Onderzoeksmethoden

Schaalmethoden

Versie 1.2

Jaap Huinink	s0276197	J.Huinink@student.kun.nl
Johan Stortelder	s0355593	JohanStortelder@student.kun.nl

Inhoud

Inleiding

Iedereen die wel eens een enquête heeft ingevuld, heeft wel eens een schaal gezien. En iedereen die wel eens een enquête heeft ingevuld, heeft zich waarschijnlijk ook wel eens fronsend afgevraagd of het nu echt nodig is om zoveel vragen te stellen.

Je kunt je ook afvragen wat er nu zo moeilijk kan zijn aan een uitspraak waarop “eens” en “oneens” (en alles wat daar tussen zit) geantwoord kan worden en of er wel zoveel te vertellen valt over schaalmethoden.

Het antwoord hierop is een eenduidig “Ja”; er valt een heleboel te vertellen over schaalmodellen, zoveel zelfs dat dit dictaatje slechts een summiere (en derhalve onvolledige) beschrijving geeft van alles wat er over schaalmethoden valt te vertellen. Toch wordt in dit dictaat geprobeerd een zo volledig mogelijk overzicht te geven van de verschillende soorten methoden, gesorteerd op eigenschappen van die methoden.
Ook zullen er een aantal expliciete voorbeelden worden gegeven van (in het verleden) veel gebruikte schaalmethoden.

Waarom schalen?

Hierboven wordt de vraag gesteld of het bij een enquête nu echt nodig is om zoveel vragen te stellen. Wanneer je de gebruikerssatisfactie van een informatiesysteem wilt testen zou je dat ook gewoon kunnen vragen:
“Wat vindt u van het systeem waarmee u werkt?”
Hierop kan de respondent dan met een kort of lang antwoord reageren.
Wanneer je het iets meer gaat structureren, kun je met de volgende antwoordmogelijkheden komen:

heel slecht——————————————————————— *heel goed*

waarbij een kruisje op de lijn de mate van tevredenheid aangeeft.
Wanneer je het nog meer gaat structureren, kun je met hokjes gaan werken:

☐	☐	☐	☐	☐
aftands	vrij slecht	redelijk	goed	perfect

Het ligt voor de hand te denken dat de onderzoeker zichzelf (en de respondent) onnodig veel werk op de hals haalt door bovenstaande vraag op te delen in 20 of meer uitspraken.
Toch is dat niet het geval, door een uitgebreide schaal te construeren zorgt de onderzoeker er voor dat hij een veel eenduidiger en betrouwbaarder antwoord krijgt. Wij geven hiertoe drie argumenten:

- Wanneer je mensen slechts één algemene vraag stelt (“Wat vindt u van het systeem waarmee u werkt?”), ligt het in de lijn der verwachting dat de verschillende respondenten bij het beantwoorden van die vraag aan verschillende aspecten (in dit geval non-functionele requirements) denken. De één denkt hierbij wellicht aan de betrouwbaarheid van het systeem terwijl de ander de vraag meer betrekt op de gebruiksvriendelijkheid. Gevolg hiervan is dat mensen die over het algemeen hetzelfde denken over het systeem een totaal ander antwoord geven, terwijl mensen die misschien een ander mening hebben over het systeem op het zelfde antwoord uitkomen. M.a.w. het verschijnsel waarover men zijn mening moet geven is zo heterogeen dat dit niet in één vraag te dekken valt (Swanborn, 1993).
- Wanneer je slechts één vraag stelt, spelen toevalsinvloeden (respondent kan even afgeleid zijn, emotionele stemming van de respondent, slecht begrijpen van een vraag) een veel grotere rol. Wanneer echter meerdere vragen worden

gesteld zullen deze invloeden weggemiddeld worden. “Het gebruik van meerdere items verhoogt dus in principe de validiteit” (Segers, 2002).
- Je kunt bij het stellen van meerdere vragen een veel preciezer onderscheid maken tussen de respondenten, simpelweg omdat er meer deelverzamelingen geconstrueerd kunnen worden.

We zijn het er nu dus over eens dat een antwoord op een enkele vraag enig wantrouwen op zou moeten roepen en dat het gebruik van schalen dit wantrouwen weg kan nemen. Als we een kenmerk of “begrip” willen meten, doen we er verstandig aan dit begrip meervoudig te operationaliseren: we formuleren een aantal items die samen dat begrip zo goed mogelijk dekken.

Wat zijn schaalmethoden?

