der Exacte Wetenschappen
De Boelelaan 1081a
1081 HV Amsterdam

februari 2003

Voorwoord

Als onderdeel van de studie Bedrijfswiskunde en Informatica aan de Vrije Universiteit in Amsterdam moet elke student een werkstuk maken. In dit BWI-werkstuk moeten de drie disciplines Bedrijfskunde, Wiskunde en Informatica aan bod komen.
Het onderwerp van mijn BWI-werkstuk is het Rasch-model in de itemresponstheorie.
Deze testtheorie houdt zich op een bepaalde manier bezig met het construeren van meetinstrumenten zoals toetsen en examens. Het Rasch-model is een speciaal geval van de itemresponstheorie, waarmee veel gewerkt wordt in de praktijk.

Doordat zowel mijn begeleider, Geurt Jongbloed, als ik totaal onbekend waren met het onderwerp is het een spannende uitdaging gebleken om meer te weten te komen over het onderwerp en het ook toe te passen. Ik wil hem graag bedanken voor zijn begeleiding en voor de stimulerende en motiverende besprekingen. Verder wil ik Sylvia van Os en Ivette Berends bedanken voor de informatie en het materiaal die zij vanuit hun studie (toegepaste) onderwijskunde hebben geleverd.

Denise B. Kamerbeek

Amsterdam, februari 2003

Samenvatting

In de psychometrie houdt men zich bezig met het construeren van meetinstrumenten zoals toetsen en examens. Er zijn diverse testtheorieën ontwikkeld, waarvan de klassieke testtheorie en de itemresponstheorie het bekendst zijn. Het Rasch-model, dat behoort tot de laatstgenoemde theorie, is onderwerp van dit BWI-werkstuk

In de itemresponstheorie staat de term 'vaardigheid' centraal. De functie geeft aan hoe groot de kans is dat een item juist wordt beantwoord als functie van de vaardigheid van de respondent. Het Rasch-model houdt in de itemresponsfunctie ook nog rekening met de moeilijk van de items. Er zijn dus twee soorten parameters waarvan de functie afhangt: de vaardigheidsparameters en de itemparameters. Om het model toe te passen is het noodzakelijk dat de parameters waarden krijgen. Een manier om deze parameters te schatten is gebruik te maken van de grootste aannemelijkheidsschatter (Maximum Likelihood). Er zijn diverse varianten op deze methode ontwikkeld die goede schattingen als resultaat hebben.

Voor een eenvoudige situatie is er in dit werkstuk een Rasch-model ontwikkeld. De parameters zijn volgens de Joint Maximum Likelihood-methode geschat. Hierbij worden de vaardigheidsparameters en de itemparameters tegelijkertijd geschat. De parameters zijn zowel numeriek benaderd als exact berekend. Ook is er gekeken naar de betrouwbaarheid van één van de parameters door een betrouwbaarheidsinterval te construeren. Hierbij is gebruik gemaakt van de bootstraptechniek. Als laatste is er voor een specifiek geval getoetst of twee vragen in de toets even moeilijk zijn.

Bij het opstellen van het model zijn er uiteindelijk hele eenvoudige formules ontstaan om de parameters te schatten. In de praktijk wordt echter gebruik gemaakt van computerprogramma's omdat de complexiteit van de formules waarmee de parameters geschat worden toeneemt met de omvang van de toets.
Om te toetsen of twee items even moeilijk zijn kan men op twee manieren te werk gaan. Er kan met behulp van een tweesteekproef binomiale toets getoetst worden, maar er kan ook getoetst worden door een betrouwbaarheidsinterval te construeren voor een itemparameter op basis van het Rasch-model.

Er blijven na dit werkstuk genoeg zaken over die nog onderzocht kunnen worden. Zo kan er verdere verdieping in de hier behandelde theorieën plaatsvinden, maar er kan ook onderzoek worden gedaan naar andere theorieën. Het Rasch-model wat in dit werkstuk is opgesteld zal ook nog verder uitgebreid kunnen worden voor een grotere dataset. Tevens kan er gekeken worden naar andere methoden om de parameters in het model te schatten. Een laatste aanbeveling voor verder onderzoek is op het gebied van het soort items waarmee in een toets wordt gewerkt. In dit werkstuk is er alleen gekeken naar dichotome items (goed/fout), maar er zijn natuurlijk ook toetsen met open vragen.

Inhoudsopgave

1 Inleiding

Testen zijn heel populair in het dagelijks leven. Tegenwoordig maakt een psychologische test standaard deel uit van een sollicitatieprocedure. In het onderwijs wordt de test al tijden toegepast om de kennis van de leerling te toetsen. Deze laatste toepassing bestaat al lang en er is al veel onderzoek over gedaan, zodat er een speciaal vakgebied is ontstaan dat zich ermee bezighoudt, namelijk de psychometrie. Dit onderdeel van de psychologie houdt zich bezig met het construeren van meetinstrumenten voor onderwijskundige doeleinden.

Er bestaan verschillende testtheorieën, maar in de praktijk wordt er het meest gebruik gemaakt van de klassieke testtheorie en de itemresponstheorie. De eerste theorie houdt zich voornamelijk bezig met de betrouwbaarheid en de standaardfout van de test, terwijl de tweede theorie meer aandacht heeft voor de niet zichtbare kenmerken van de respondent.

In de itemresponstheorie wordt er gewerkt met een itemresponsfunctie. Deze geeft aan hoe groot de kans is dat het item juist wordt beantwoord als functie van de vaardigheid van een bepaald persoon. Een speciaal geval van de itemresponstheorie is het Rasch-model. De itemresponsfunctie is in dit geval afhankelijk van twee soorten parameters, namelijk de itemparameters en de vaardigheidsparameters. Er zijn diverse methoden bekend waarmee deze parameters in het Rasch-model geschat kunnen worden. De bekendsten zijn gebaseerd op de methode van de meest aannemelijke schatter (ML): die waarde van de parameter die de grootste waarschijnlijkheid toekent aan de waargenomen waarde x.

Na het bestuderen van de literatuur over de testtheorieën, de itemresponstheorie en het Rasch-model ontstond de uitdaging om zelf een model op te stellen. Bij het ontwikkelen van een Rasch-model wordt normaliter gebruik gemaakt van computerprogramma's, omdat het rekenwerk voor het schatten van de parameter zeer intensief is. Daarom is in dit werkstuk getracht om voor een eenvoudige situatie een model op te stellen en uitspraken te doen omtrent betrouwbaarheid en interpretatie.

In hoofdstuk 2 staat een korte uitleg over de veel gebruikte testtheorieën, om vervolgens in hoofdstuk 3 over te gaan op het Rasch-model. Daarna is er een hoofdstuk (4) gewijd aan het opstellen van een model op basis van praktijkgegevens. Verder is er nog een toets uitgevoerd en is er gekeken naar de betrouwbaarheid van een van de schatters. De conclusies en aanbevelingen voor verder onderzoek zijn beschreven in hoofdstuk 5.

2 Testtheorieën

2.1 Algemeen

Tests worden in het dagelijks leven regelmatig gebruikt om metingen te doen. Dit kan bijvoorbeeld in de vorm van een psychologische test voor de selectie van sollicitanten, maar ook als vorderingentest in het onderwijs. Dit zijn toch twee verschillende soorten tests met twee verschillende doelen. De psychometrie is een onderdeel van de psychologie die zich bezighoudt met de *educational* tests. Het vakgebied psychometrie kan ook wel omschreven worden als de theorie met betrekking tot het construeren van meetinstrumenten zoals toetsen en examens (2002).
Hieronder volgt een heldere definitie van het begrip 'test':

> Een test is een systematische meetprocedure waarmee getracht wordt zo efficiënt, betrouwbaar en valide mogelijk een bepaalde karakteristiek (attribuut, trek) van een persoon of een groep personen te meten (vaardigheid, attitude, geschiktheid, enzovoort) door uit te gaan van een objectieve verwerking van de reacties van de persoon (in vergelijking met die van anderen) op een aantal gestandaardiseerde zorgvuldig gekozen stimuli (meerdere items). (Wierstra, 2000)

Er zijn twee testtheorieën die heel bekend zijn en regelmatig worden gebruikt. Dit is allereerst de klassieke testtheorie, die vooral draait om de begrippen betrouwbaarheid en standaardmeetfout. Later is ook de latente trek-theorie of de itemresponstheorie ontwikkeld, die wat eleganter is en praktische problemen beter aankan. Deze theorie draait meer om de vaardigheid van de respondent. Hieronder volgt voor elke testtheorie een korte toelichting waarbij ook de verschillen tussen de theorieën worden behandeld.
Onderstaande paragrafen zijn beiden gebaseerd op Eggen (1993) en Verhelst (1992).

2.2 Klassieke testtheorie

Indien bij een persoon meerdere malen een toets wordt afgenomen, zal niemand verwachten dat het resultaat elke keer precies hetzelfde zal zijn, ook niet in gevallen waar geheugen- of leereffecten geen enkele rol spelen. Men houdt er dus rekening mee dat de toetsscore (X) niet perfect de kennis of vaardigheid van de geteste persoon weerspiegelt, of met andere woorden dat er een meetfout (E) is gemaakt. In de klassieke testtheorie is men geïnteresseerd in de 'ware' score (T), en men hanteert als basisvergelijking

$$X = T + E \tag{1}$$

waarbij men eist dat de gemiddelde of verwachte fout gelijk is aan nul (zodat de test betrouwbaar is).

De ware score zou men in principe kunnen achterhalen door de toets een zeer groot aantal keren bij dezelfde persoon af te nemen onder exact dezelfde condities en van al de geobserveerde scores het gemiddelde te bepalen. Zo'n procedure is in de praktijk natuurlijk niet uitvoerbaar. In de klassieke testtheorie heeft men slimme manieren gevonden om met andere, wel uitvoerbare procedures, iets zinnigs te kunnen zeggen over de (gemiddelde) grootte van de meetfout (hier wordt niet op ingegaan).

2.3 Itemresponstheorie

In de itemresponstheorie (IRT) speelt het begrip 'ware toetsscore' een zeer ondergeschikte rol. Centraal in de IRT staat een abstract begrip dat men zou kunnen aanduiden als 'vaardigheid'. Men gaat ervan uit dat personen verschillen in vaardigheid en dat het op één of andere manier mogelijk moet zijn aan elke persoon een getal toe te kennen dat zijn vaardigheid adequaat weerspiegelt. Bovendien is men ervan overtuigd dat de hoeveelheid vaardigheid van een persoon niet direct waarneembaar is (latent).
De theorie is ontwikkeld zonder enige referentie aan een of andere populatie. Bovendien staat niet de toetsscore centraal, maar het item en het antwoord op het item. Dit verklaart ook de naam van deze theorie: itemresponstheorie oftewel latente-trektheorie (latent trait theory).

Een IRT is dus een geheel van uitspraken over de samenhang tussen één bepaald kenmerk (latent) van een persoon en zijn/haar antwoordgedrag op een verzameling items.
De uitspraken in zo'n theorie zijn meestal niet heel specifiek: de voorspellingen over het gedrag hangen af van kenmerken van de items en van de personen.

Status latente variabele: Aan elke persoon kan een getal worden toegevoegd dat een uitdrukking is van de mate waarin die persoon over de vaardigheid beschikt : θ.
De getalswaarde die aan persoon *v* is toegekend, wordt aangeduid als $\theta_{v,}$ waarbij $-\infty \leq \theta \leq \infty$.

X_i is het antwoord op item *i* (aanname: X_i is dichotoom)

$$X_i = \begin{cases} 1 & \text{indien het antwoord op item i correct is,} \\ 0 & \text{indien het antwoord op item i fout is.} \end{cases}$$

Per item wordt het antwoordproces beschouwd als een kansexperiment (stochastisch experiment). Namelijk: de kans dat een bepaald persoon een 'goed' antwoord geeft (=score 1 op het item), is afhankelijk van een *persoonsparameter* (de 'vaardigheid') en een *itemmoeilijkheidsparameter*. Hoe moeilijker het item, hoe kleiner de kans om het item goed te hebben. Hoe vaardiger de persoon hoe groter (in ieder geval niet kleiner) de kans om het item goed te hebben.

De itemresponsfunctie $f_i(\theta)$ drukt uit hoe groot de kans is dat het item juist wordt beantwoord als functie van de vaardigheid (de conditionele kans op een juist antwoord gegeven de waarde van θ) :

$$f_i(\theta) = P(X_i = 1|\theta) . \quad (2)$$

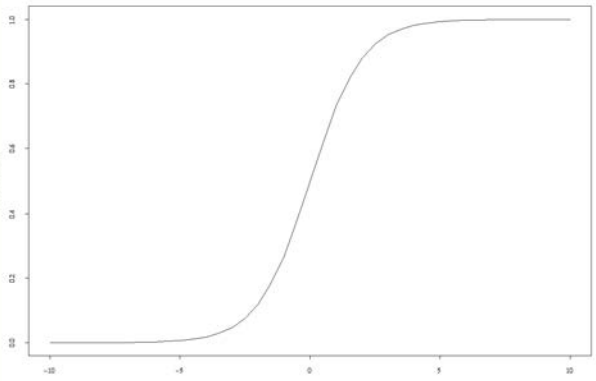
Voor de itemresponsfunctie gelden de volgende eisen:

(1) $0 \leq f_i(\theta) \leq 1$;

(2) de functie is strikt stijgend (de kans op een juist antwoord wordt nooit kleiner als de vaardigheid toeneemt);

(3) de functie moet een vloeiend verloop hebben, of exacter uitgedrukt: de functie moet overal differentieerbaar zijn.

Een voorbeeld van een itemresponsfunctie staat in figuur 1.



Figuur1. Itemresponsfunctie

Er blijven nog veel functies over die aan deze eisen voldoen. Door één specifieke functie te kiezen perkt men de theorie verder in tot één speciaal geval: een IRT model.

Enkele bekende IRT modellen:

- Rasch-model
- Lineair-logistische testmodel
- Unidimensionale modellen voor dichotome items
- Unidimensionale modellen voor polytome items
- Multidimensionale IRT-modellen

In het volgende hoofdstuk zal nader ingegaan worden op het Rasch-model.

3 Rasch-model

3.1 Algemeen

Dit hoofdstuk is gebaseerd op Eggen (1993).

Een eenvoudig IRT model dat in de literatuur veel aandacht heeft gekregen is het Rasch-model. Het werd in 1960 voorgesteld door de Deense statisticus G. Rasch.

In het Rasch-model is de itemresponsfunctie een logistische functie:

$$f(y) = \frac{\exp(y)}{1 + \exp(y)} \tag{3}$$

Een eenvoudig functieonderzoek levert op dat de logistische functie *f(y)* altijd tussen 0 en 1 ligt (teller is steeds positief en de noemer is groter dan de teller).
Verder geldt dat

$$f(0) = 0.5$$
$$\lim_{y \to \infty} f(y) = 1,$$
$$\lim_{y \to -\infty} f(y) = 0.$$

Het argument in de itemresponsfunctie in het Rasch-model is het verschil (θ - β_i), waarbij β_i een kengetal is dat item *i* karakteriseert (moeilijkheid van item *i*). Dus de itemresponsfunctie kan ook geschreven worden als

$$f_i(\theta - \beta_i) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}. \tag{4}$$

Omdat er in het Rasch-model met elk item slechts één parameter gemoeid is, wordt β_i ook vaak kortweg de itemparameter genoemd. Indien de vaardigheid precies gelijk is aan het getal β_i dan is de kans op een juist antwoord precies 0.5.

Verder geldt dat voor zeer kleine waarden van θ de kans bijna 0 is dat een correct antwoord wordt gegeven. Dat betekent dat het Rasch-model eigenlijk ongeschikt is voor items waarvan het juiste antwoord door raden tot stand komt. Dit betekent dat extra voorzichtigheid geboden is wanneer het Rasch-model wordt toegepast bij meerkeuze-items.

Formule (4) beschrijft het gedrag van iemand met vaardigheid θ op één item. Er moet echter ook nog iets gezegd worden over het gedrag, indien meer items moeten worden beantwoord. Hiervoor wordt er gebruik gemaakt van het axioma der lokale stochastische onafhankelijkheid.

> ***Axioma der lokale stochastische onafhankelijkheid***
> De kans op een juist antwoord hangt alleen af van de vaardigheid en de moeilijkheid van het item, dus als het gaat om items met dezelfde moeilijkheid en om personen met dezelfde vaardigheid moeten die kansen gelijk zijn.

$$P(X_i = 1 | \theta \text{ en } X_j = 1) = P(X_i = 1 | \theta) = f_i(\theta) \quad (5)$$

of

$$P(X_i = 1 \text{ en } X_j = 1 | \theta) = P(X_i = 1 | \theta) P(X_j = 1 | \theta) = f_i(\theta) f_j(\theta) \quad (6)$$

De beperking 'lokaal' wijst erop dat X_i en X_j alleen onafhankelijk zijn bij gelijke θ.
Daaruit volgt niet dat X_i en X_j onafhankelijk zijn van elkaar. Het axioma van de lokale stochastische onafhankelijkheid is zeer belangrijk in de IRT, maar het is erg moeilijk om te controleren of eraan voldaan is.

3.2 Schatting parameters

Om het Rasch-model als model voor het beantwoorden van de items aan te nemen moeten er getalswaarden worden ingevuld in (4) voor de parameters θ en β_i. Deze waarden zijn onbekend en zullen dus moeten worden geschat uit de observaties.
Er zijn verschillende manieren om parameters te schatten. De meest gebruikte is de grootste-aannemelijkheidsmethode (Engels: Maximum Likelihood, afgekort als ML). Deze methode wordt verreweg het meest gebruikt in de IRT-literatuur.

De methode wordt hieronder uitgelegd aan de hand van een voorbeeld. Daarna volgen nog een paar methodes die gebaseerd zijn op de ML-methode.

3.2.1 Grootste aannemelijkheidsschatter

Deze paragraaf begint met een voorbeeld van de Maximum Likelihood methode.
Een onzuiver muntstuk wordt vijf maal opgegooid, waarbij de uitkomst munt als een succes beschouwd wordt en de uitkomst kruis als een mislukking.