De vraag die nu rest is “wat wordt er nu precies bedoelt met schalen, en wat is een schaalmethode?”. De laatste zin van de vorige paragraaf omschrijft in feite al gedeeltelijk wat schalen inhoudt, toch willen we wat meer duidelijkheid geven.
Het antwoord op de vraag “wat is schalen?” is vierledig; schalen heeft in feite vier betekenissen:

1. *Het aantal mogelijke antwoorden per item* die de respondent tot zijn beschikking heeft (bijvoorbeeld lopend van “aftands” tot “perfect”, zie boven).
2. *Het totale bereik aan mogelijke scores* dat op de betreffende set items als totaal kan worden behaald (mede afhankelijk van de manier waarop de antwoorden op de diverse items worden gecombineerd).
3. *De set items* die tezamen een (voldoende) betrouwbare en valide schaal blijken te vormen. De set items waarmee een begrip gemeten wordt, wordt vaak een schaal genoemd. Dat doen we echter pas als die set blijkt te voldoen aan de eisen van een bepaalde schaalmethode. We spreken dan bijvoorbeeld van een Likertschaal (zie verderop in dit dictaat). Als we de items samen een schaal noemen, bedoelen we met schaal dus een meetinstrument, dat aan bepaalde kwaliteitseisen voldoet.
4. Alle voorgaande betekenissen aangevuld met de *procedures* en *criteria* die moeten worden gehanteerd om tot schaalwaarden te komen.

Over het algemeen geldt dat de verschillende betekenissen van “schalen” in principe geen problemen hoeft op te leveren, meestal blijkt uit de context wel welke betekenis bedoeld wordt, wanneer dat niet het geval zullen wij dat expliciet aangeven.
D.m.v. schaalconstructie komen we aan een set items die samen een schaal vormen. Met schaalconstructie wordt het genereren van items (op basis van kennis van het te maten domein) en het nagaan of de items samen voldoen aan de eisen van het gekozen schaalmodel bedoeld.
Een schaalmodel is eigenlijk een minitheorie omtrent de te meten eigenschap. Het wijkt in feite niet af van de betekenis van modelleren die we al kennen: Het in kaart brengen van een bepaald subject of eigenschap. Vervolgens moeten aan de objecten (die ontstaan als gevolg van het modelleren) op een correcte manier schaalwaarden worden toegekend (Mokken, 1971).
Dan nu het antwoord op de vraag “wat is een schaalmethode?”:
Een schaalmethode bevat een theoretisch model, een stel procedures, en een set criteria. Een model is bijvoorbeeld: antwoorden op kennistoetsvragen kunnen worden opgeteld tot een zinvolle totaalscore; het betreffende kennisdomein wordt kennelijk opgevat als één geheel van gelijkwaardige onderdelen. De procedures hebben betrekking op de constructie van het meetinstrument, de wijze waarop moet worden nagegaan of aan de criteria wordt voldaan, de wijze waarop de gegevens worden verzameld, en de wijze waarop de schaalwaarden moeten worden toegekend. Een criterium kan bijvoorbeeld zijn: voorkeursordeningen moeten consistent zijn (transitiveit; als A>B en B>C dan A>C).

Schaalmethoden; een introductie

In dit hoofdstuk zullen Wij de belangrijkste achterliggende theorieën en begrippen die van belang zijn bij schaalmethoden, bespreken. Het vormt een verdere uitbreiding van wat in de inleiding besproken is en tevens zorgt het voor een kennisbasis m.b.t. de voorbeelden die in het volgende hoofdstuk aan bod zullen komen.