De toevalsvariabelen X_i worden gedefinieerd als

$$X_i = \begin{cases} 1 & \text{indien munt bij de } i\text{-de beurt,} \\ 0 & \text{indien kruis bij de } i\text{-de beurt, } (i = 1, .., 5). \end{cases}$$

Kans op succes is π, waarbij π een getal is tussen 0 en 1.
Nu wordt π geschat door gebruik te maken van een experimentje.
Stel dat de volgende uitkomst waargenomen wordt: (1 0 1 1 0).
De kans op die uitkomst is

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0; \pi) = \pi(1-\pi)\pi\pi(1-\pi) = \pi^3(1-\pi)^2 \quad (7)$$

Nu kan (7) beschouwd worden als een functie van π waarbij de uitkomst van het experiment als gegeven wordt beschouwd. Voor elke waarde van π die ingevuld wordt, wordt er een uitkomst verkregen hoe waarschijnlijk de observaties zijn, als π die waarde aanneemt (aannemelijkheidsfunctie):

$$L(\pi; (10110)) = P((10110); \pi) \quad (8)$$

De ML-schatting van π is die waarde van π waarvoor de aannemelijkheidsfunctie zo groot mogelijk wordt, d.w.z. die waarde waarvoor de gegeven observaties de grootste waarschijnlijkheid hebben.

Als de uitkomsten van een experiment voorgesteld worden als $x = (x_1, .., x_n)$, dan wordt de algemene uitdrukking voor de aannemelijkheidsfunctie:

$$L(\pi; x) = \pi^s (1-\pi)^{n-s}, \text{ waarbij } s = \sum_{i=1}^{n} x_i \qquad n = \#\text{ worpen}, s = \#\text{ successen} \quad (9)$$

Om het maximum hiervan te bepalen kiest men gewoonlijk een andere functie waarvan men weet dat ze monotoon is met de aannemelijkheidsfunctie. De meest gebruikte is de logaritme van de aannemelijkheidsfunctie:

$$\ln L(\pi; x) = s \ln \pi + (n-s) \ln(1-\pi). \quad (10)$$

Nemen we hier de afgeleide van:

$$\frac{d \ln L(\pi; x)}{d\pi} = \frac{s}{\pi} - \frac{n-s}{1-\pi}. \quad (11)$$

Gelijkstellen van (11) aan 0 geeft als oplossing

$$\hat{\pi} = \frac{s}{n}. \quad (12)$$

De oplossing garandeert echter alleen dat de eerste afgeleide 0 is indien $\pi = s/n$. Om te bepalen of dit ook met een maximum overeenkomt moeten de hogere afgeleiden onderzocht worden.

De tweede afgeleide van de log-aannemelijkheidsfunctie is gegeven door:

$$\frac{d^2 \ln L(\pi; x)}{d\pi^2} = -\frac{s}{\pi^2} - \frac{n-s}{(1-\pi)^2} \tag{13}$$

Deze functie is negatief voor alle waarden van π in het interval (0,1) (waarbij $\pi = 0$ en $\pi = 1$ buiten beschouwing worden gelaten). Dus de oplossing $\hat{\pi} = s/n$ komt overeen met een maximum van de aannemelijkheidsfunctie.

3.2.2 JML-schatting

Om een schatting te maken voor de parameters in het Rasch-model moet de methode hierboven worden aangepast: er moet nu niet één parameter geschat moet worden, maar verschillende parameters tegelijkertijd. Bij een toets bestaande uit *k* items en afgenomen bij *n* personen, moeten er dus *n* θ-parameters en *k* itemparameters geschat worden (*n* + *k* parameters).

De J in JML staat voor 'joint'. Men gebruikt deze aanduiding niet om aan te geven dat er meer parameters geschat moeten worden, maar om aan te geven dat de twee soorten parameters, persoonparameters en itemparameters, tegelijkertijd geschat worden.

Om de aannemelijkheidsfunctie op te stellen moet de notatie uitgebreid worden.
De toevalsvariabele X_{vi} verwijst naar het antwoord van persoon *v* op item *i*. De waarden die die toevalsvariabele kan aannemen, 0 of 1, zullen in het algemeen aangeduid worden met x_{vi}. Wordt er verwezen naar antwoorden van persoon *v*, dan wordt dit aangeduid met x_v, en wordt er verwezen naar alle antwoorden van alle personen in de steekproef dan wordt dit aangeduid met *X*.
Hieronder volgt een voorbeeld als verduidelijking.

Voorbeeld
Er is één persoon, *v*, met $\theta = \theta_v$ en een toets van $k = 3$ items. Veronderstel dat de antwoorden (1,0,1) zijn geobserveerd. Er moeten dus vier parameters geschat worden: θ_v, β_1, β_2 en β_3.

Gebruik makend van het principe van de lokale stochastische onafhankelijkheid kan de aannemelijkheidsfunctie voor dit antwoordpatroon geschreven worden als

$$L(\beta_1, \beta_2, \beta_3, \theta_v; (1,0,1)) = f_1(\theta_v - \beta_1)(1 - f_2(\theta_v - \beta_2)) f_3(\theta_v - \beta_3)\,. \tag{14}$$

Het rechterlid kan ook geschreven worden als:

$$\prod_{i=1}^{3} \left[f_i(\theta_v - \beta_i)\right]^{x_{vi}} \left[1 - f_i(\theta_v - \beta_i)\right]^{1-x_{vi}} \tag{15}$$

Indien $x_{vi} = 1$ is dit product voor dit item i gelijk aan $f_i(\theta_v - \beta_i)$ en indien $x_{vi} = 0$ is het product gelijk aan $(1-f_i(\theta_v - \beta_i))$.
Wordt nu de vector $(\beta_1,..., \beta_k)$ met β aangeduid, dan wordt algemene formulering:

$$L(\beta, \theta_v; x_v) = \prod_{i=1}^{k} \left[f_i(\theta_v - \beta_i)\right]^{x_{vi}} \left[1 - f_i(\theta_v - \beta_i)\right]^{1-x_{vi}} \tag{16}$$

Nu generaliseren naar een steekproef van n personen. Onder de aanname dat de antwoorden van de personen onafhankelijk zijn van elkaar (antwoorden van ene persoon bevat geen informatie over de antwoorden van een andere persoon, ook wel experimentele onafhankelijkheid).

$$L(\beta, \theta; X) = \prod_{v=1}^{n} \prod_{i=1}^{k} \left[f_i(\theta_v - \beta_i)\right]^{x_{vi}} \left[1 - f_i(\theta_v - \beta_i)\right]^{1-x_{vi}} \tag{17}$$

Totnogtoe lijkt het alsof er bij een toets van k items ook k parameters moeten worden geschat. Dit is echter niet helemaal juist. Uit (3) en (4) komt naar voren dat de itemresponsfunctie in het Rasch-model de logistische functie is met als argument $\theta - \beta_i$. Als dit verschil vastligt, ligt de functiewaarde vast en als de functiewaarde vastligt ligt het verschil ook vast. Maar als het verschil $\theta - \beta_i$ vastligt, betekent dit niet dat θ en β_i allebei vastliggen.

Stel dat van alle personen de θ_v bekend is en van alles items de β_i. Een andere, doch evenwaardige oplossing ontstaat als aan elke persoon v het getal $\theta^*_v = \theta_v + c$ en aan elk item het getal $\beta^*_i = \beta_i + c$ toegekend wordt, waarbij c een willekeurige constante is.
Dan geldt natuurlijk dat $\theta^*_v - \beta^*_i = \theta_v - \beta_i$, en dus blijft de itemresponsfunctie onveranderd welke waarde er ook aan c gegeven wordt. Wil er zinvol over de parameters gesproken kunnen worden dan moet de waarde van c vastgelegd worden, of met andere woorden, het nulpunt van de schaal moet vastgelegd worden. Dit kan gedaan worden door bijvoorbeeld één van de parameters (bijvoorbeeld β_1) gelijk te stellen aan nul. In dat geval zijn er nog maar k –1 vrije itemparameters over. Het kiezen van het nulpunt noemt men normaliseren. De meest gebruikte normalisatie is het nulpunt zo te kiezen dat $\sum_{i=1}^{k} \beta_i = 0$.

De JML-methode bestaat uit heel veel rekenwerk, maar er is er nog een ander probleem: Consistentie. Ruwweg betekent consistentie dat, hoe meer informatie men verzamelt over een parameter door de steekproef steeds groter te maken, des te nauwkeuriger de schatting moet zijn en in de limiet, bij $n \to \infty$ is de kans dat men de parameter juist schat gelijk aan 1. In het geval van het Rasch-model treedt er echter een complicatie op: om meer informatie te verzamelen over itemparameters dient men de toets steeds bij nieuwe personen af te nemen, doch elke persoon die men aan de steekproef toevoegt brengt zijn eigen onbekende θ-parameter mee. Dit wil zeggen dat de omvang van het probleem, het aantal te schatten parameters, even snel groeit als het aantal personen in de steekproef.
Dit maakt de JML-schattingsmethode oninteressant.

Als men echt in de itemparameters is geïnteresseerd, dan is het veel handiger naar een schattingsmethode te zoeken waarbij men geen last meer heeft van het steeds groeiende aantal θ-parameters. Deze parameters, waar men in eerste instantie niet zo in geïnteresseerd is, maar die toch in het model aanwezig zijn worden in de literatuur aangeduid met de term 'nuisance parameters'. De andere parameters waarin men wel is geïnteresseerd worden structurele parameters genoemd.

3.2.3 CML-schatting

De CML-schatting is een methode om de 'nuisance parameters' die in de vorige paragraaf besproken zijn kwijt te raken.

Het idee hierachter is om de totale steekproef op te delen in homogene scoregroepen, dat wil zeggen in groepen personen die een zelfde aantal items correct hebben, en de aannemelijkheid van een bepaald antwoordpatroon te bekijken binnen elke scoregroep afzonderlijk. Technisch drukt men dat uit door te zeggen dat men conditioneert op de score (de C van CML staat voor conditional).

Voorbeeld
Veronderstel $k = 3$ en beschouw het antwoordpatroon (1 0 1). De score s van dit antwoordpatroon is 2. Er zijn exact drie antwoordpatronen met score 2, namelijk (1 0 1), (1 1 0) en (0 1 1).

$$P(1 \quad 0 \quad 1 \mid s = 2, \theta) = \frac{P(1 \quad 0 \quad 1 \mid \theta)}{P(1 \quad 0 \quad 1 \mid \theta) + P(1 \quad 1 \quad 0 \mid \theta) + P(0 \quad 1 \quad 1 \mid \theta)} \tag{18}$$

Twee equivalente formules voor het Rasch-model:

$$P(X_i = 1 \mid \theta) = f_i(\theta - \beta_i) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \tag{19}$$

en

$$P(X_i = 0 \mid \theta) = 1 - f_i(\theta - \beta_i) = \frac{1}{1 + \exp(\theta - \beta_i)}. \tag{20}$$

Als de aannemelijksheidsfunctie opgesteld moet worden, moeten de producten genomen worden van deze uitdrukkingen. Merk op dat de noemers identiek zijn, deze zijn dus onafhankelijk van het specifieke antwoordpatroon (noem deze K). Nu is de kans op antwoordpatroon (1 0 1):

$$P(1 \quad 0 \quad 1 \mid \theta) = \frac{\exp(\theta)\exp(-\beta_1)\exp(\theta)\exp(-\beta_3)}{K} = \frac{\exp(2\theta)\exp(-\beta_1 - \beta_3)}{K}. \tag{21}$$

Dus geldt er ook:

$$P(1\quad 0\quad 1 \mid s=2,\theta)=\frac{\frac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K}}{\frac{\exp(2\theta)\exp(-\beta_1-\beta_3)}{K}+\frac{\exp(2\theta)\exp(-\beta_1-\beta_2)}{K}+\frac{\exp(2\theta)\exp(-\beta_2-\beta_3)}{K}} \tag{22}$$

$$=\frac{\exp(-\beta_1-\beta_3)}{\exp(-\beta_1-\beta_3)+\exp(-\beta_1-\beta_2)+\exp(-\beta_2-\beta_3)} \tag{23}$$

Uitdrukking (23) beschouwd als een functie van de β-parameters, wordt de conditionele aannemelijkheidsfunctie genoemd voor het patroon (1 0 1).
De conditionele aannemelijkheidsfunctie voor alle geobserveerde antwoordpatronen samen is het product van soortgelijke uitdrukkingen. De conditionele grootste-aannemelijkheids-schattingen zijn dan die waarden van de β-parameters die dit product zo groot mogelijk maken. Het vinden van die waarden is behoorlijk ingewikkeld, maar daar zijn computerprogramma's voor ontwikkeld (OPLM).

Er is aangetoond dat met een bepaalde berekeningsmethode zeer nauwkeurige resultaten verkregen worden: bij k=5000 zijn slechts de laatste vier decimalen van het resultaat dat in veertien decimalen wordt weergegeven aangetast door afrondingsfouten.

Er is nog niets gezegd over de manier waarop de getoetste personen uit de populatie getrokken dienen te worden. Dit is met opzet gebeurd. Er is niet stilzwijgend verondersteld dat de steekproef een aselecte trekking moet zijn uit de populatie. Integendeel, door gebruik te maken van de CML-methode maakt het in principe niet uit hoe de steekproef uit de populatie is getrokken. Immers de CML-methode wordt gebruikt om iets te kunnen zeggen over de itemparameters en niet over de populatie van personen.

3.2.4 MML-schatting

Een tweede methode om de individuele θ-parameters kwijt te raken bestaat eruit ze een andere status te geven. De status van de θ-waarden is het standpunt van waaruit men de gegevens beschouwt. Totnogtoe is er eigenlijk impliciet aangenomen dat, als Jan en Piet tot de steekproef behoren, we ter zelfder tijd geïnteresseerd zijn in de waarde van de itemparameters en in de θ-waarde van Jan en Piet en van alle andere personen die tot de steekproef behoren. Een ander standpunt is dat het ons eigenlijk niet kan schelen wie er in de steekproef zit, omdat we alleen maar geïnteresseerd zijn in de itemparameters. Dit impliceert dat we de steekproef als een aselecte steekproef uit een of andere populatie beschouwen, en dat we de gedragingen van die toevallige steekproef willen gebruiken om de itemparameters te schatten. Dit standpunt biedt de mogelijkheid om θ kwijt te raken op de volgende manier.

Veronderstel dat θ slechts drie verschillende waarden kan aannemen in de populatie, namelijk -1, 0 en 1, en veronderstel dat deze waarden in de populatie voorkomen met een proportie van respectievelijk 0.25, 0.35 en 0.40. We beschouwen nu de kans dat we het antwoordpatroon x= (1 0 1) observeren bij aselecte trekking van een persoon uit de populatie. Deze kans is gegeven door:

$$P(x) = 0.25 \times P(x \mid \theta = -1) + 0.35 \times P(x \mid \theta = 0) + 0.40 \times P(x \mid \theta = 1). \qquad (24)$$

Dat wil zeggen, als θ onbekend is, kunnen we alle conditionele kansen als $P(x|\theta)$ het ware gaan middelen door te vermenigvuldigen met de kans dat die θ optreedt, en die gewogen conditionele kansen op te tellen. Het resultaat noemt men marginale kans. Vandaar de eerste M in MML.

Nu de situatie waarin het aantal verschillende waarden dat θ kan aannemen gelijk is aan W:

$$P(x) = \sum_{j=1}^{W} P(x \mid \theta_j) P(\theta_j). \qquad (25)$$

Hier zijn de waarden voor $P(\theta_j)$ onbekend alsmede alle θ_j's. Hoe groter W, des te meer onbekende parameters er zijn. Als we nu θ als continue stochastische variabele modelleren met een normale (μ,σ^2)-verdeling, reduceert het aantal onbekende parameters weer tot twee. Dit resulteert in de dichtheidsfunctie (van in dit geval de normale verdeling) g_{μ,σ^2}:

$$g_{\mu,\sigma^2}(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2} \right], \qquad (26)$$

waarin π = 3,14159...

De marginale kans van antwoordpatroon x in het geval er een normale verdeling verondersteld wordt van θ, is gegeven door

$$\begin{aligned} P(x) &= \int_{-\infty}^{\infty} P(x \mid \theta) g_{\mu,\sigma^2}(\theta) d\theta \qquad (27) \\ &= \int_{-\infty}^{\infty} P(x \mid \theta) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2} \right] d\theta. \end{aligned}$$

Formule (27) is niet meer afhankelijk van θ, maar wel van de itemparameters en van de twee verdelingsparameters μ en σ^2.

Beschouw nu de marginale kans als functie van die parameters, dan krijgen we de marginale aannemelijkheidsfuncie voor het antwoordpatroon x.

$$L(\beta, \mu, \sigma^2; X) = \prod_{v=1}^{n} \int_{-\infty}^{\infty} P(x_v \mid \theta) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2} \right] d\theta \qquad (28)$$

En de log-aannemelijkheidsfunctie wordt dan:

$$\ln L(\beta, \mu, \sigma^2; X) = \sum_{v=1}^{n} \ln \int_{-\infty}^{\infty} P(x_v \mid \theta) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2} \right] d\theta \,. \qquad (29)$$

Vergelijking van CML- en de MML-methode

Bij CML wordt geen enkele veronderstelling gemaakt over de verdeling van θ in de populatie, terwijl dat bij MML wel wordt gedaan.
Het is bij MML helemaal niet noodzakelijk een normale verdeling te veronderstellen; men zou ook een andere verdeling kunnen veronderstellen. Belangrijk is echter in te zien dat de veronderstelling over de verdeling nu deel gaat uitmaken van het model. Dus als het MML wordt toegepast (met de normale verdeling) dan worden als het ware twee modellen vermengd: het Rasch-model (dat iets vertelt over de antwoorden gegeven θ) en de normale verdeling (die vertelt hoe de θ's in de populatie zijn verdeeld). Een gebruiker die MML gebruikt stelt zich iets kwetsbaarder op, omdat een fout in de veronderstelling over de normale verdeling hetzij omdat θ niet normaal verdeeld is, hetzij omdat de steekproef niet aselect uit de normale verdeling is getrokken, heeft als gevolg dat er ook systematische fouten geïntroduceerd worden in de schatting van de itemparameters.
Voordeel van MML is wel dat de verdelingsparameters gelijktijdig met de itemparameters geschat kunnen worden.

4 Case Study

4.1 Inleiding

Om de toepassing in de praktijk te zien van het Rasch-model zijn er vele computerprogramma's ontworpen en op de markt gebracht. Een van de bekendste is wel het op het CITO ontwikkelde programmapakket OPLM (éénparameter logistisch model). Dit pakket bevat een aantal programma's waarmee data kunnen worden geanalyseerd volgens het Rasch-model en enkele aanpassingen daarop.

Bij het gebruiken van zo'n computerpakket gaat de theorie erachter grotendeels verloren en wordt er alleen naar de resultaten gekeken. De theorie achter het model en het toewerken naar resultaten is echter zodanig interessant dat in dit hoofdstuk wat simpele situaties worden nagebootst om respectievelijk het model op te stellen, toe te passen en uitspraken te doen over de resultaten.

Om het Rasch-model in de praktijk te brengen is er in dit hoofdstuk gebruik gemaakt van een rekentoets die is afgenomen bij diverse pabostudenten. De inhoud van de toets is onbelangrijk voor het opstellen van het model. Alleen de antwoorden op de dichotome vragen zijn van belang. In bijlage C vindt u de uitslagen van de toets.