Representerend tegenover indicerend meten

Wanneer we spreken over schaalmethoden is het belangrijk een onderscheid te maken in de manier van meten die gebruikt wordt in een bepaalde schaalmethode. Wanneer je weet welke manier van meten in een specifiek geval gebruik wordt kun je zo achterhalen welke schaalmethode(n) hiervoor geschikt zijn en welke afvallen. We maken hier onderscheid tussen twee manieren van meten:
Bij representerend meten bestaat er een heen-en-weer relatie tussen het empirische en het numerieke stelsel; wanneer we een aantal reacties c.q. antwoorden op stellingen / vragen hebben, kunnen we op grond hiervan voorspellingen doen over de overige uitslagen. In praktijk: Wanneer je niet weet hoe zwaar Kees is, maar wel dat hij zwaarder is dan Daan, en wanneer je niet weet dat hoe zwaar Jan is, maar wel dat hij zwaarder is dan Kees, dan kun je daaruit logischerwijs concluderen dat Jan zwaarder is dan Daan.
M.a.w., “bij representerend meten is er toetsing mogelijk op de correctheid van de afbeelding van het empirische stelsel” (Swanborn, 1993).
In het geval van indicerend meten is de relatie tussen het empirische stelsel en het numerieke stelsel niet volledig. Een voorbeeld hiervan is bijvoorbeeld de stand in de Eredivisie in het betaald voetbal. Het feit dat PSV op de eerste plek staat suggereert dat PSV van alle tegenstanders zou winnen. Toch weten we dat dit niet het geval is. PSV zal weliswaar de meeste wedstrijden winnen, maar het is ook mogelijk dat het verliest van het veel lager geklasseerde NEC. Kortom er bestaat geen één op één relatie tussen het empirisch- en numerieke stelsel, om de doodeenvoudige reden dat het empirische stelsel niet (geheel) bekend is. Wanneer Louis van Gaal aldus stelt dat AJAX de beste club van Nederland, Europa en de Wereld is, verward hij het empirisch- en het numeriek stelsel en doet hij een niet gestaafde uitspraak omdat het hier immers een indicerende meting betreft.
Bij indicerend meten wordt de keuze van getallen voornamelijk bepaald door hun praktische bruikbaarheid. Bij representerend meten daarentegen bepaald het meetniveau de keuze van de getallen en de hoeveelheid toelaatbare algebraïsche transformaties.
Om te voorkomen dat we te ver afwijken van het onderwerp, zullen we hieronder slecht ter verduidelijking een korte samenvatting geven van de vier te onderscheiden meetniveau.

- **nominaal niveau**: Het antwoord van de respondent valt in één categorie, van een aantal categorieën die kwalitatief van elkaar verschillen. Bijvoorbeeld vragen over een politieke partij: men behoort tot een bepaalde categorie of niet, meer valt er, meettechnisch, niet van te zeggen.
- **ordinaal niveau:** De vraag betreft een begrip of kenmerk waarbij sprake is van meer of minder; een voorbeeld is de mate van tevredenheid: de ene persoon is ‘meer’ tevreden dan de andere persoon.
- **interval niveau**: Ook hier is sprake van meer-minder, maar nu meer precies: de afstanden tussen de antwoordmogelijkheden (‘schaalpunten’) zijn even groot. Een bekend voorbeeld is de meting van temperatuur in bijv. Graden Celsius: het verschil in warmte tussen 5 en 10 Graden is (objectief gesproken) even groot als dat tussen 10 en 15 graad.
- **rationiveau (verhoudingsschaal)**: Dit betreft vragen naar zaken waarbij sprake is van een zogeheten ‘absoluut nulpunt’. Dat is bij temperatuurmeting in graden Celsius zoals bekend niet het geval, men kan niet zeggen dat 10 graad Celsius tweemaal zo warm is als 5 graad Celsius. Een voorbeeld waarbij sprake is van meten op rationiveau is het meten van snelheid: het is zinvol om te spreken over een snelheid

van 0, en een snelheid van 10 is bovendien inderdaad tweemaal zo snel als een snelheid van 5.

Wanneer we in de volgende paragraven verschillende schaalmethoden bespreken, zullen we hierin aangeven of er gebruikt wordt gemaakt van representerend dan wel indicerend meten. In het geval van representerend meten zullen we tevens aangeven voor welk meetniveau de schaalmethode geschikt is.

Enkele begrippen

Wanneer je met behulp van schaalmethoden een verschijnsel probeert te meten, ontstaan er een hoop verschillende uitkomsten (per respondent).
De mogelijke uitkomsten die uit bijvoorbeeld een enquête kunnen komen, kun je uitzetten op een gegevensruimte, een continuüm. In het geval van de gebruikerssatisfactie die in voorgaande paragraven aan de orde kwam, kan dit lopen van aftands tot perfect. Over het algemeen loopt een continuüm van zeer negatief tot zeer positief. De meningen van de verschillende respondenten zijn op een bepaalde manier verdeeld over dit continuüm dat voor te stellen is als een rechte lijn. Hierbij valt te denken aan een normale verdeling of juist een sterk gepolariseerde verdeling.
In eerdere paragraven kwam het begrip *item* al een paar keer aan bod. Een item is eigenlijk niets meer dan een uitspraak over het te meten verschijnsel. Wanneer je naast verbale uitspraken ook nog foto´s of andere modaliteiten als uitspraak over een verschijnsel opneemt, dan noemen we dit *stimuli*.
Net als respondenten kunnen ook items en andere stimuli uitgezet worden op een continuüm.
Er zijn sterk negatieve, “neutrale” en sterk positieve stimuli. De uitspraak