In paragraaf 4.2 wordt de aanpak van de diverse situaties omschreven die gesimuleerd of onderzocht gaan worden. Deze omschrijvingen bestaan uit een situatieschets, de methoden en technieken die gebruikt gaan worden en eventueel worden de aannames besproken.

4.2 Aanpak

4.2.1 Model opstellen

Als eerste wordt er een Rasch-model ontwikkeld voor een fictieve situatie. Hiervoor is gekozen omdat de omvang van de praktijkgegevens in bijlage C groot is en daardoor het model rekenintensief wordt.

De volgende aannames worden gemaakt:
Voor het opstellen van een model gaan we uit van een eenvoudige situatie en maken we gebruik van de theorie beschreven in paragraaf 3.1. Voor het schatten van de parameters is er gekozen voor de JML-schatting, paragraaf 3.2.2.
Stel dat er 2 items zijn die afgenomen worden bij 10 studenten. De twee items hebben beide een moeilijkheidsgraad van respectievelijk β_1 en β_2. De studenten worden geacht over een eigen vaardigheidsniveau te beschikken, nl. θ_1, θ_2,.., θ_{10}.

In het Rasch-model, beschreven in paragraaf 3.1, is de itemresponsfunctie een logistische functie. Met behulp van deze functie kunnen de kansen bepaald worden voor het al dan niet goed beantwoorden van de items door student i:

$$P_{i1} = \frac{e^{\theta_i - \beta_1}}{1 + e^{\theta_i - \beta_1}} \qquad P_{i2} = \frac{e^{\theta_i - \beta_2}}{1 + e^{\theta_i - \beta_2}} \,. \tag{30}$$

Zoals uitgelegd in paragraaf 3.2.2 is het noodzakelijk dat er genormaliseerd wordt. Hiervoor kiezen we β_1 gelijk aan 0. En omdat β_2 nog maar de enige β-parameter is noemen we deze vanaf nu gewoon β. Hierdoor is het mogelijk om P_{i1} en P_{i2} nog verder te vereenvoudigen tot

$$P_{i1} = \frac{e^{\theta_i}}{1 + e^{\theta_i}} \qquad P_{i2} = \frac{e^{\theta_i - \beta}}{1 + e^{\theta_i - \beta}} \,. \tag{31}$$

Om het model toe te kunnen passen moeten de parameters worden geschat die in het model gebruikt worden, nl. θ_1, .., θ_{10} en β. Hiervoor maken we gebruik van de JML-methode. Deze methode staat beschreven in paragraaf 3.2.2. De methode is gebaseerd op de meest aannemelijke schatter.

> **Definitie**. De *meest aannemelijke schatter* (engels: maximum likelihood estimator) voor θ is die waarde $T(X) \in \Theta$ die de functie $\theta \rightarrow L(\theta;X)$ maximaliseert. Oosterhoff (2000)

De log aannemelijkheidsfunctie kan als volgt worden gedefinieerd:

$$\begin{aligned} \log L(\theta_1,..,\theta_{10},\beta) &= \sum_{i=1}^{10} \left(x_{i1} \log P_{\theta,\beta}(X_{i1} = x_{i1}) + x_{i2} \log P_{\theta,\beta}(X_{i2} = x_{i2}) \right) \\ &= \sum_{i=1}^{10} \left\{ x_{i1}\theta_i - \log(1 + e^{\theta_i}) + x_{i2}(\theta_i - \beta) - \log(1 + e^{\theta_i - \beta}) \right\} \end{aligned} \tag{32}$$

De ML-schatting van de parameters zijn die waarden waarvoor de aannemelijkheids-functie zo groot mogelijk wordt. We zoeken dus naar een maximum van (32). Hiervoor moet er gedifferentieerd worden. De vraag is nu alleen naar welke soort parameter er gedifferentieerd moet worden: naar de θ_i's of naar β?

In totaal moeten er 11 parameters geschat worden (10 θ_i'2 en 1 β). Wordt er gekozen voor een vaste β dan kunnen de θ_i's geschat worden en kan er gezocht worden naar een optimale β, dus dit is in ons geval een 1-dimensionaal probleem.
Er kan ook gekozen worden om de θ_i's vast te nemen en dan de β daarvan af te leiden. Hierna kan er gezocht worden naar de optimale waarden voor de θ_i's. Dit zijn echter nog 10 parameters, dus het probleem wordt 10-dimensionaal.

De keuze valt dus op de eerste strategie, omdat het 11-dimensionale probleem gereduceerd wordt tot een 1-dimensionaal probleem in tegenstelling tot de reductie naar een 10-dimensionaal probleem bij de tweede strategie. En tevens zijn we meer geïnteresseerd in de itemparameters dan in de vaardigheidsparameters.

Dus we gaan op zoek naar een maximum van (32) op θ_i:

$$\frac{\partial}{\partial \theta_i} = x_{i1} - \frac{1}{1+e^{\theta_i}} e^{\theta_i} + x_{i2} - \frac{1}{1+e^{\theta_i-\beta}} e^{\theta_i-\beta}$$

$$= x_{i1} + x_{i2} - \frac{1}{1+e^{-\theta_i}} - \frac{1}{1+e^{-\theta_i+\beta}} \qquad (33)$$

Om (32) te maximaliseren moet het nulpunt van deze afgeleide gezocht worden.
Twee problemen die aan de orde zijn bij het bepalen van het nulpunt van (33):

- Hoe wordt het nulpunt gezocht als er twee typen parameters onbekend zijn (θ_i's en β)?
- Is het nulpunt van de afgeleide van de aannemelijkheidsfunctie (33) wel een maximum van de aannemelijkheidsfunctie (32)?

Het eerste probleem wordt opgelost door van een vaste β uit te gaan.
Op basis van deze vaste β worden de θ_i's van de studenten bepaald. Doordat dan alle benodigde parameterwaarden (schattingen) aanwezig zijn kan de waarde van de log aannemelijkheidsfunctie bepaald worden.

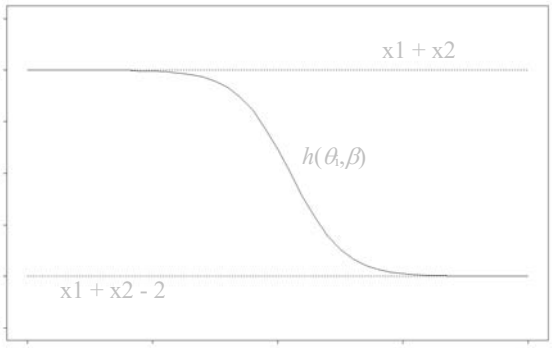
Om vast te stellen dat het nulpunt van de afgeleide van de log aannemelijkheidsfunctie wel een maximum voorstelt voor de log aannemelijkheidsfunctie moet er gekeken worden naar het functieverloop van (33). Deze functie kan ook wel geschreven worden als functie $h(\theta_i,\beta)$:

$$h(\theta_i, \beta) = x_{i1} + x_{i2} - \frac{1}{1+e^{-\theta_i}} - \frac{1}{1+e^{-\theta_i+\beta}} \qquad (34)$$

Deze functie is een monotoon dalende functie. De derde en de vierde term bezitten beiden een negatieve exponentiele functie $e^{-\theta}$, waardoor de noemer dus ook dalend is. De hele breuk is echter weer een stijgende functie, maar het minteken zorgt er uiteindelijk voor dat beide termen dalend zijn.

Het asymptotisch gedrag van (34):
$\lim_{\theta_i \to -\infty} h(\theta_i, \beta) = x_{i1} + x_{i2}$
$\lim_{\theta_i \to \infty} h(\theta_i, \beta) = x_{i1} + x_{i2} - 2$

Een grafische weergave van $h(\theta_i,\beta)$ staat in figuur 2.



Figuur 2. Asymptotisch gedrag van $h(\theta_i,\beta)$

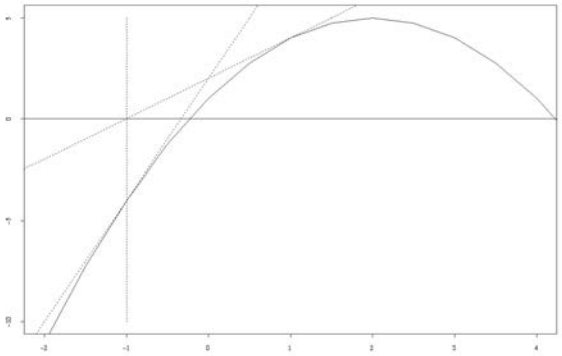
Aangezien er maar vier antwoordencombinaties mogelijk zijn, nl. $(x_1, x_2) = \{(0,0), (1,0), (0,1), (1,1)\}$ en de combinaties (0,0) en (1,1) ervoor zorgen dat de θ_i's respectievelijk naar $-\infty$ en ∞ gaan, zijn we alleen geïnteresseerd in de studenten die precies een van beide vragen correct heeft beantwoord (totale score van 1).

Vervangen we $x_{i1} + x_{i2}$ door 1 in $h(\theta_i,\beta)$ en wordt (34) gelijk aan nul gesteld, dan zal dit de gezochte θ_i's opleveren.

Doordat er twee soorten parameters onbekend zijn lijkt exact oplossen zeer complex te worden. Daarom wordt er eerst gestart met numeriek benaderen en daarna zal er ook naar een exacte oplossing worden gezocht.

Numerieke oplossing

Dit gebeurt door gebruik te maken van lineaire Taylor-ontwikkeling rond het nulpunt.
De situatie kan men zich voorstellen zoals weergegeven in figuur 3.
.



Figuur 3. Taylor-ontwikkeling rond nulpunt

Kies nu een startwaarde $\theta_i^{(0)}$ en benader $f(\theta_i)$ door lineaire Taylorontwikkeling in $\theta_i^{(0)}$.

$$t(\theta;\theta_i^{(0)}) = f(\theta_i^{(0)}) + (\theta - \theta_i^{(0)}) f'(\theta_i^{(0)})$$

Het nulpunt van deze raaklijn wordt nu als volgt bepaald:

$$\begin{aligned} 0 &= f(\theta_i^{(0)}) + (\theta - \theta_i^{(0)}) f'(\theta_i^{(0)}) \\ f(\theta_i^{(0)}) &= -(\theta - \theta_i^{(0)}) f'(\theta_i^{(0)}) \\ (\theta - \theta_i^{(0)}) &= -\frac{f(\theta_i^{(0)})}{f'(\theta_i^{(0)})} \\ \theta &= \theta_i^{(0)} - \frac{f(\theta_i^{(0)})}{f'(\theta_i^{(0)})} \end{aligned} \tag{35}$$

waarbij

$$f'(\theta_i) = \frac{-e^{-\theta_i}}{(1+e^{-\theta_i})^2} - \frac{e^{-\theta_i+1}}{(1+e^{-\theta_i+1})^2} \tag{36}$$

4.2.2 Betrouwbaarheidsinterval β

Om uitspraken te kunnen doen over de betrouwbaarheid van de parameterschatting β, wordt er gebruik gemaakt van de bootstraptechniek. Deze techniek werkt als volgt: je neemt alle waarnemingen en berekent hiervan de parameter(s) die je wilt weten. Vervolgens heb je dus een verdeling zonder onbekenden (de onbekende(n) die je had heb je ingevuld met het getal(len) berekend uit de sample). Vervolgens laat je een computer random samples nemen uit de verdeling die je hebt gemaakt en van al deze samples bereken je de toetsingsgrootheid waarvan je het betrouwbaarheidsinterval wilt weten.

Er zal in dit geval dus een bootstrapfunctie geschreven moeten worden die herhaaldelijk de waarde van β berekent. Hier uit zal dan een betrouwbaarheidsinterval voor β geconstrueerd worden.

Om een nauwkeurig resultaat te krijgen wordt er gebruik gemaakt van een grote hoeveelheid gegevens. Er zal een betrouwbaarheidsinterval geconstrueerd worden voor de itemparameter van item 2. De gegevens die gebruikt zullen worden om hiertoe te komen zijn de eerste twee kolommen uit tabel 2 in Bijlage C.

4.2.3 Toets

De toets die vooraf ging aan de resultaten in Bijlage C bestaat uit 11 vragen. Voordat er naar de resultaten wordt gekeken, vraagt men zich af of vraag 3 en 11 even moeilijk (of makkelijk) zijn.

Om een uitspraak te doen over de moeilijkheid van twee vragen ten opzichte van elkaar, maken we gebruik van de tweesteekproef binomiale toets, Larsen (2001).

Laat x en y het aantal successen voorstellen die geobserveerd zijn in twee onafhankelijke verzamelingen van n en m vragen.

Dan geldt:

$$\hat{p} = \frac{x+y}{n+m} \qquad \text{en} \qquad z = \frac{\frac{x}{n} - \frac{y}{m}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{p}(1-\hat{p})}{m}}}.$$

Om nu H_0: $p_x = p_y$ tegen H_1: $p_x \neq p_y$ te toetsen met een significantieniveau van $\alpha = 0.05$, verwerpen we H_0 als
(1) $z \leq -z_{\alpha/2}$ of
(2) $z \geq z_{\alpha/2}$.

In dit geval:
Laat p_3 en p_{11} het percentage goede antwoorden voorstellen op respectievelijk de vragen 3 en 11. De hypothesen die getoetst moeten worden zijn:

H_0: $p_3 = p_{11}$
H_1: $p_3 \neq p_{11}$

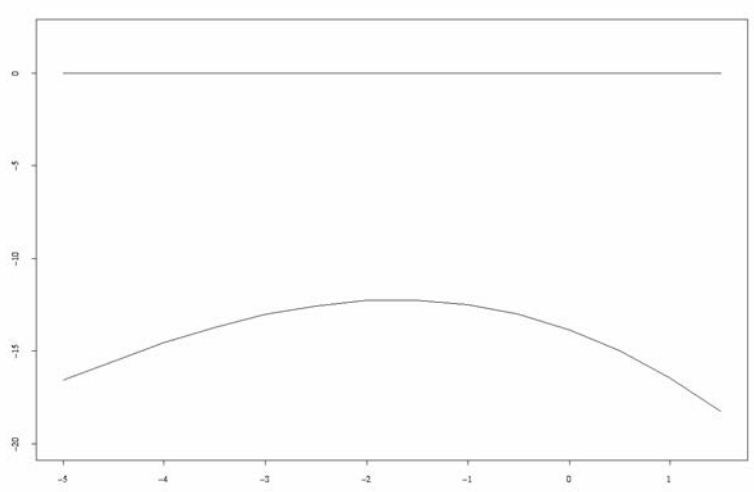
4.3 Resultaten

4.3.1 Model opstellen

Numerieke benadering
De simulatie die gaat plaatsvinden gaat uit van de situatie waarin drie studenten de eerste vraag goed hebben beantwoordt en de tweede vraag fout en zeven studenten op de eerste vraag het verkeerde antwoord hebben gekregen en de twee vragen juist correct hebben beantwoord. De resultaten kunnen dus worden weer gegeven als

$$m = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

Door gebruik te maken van de geprogrammeerde functie ‘likelihood’ in Bijlage A, is de waarde van (32) veelvuldig uitgerekend bij verschillende startwaarden van β.
Deze waarden zijn grafisch weergegeven in figuur 4, waarbij β op de horizontale as staat weergegeven en de daarbij behorende likelihoodwaarde op de verticale as staat.



Figuur 4. Likelihoodwaarde als functie van β

Uit figuur 4 kan afgelezen worden dat de waarde van $\hat{\beta}$ rond –1,5 ligt.

De gebruikte functie ‘likelihood’ uit bijlage A maakt op zijn beurt weer gebruik van de functie ‘nulpunt’, die ook is weergegeven in bijlage A. Met behulp van deze functie kunnen de afzonderlijke waarden van θ_i voor elke student i worden uitgerekend.
Wat opvalt is dat elke θ_i precies de helft is van β.

<u>Exacte oplossing</u>
Om de precieze waarde van β te bepalen moet de waarde van β direct gevonden worden door (34) gelijk te stellen aan 0 en $x_{i1} + x_{i2}$ te vervangen door 1:

$$1 - \frac{1}{1+e^{-\theta_i}} - \frac{1}{1+e^{-\theta_i+\beta}} = 0$$

$$\frac{2 + e^{-\theta_i+\beta} + e^{-\theta_i}}{1 + e^{-\theta_i} + e^{-\theta_i+\beta} + e^{-2\theta_i+\beta}} = 1$$

$$e^{-2\theta_i+\beta} = 1$$

$$\theta_i = \frac{1}{2}\beta$$

Ditzelfde resultaat is ook gevonden bij de numerieke benadering.
Deze uitdrukking van θ in termen van β kan nu gebruikt worden in de likelihoodfunctie.

De likelihoodfunctie wordt dan

$$\widetilde{L}(\beta) = \sum_{i=1}^{10} x_{i1}\frac{1}{2}\beta - \log(1+e^{\frac{1}{2}\beta}) - x_{i2}\frac{1}{2}\beta - \log(1+e^{-\frac{1}{2}\beta})$$

$$= \frac{1}{2}\beta\sum_{i=1}^{10}(x_{i1}-x_{i2}) - I*\left[\log(1+e^{\frac{1}{2}\beta}) + \log(1+e^{-\frac{1}{2}\beta})\right], \text{ waarbij I= \# studenten met score gelijk aan 1}$$

$$= \frac{1}{2}\beta\sum_{i=1}^{10}(x_{i1}-x_{i2}) - I*\log(1+e^{\frac{1}{2}\beta})(1+e^{-\frac{1}{2}\beta})$$

$$= \frac{1}{2}\beta\sum_{i=1}^{10}(x_{i1}-x_{i2}) - I*\log(2+e^{\frac{1}{2}\beta}+e^{-\frac{1}{2}\beta}) \quad (37)$$

$$\frac{\partial\widetilde{L}(\beta)}{\partial\beta} = \frac{1}{2}\sum_{i=1}^{10}(x_{i1}-x_{i2}) - \frac{I(e^{\frac{1}{2}\beta}-e^{-\frac{1}{2}\beta})}{4+2e^{\frac{1}{2}\beta}+2e^{-\frac{1}{2}\beta}}\ . \quad (38)$$

Nu het linkerlid gelijkstellen aan 0:

$$0 = \frac{1}{2}\sum_{i=1}^{10}(x_{i1}-x_{i2}) - \frac{I(e^{\frac{1}{2}\beta}-e^{-\frac{1}{2}\beta})}{4+2e^{\frac{1}{2}\beta}+2e^{-\frac{1}{2}\beta}}. \quad (39)$$

$$\frac{1}{2}\sum_{i=1}^{10}(x_{i1}-x_{i2}) = \frac{I(e^{\frac{1}{2}\beta}-e^{-\frac{1}{2}\beta})}{4+2e^{\frac{1}{2}\beta}+2e^{-\frac{1}{2}\beta}}$$

$$(2+e^{\frac{1}{2}\beta}+e^{-\frac{1}{2}\beta})\sum_{i=1}^{10}(x_{i1}-x_{i2}) = I(e^{\frac{1}{2}\beta}-e^{-\frac{1}{2}\beta})$$

Neem nu $c = \frac{1}{2}\sum_{i=1}^{10}(x_{i1}-x_{i2})$, dan geldt:

$$2c(2+e^{\frac{1}{2}\beta}+e^{-\frac{1}{2}\beta}) = I(e^{\frac{1}{2}\beta}-e^{-\frac{1}{2}\beta}) \quad (40)$$

$$2c(2e^{\frac{1}{2}\beta}+e^{\beta}+1) = I(e^{\beta}-1)$$

$$(2c-I)e^{\beta} + 4ce^{\frac{1}{2}\beta} + 2c + I = 0$$

Neem nu $\nu = e^{\frac{1}{2}\beta}$, dan geldt:

$$(2c-I)\nu^2 + 4c\nu + 2c + I = 0 \quad (41)$$

Als deze (41) een oplossing heeft met $\nu > 0$ dan kunnen we β schatten:

$$\hat{\beta} = 2\log\nu \tag{42}$$

Voorbeeld
Stel er zijn 10 studenten met als scores 3 keer (1,0) en 7 keer (0,1).

$$c = \frac{1}{2}\sum_{i=1}^{10}(x_{i1} - x_{i2}) = \frac{1}{2}(3-7) = -2$$
$$I = \#\{j : x_{j1} \neq x_{j2}\} = 10$$

Dan geldt verder:
$$(-4-10)\nu^2 - 8\nu - 4 + 10 = 0$$
$$-14)\nu^2 - 8\nu + 6 = 0$$
$$\nu^2 + \frac{4}{7}\nu - \frac{3}{7} = 0$$
$$(\nu + \frac{2}{7})^2 - \frac{4}{49} - \frac{3}{7} = 0$$
$$(\nu + \frac{2}{7})^2 = \frac{25}{49}$$
$$\nu + \frac{2}{7} = \pm\frac{5}{7}$$
$$\nu_1 = \text{-1 of} \qquad \nu_2 = \frac{3}{7}$$

Om β te kunnen schatten geldt de voorwaarde dat $\nu > 0$, dus alleen ν_2 voldoet.
Dit levert het volgende resultaat op:

$$\hat{\beta} = 2\log\frac{3}{7} \approx -1{,}694596\,.$$

En deze waarde komt overeen met de waarde die is af te lezen uit figuur 4.

In onderstaande tabel staan nog meer resultaten voor verschillende antwoordpatronen:

# (1,0)	# (0,1)	β
0	10	-∞
1	9	-4,394449
2	8	-2,772589
3	7	-1,694596
4	6	-0,810930
5	5	0
6	4	0,810930
7	3	1,694596
8	2	2,772589
9	1	4,394449
10	0	∞

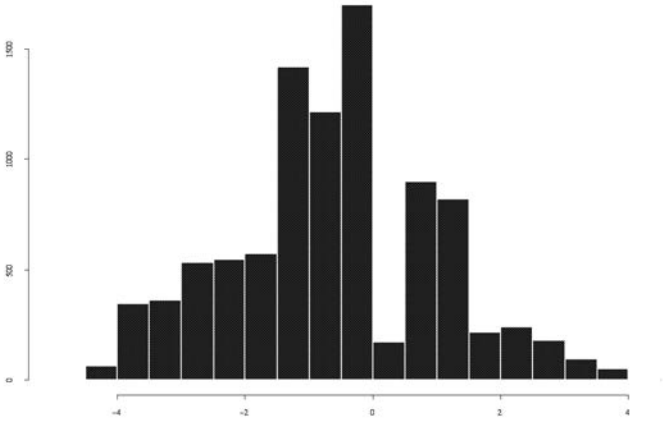
Tabel 1. β-waarden voor diverse antwoordpatronen

4.3.2 Betrouwbaarheidsinterval β

Allereerst is de waarde van β bepaald op basis van de originele data (eerste twee kolommen uit de tabel in Bijlage C):

β = -0,3646431

Het resultaat na het uitvoeren van de bootstrapanalyse is een vector met β-realisaties. Uit deze vector zijn de β's met de waarden -∞ en ∞ uit deze vector verwijderd. De β-realisaties die over zijn gebleven zijn weergegeven in figuur 5.



Figuur 5. Histogram β-realistaties na bootstrapanalyse

Doordat de resultaatvector is opgeschoond kunnen er hier geen uitspraken gedaan worden over het gemiddelde en de standaarddeviatie van de berekende β-waarden.

Het 95% betrouwbaarheidsinterval voor β kan echter wel worden bepaald, met behulp van de kwantielen. Daarvoor wordt de vector met β-realisaties op volgorde gezet en is er gekeken naar de grenzen van respectievelijk 2,5% en 97,5% van de waarnemingen. Dit heeft geleid tot het 95%-betrouwbaarheidsinterval voor β:

[-3.583519, 2.772589]

4.3.3 Toets

Neem α =0,05. Dan zullen $\pm z_{.025}$ = 1,96 de twee kritieke waarden zijn. En er geldt $m = n$.

Dan geldt:

$$\hat{p} = \frac{x+y}{2n} = \frac{21+15}{2*33} = 0{,}545$$

$$z = \frac{\frac{x-y}{2*n}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{p}(1-\hat{p})}{n}}} = \frac{\frac{21-15}{33}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{33} + \frac{\hat{p}(1-\hat{p})}{33}}} = 1{,}730$$

De conclusie is duidelijk: we kunnen de nulhypothese niet verwerpen. Deze gegevens bevatten niet genoeg bewijs dat het aantal goed antwoorden voor vraag 3 anders is dan voor vraag 11.

4.4 Conclusies

4.4.1 Itemparameter

Uit de numerieke benadering komt naar voren dat voor elke student *i* geldt $\theta_i = ½\ \beta$.
Dit valt echter wel te verklaren. Welke waarde voor β men ook als (start)waarde kiest, elke student heeft 1 vraag goed en 1 vraag verkeerd beantwoord. En aangezien de β van de eerste vraag gelijk is aan 0 en die van de tweede β bepaald wordt aan de hand van de scores, is het niet verwonderlijk dat de studenten een θ krijgen die precies de helft is van deze bepaalde β. Toch zou je verwachten dat de θ wel voor studenten verschillend zal zijn als ze een moeilijker item wel goed beantwoorden en een makkelijk item fout beantwoorden. Dat is in werkelijkheid ook het geval, maar de strategie die hier gekozen is (zie paragraaf 4.2.1) is hoofdzakelijk geïnteresseerd in de verhouding van de moeilijkheidsgraad tussen beide items. Hierdoor wordt eenmalig een θ bepaald voor elke student aan de hand van een vooraf gekozen β (startwaarde) en na het vinden van de optimale β worden de θ's van de studenten niet meer aangepast.

De tekens van de verkregen waarden voor β (β_2) zijn eenvoudig te interpreteren:
Als $\beta < 0$ vraag twee is moeilijker dan vraag één,
Als $\beta = 0$ vraag één en twee zijn van dezelfde moeilijkheidsgraad,
Als $\beta > 0$ vraag twee is makkelijker dan vraag één.

Voor de exacte waarde kan het bepalen van β op een redelijk eenvoudige manier:

$$s_1 = \sum_{j:x_{j1} \neq x_{j2}} x_{i1} \text{ en} \qquad s_2 = \sum_{j:x_{j1} \neq x_{j2}} x_{i2}$$

$$\hat{\beta} = 2\log\frac{s_1}{s_2}$$

4.4.2 Items even moeilijk?

Om te onderzoeken of twee items van dezelfde moeilijkheidsgraad zijn kan er op twee manieren te werk gegaan worden.

De methode, waarvan in paragraaf 4.3.3 de resultaten staan, vergelijkt van beide vragen de moeilijkheidsgraad met elkaar. Er wordt getoetst of de moeilijkheidsgraden aan elkaar gelijk zijn.
Een andere manier om een uitspraak te doen of de itemparameters en aan elkaar gelijk zijn is gebruik te maken van het opgestelde Rasch-model. In paragraaf 4.3.1 is een model opgesteld voor twee items waarvan de eerste een moeilijkheid heeft van $\beta_1=0$. Met behulp van een bootstrapanalyse, waarvan resultaten staan in paragraaf 4.3.2, kan dan een betrouwbaarheidsinterval geconstrueerd worden. Om de nulhypothese te verwerpen dat de twee items even moeilijk zijn, zal $\beta_2=0$ niet in het 95%-betrouwbaarheidsinterval mogen liggen.

Er zijn dus twee manieren om de nulhypothese te testen dat twee items even moeilijk zijn. Dit kan met behulp van een tweesteekproef binomiale toets waarbij de moeilijkheid van het ene item tegen opzichte van de andere wordt getoetst. Maar men kan ook gebruik maken van het Rasch-model, waarbij de moeilijkheid van het ene item 0 wordt genomen en voor de andere een betrouwbaarheidsinterval wordt geconstrueerd. Ligt 0 in dit betrouwbaarheidsinterval dan kan de nulhypothese niet verworpen worden.

5 Conclusies & Aanbevelingen

Het Rasch-model, dat een speciaal geval is van de itemresponstheorie, is gebaseerd op de itemresponsfunctie. Deze functie is in het geval van het Rasch-model zeer eenvoudig en hangt voor één persoon maar af van twee soorten parameters: de vaardigheid van de persoon en de moeilijkheid van een items. Er zijn diverse methodes bekend om deze parameters te schatten, waarbij de bekendsten allen zijn gebaseerd op de methode van de meest aannemelijke schatter. Er zijn diverse computerprogramma's ontwikkeld (waaronder OPLM bij het Cito) om deze schatters uit te rekenen. Het handmatig berekenen van de gezochte parameters is zeer rekenintensief. In de meeste gevallen is er sprake van een groot aantal vaardigheidsparameters, omdat er voor een betrouwbaar resultaat een grote steekproefomvang vereist is. En het blijkt dat de parameters vaak niet tegelijk en rechtstreeks te berekenen zijn, maar successievelijk moeten worden benaderd.

Tijdens de Case Study (hoofdstuk 4) in dit werkstuk is getracht om voor een kleine overzichtelijke situatie een model op te stellen. Voor het schatten van de parameters is gebruik gemaakt van het JML-model, beschreven in paragraaf 3.2.2, waarbij de persoonsparameters en de itemparameters gelijktijdig geschat worden. Deze manier van parameterschatten was hier goed uitvoerbaar, omdat de dataset klein was. Uiteindelijk zijn er twee eenvoudige formules voor de twee parameters naar voren gekomen, zie paragraaf 4.4.1. Verder is het mogelijk gebleken om uitspraken te doen over de betrouwbaarheid van de te schatten parameter(s) en kunnen er zowel met behulp van het Rasch-model als zonder dit model hypothesen getoetst worden.

Verder onderzoek

Het onderwerp testtheorieën is zeer divers. Vanwege de beperkte tijd is er in dit BWI-werkstuk alleen gekeken naar het Rasch-model in de itemresponstheorie. Er zijn nog meerdere modellen bekend binnen deze theorie en er zijn ook nog andere testtheorieën die ook zeker het onderzoeken waard zijn.

In dit werkstuk is er gewerkt met een kleine vereenvoudigde situatie om een model op te stellen. Het verdient aanbeveling om de situaties waarvoor een model moet worden opgesteld uit te breiden. Dit betekent meer items en meer personen. Het model wordt dan multidimensionaal. Dit zal meer rekentijd gaan kosten, maar hetzelfde principe als in hoofdstuk 4, van de meest aannemelijke schatter, kan worden toegepast. Er wordt nu echter meer dimensionaal geoptimaliseerd. Verder is het ook interessant om de overige schatters die beschreven staan in hoofdstuk 3 (CML- en MML-schatter) te gaan gebruiken in praktijksituaties door ze zelf te implementeren.

In de modellen in dit werkstuk is gewerkt met dichotome (0-1) items. Het gebruik van dichotome items in een toets komt veel voor, maar open vragen bij toetsen zijn ook heel populair. Een aanpassing van het Rasch-model zal er ongetwijfeld toe leiden dat ook voor andere typen items het Rasch-model zal kunnen worden toegepast. Onderzoek naar modellen voor andere typen items zal zeker een uitdaging zijn.

6 Literatuurlijst

(2002), Psychometrisch Onderzoekscentrum, Citogroep Arnhem.
<*http://www.cito.nl/pok/poc/eind_fr.htm*>

Eggen, T.J.H.M., en Sanders, P.F. (1993), Psychometrie in de praktijk, Citogroep, Arnhem, hoofdstuk 4: *Itemresponstheorie* (Verhelst, N.D.)

Larsen, Richard J., en Marx, Morris L. (2001), *An introduction to Mathematical Statistics and Its Applications*, Prentice Hall, pagina 505-508

Oosterhoff, J., en Vaart, A.W. van der (2000), *Algemene Statistiek*. Collegedictaat, *Vrije* Universiteit Amsterdam, pagina 13.

Verhelst, N.D. (1992), *Het Eenparameter Logistisch Model (OPLM)* - Een theoretische inleiding en een handleiding bij het computerprogramma, Cito Arnhem, pagina 1-19.

Wierstra, R.F.A. (2000), *Toets-en item-analyse in onderwijsonderzoek*. Collegedictaat, Universiteit Utrecht, Utrecht, pagina 3.

BIJLAGEN

Bijlage A: functies (I)

```
nulpunt
function(m, i, beta, startwaarde)
{
	t0 <- startwaarde
	t1 <- t0 - f(m, i, t0, beta)/afgeleide(t0, beta)
	while(abs(f(m, i, t0, beta)) > 1e-05) {
		t0 <- t1
		t1 <- t0 - f(m, i, t0, beta)/afgeleide(t0, beta)
	}
	t1
}


f
function(m, i, theta, beta)
{
	hulp <- 0
	for(j in 1:ncol(m)) {
		hulp <- hulp + m[i, j]
	}
	hulp - 1/(1 + exp( - theta)) - 1/(1 + exp( - theta + beta))
}


afgeleide
function(t, beta)
{
 - exp( - t)/((1 + exp( - t)) * (1 + exp( - t))) - exp( - t + beta)/((1 + exp(
		 - t + beta)) * (1 + exp( - t + beta)))
}
```

```
likelihood
function(m, beta, startwaarde)
{
	l <- 0
	for(i in 1:nrow(m)) {
		totaal <- 0
		for(j in 1:ncol(m)) {
			totaal <- totaal + m[i, j]
		}
		if((totaal > 0) && (totaal < ncol(m))) {
			n <- nulpunt(m, i, beta, startwaarde)
			l <- l + m[i, 1] * n - log(1 + exp(n)) + m[i, 2] * (
				n - beta) - log(1 + exp(n - beta))
		}
	}
	l
}
```

Bijlage B: functies (II)

```
bsr
function(m)
{
	bet <- mtobeta(m)
	B <- 10000
	res <- numeric(B)
	for(i in 1:B) {
		k <- rmat(m, bet)
		res[i] <- mtobeta(k)
	}
	res
}


mtobeta
function(m)
{
	s1 <- 0
	s2 <- 0
	for(i in 1:nrow(m)) {
		if(m[i, 1] + m[i, 2] == 1) {
			if(m[i, 1] == 1) {
				s1 <- s1 + 1
			}
			else s2 <- s2 + 1
		}
	}
	if((s1>0) && (s2 > 0)) {
		beta <- 2 * log(s1/s2)
		beta
	}
	else {
		beta <- 100
		beta
	}
}
```

```
rmat
function(m, bet)
{
	k <- matrix(0, nrow(m), ncol(m))
	k[, 1] <- rbinom(nrow(m), 1, exp(bet/2)/(1 + exp(bet/2)))
	k[, 2] <- rbinom(nrow(m), 1, exp( - bet/2)/(1 + exp( - bet/2)))
	for(i in 1:nrow(m)) {
		if(m[i, 1] + m[i, 2] == 2) {
			k[i,  ] <- 1
		}
		else if(m[i, 1] + m[i, 2] == 0) {
			k[i,  ] <- 0
		}
	}
	k
}


sv
function(v)
{
	res <- numeric(aantal(v))
	j <- 1
	for(i in 1:length(v)) {
		if(v[i] < 100) {
			res[j] <- v[i]
			j <- j + 1
		}
	}
	res
}
```

Bijlage C: uitslagen toets

Stnr.	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	0	1	1	0	1	1	1
2	1	1	1	1	0	1	0	0	1	1	1
3	0	0	1	1	0	1	0	0	1	1	0
4	1	0	1	1	0	1	1	0	1	1	0
5	1	1	1	1	0	1	1	0	1	1	0
6	0	0	0	1	0	1	0	0	1	1	0
7	1	1	1	1	0	1	1	0	1	1	0
8	1	0	0	1	0	1	0	0	1	1	0
9	0	1	0	0	0	0	0	0	0	0	0
10	0	0	0	1	0	0	0	0	1	1	1
11	1	1	0	1	0	0	0	0	0	1	1
12	1	0	0	0	0	0	0	0	1	1	0
13	0	1	0	1	0	0	0	0	0	1	0
14	1	1	1	1	1	1	0	1	1	1	1
15	1	0	1	1	0	1	1	0	0	1	1
16	1	0	0	1	0	0	0	0	1	1	0
17	0	0	1	1	0	1	1	0	1	1	1
18	1	1	0	1	0	0	0	1	1	1	0
19	0	0	1	1	0	0	0	0	0	1	1
20	0	1	1	1	0	1	1	1	1	1	1
21	1	1	1	1	0	1	0	0	1	1	0
22	0	0	1	1	0	0	0	0	0	1	0
23	1	1	1	1	0	0	1	1	0	1	1
24	0	0	1	1	0	1	0	0	0	0	0
25	1	1	0	1	0	1	1	1	1	1	0
26	0	0	1	1	0	0	0	0	0	1	0
27	0	1	1	1	0	1	1	1	1	1	1
28	0	1	0	1	0	1	0	0	0	1	0
29	0	0	1	1	0	1	0	0	1	1	1
30	1	1	0	1	0	1	1	0	0	1	1
31	0	1	1	1	1	1	0	0	1	1	0
32	1	1	1	0	0	1	1	0	1	1	1
33	0	0	1	0	0	0	0	0	1	1	1

Tabel 2. Antwoorden op toets. Bron: Van Os, S (2002), *Gecijferdheid beïnvloed,* Samenhang tussen kenmerken en prestaties bij gecijferdheid van eerstejaars pabostudenten.