“De userinterface is volstrekt onlogisch”

is op het continuüm “tevredenheid van de gebruiker t.o.v. een het informatiesysteem” te kwalificeren als een erg negatief item, terwijl een item als

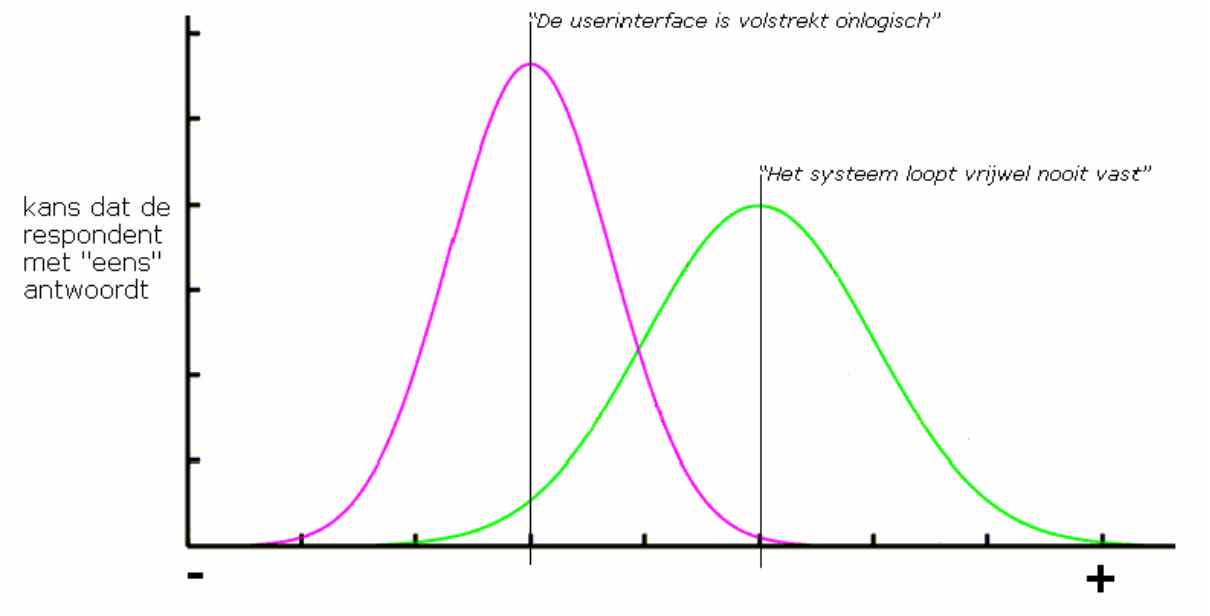
“Het systeem loopt vrijwel nooit vast”

een veel positievere uitspraak over de werking van een systeem is.
Het continuüm wordt nu dus een gemeenschappelijke ruimte van respondenten en stimuli. In dit dictaat zullen wij die ruimte als een ééndimensionale beschouwen. Natuurlijk ligt het voor de hand dat ingewikkelde houdingen van mensen ten opzichte van een verschijnsel in sommige gevallen beter gerepresenteerd wordt wanneer de verschillende stimuli worden uitgezet op meerdere dimensies. *Multidimensionale schaalmethoden* wijken echter behoorlijk af van de *ééndimensionale schaalmethoden* die wij in dit dictaat bespreken en vallen derhalve buiten de scope van dit dictaat.
Voor meer informatie over Multidimensionale schaalmethoden verwijzen wij naar (Kruskal, J.B., W. Myron (1978)).

Itemkarakteristieken

Het is nu interessant om te bekijken hoe onze respondenten en stimuli zich tot elkaar verhouden in een ééndimensionale ruimte. M.a.w. welke respondenten op welke vraag positief antwoorden, en welke (op welke) negatief (Swanborn, 1993). Dit laatste kan uitéengezet worden in zogenaamde itemkarakteristieken. In de rest van deze paragraaf zullen we aan de hand van een eenvoudig model proberen uit te leggen wat een itemkarakteristiek is.
In dit model worden twee bovenstaande stimuli (“De userinterface is volstrekt onlogisch” en “Het systeem loopt vrijwel nooit vast”) voorgelegd aan de respondenten. Voor het gemak gaan we er nu even van uit dat de respondenten slechts kunnen antwoorden met “eens” en “oneens”. Elke respondent bevindt zich op een bepaalde plek in het continuüm en datzelfde geld voor beide items, wat we nu willen uitdrukken is de kans dat een respondent het eens is met een stimulus (item), gegeven hun beider positie op het continuüm.

In figuur 1, die hieronder te vinden is de positie van het eerste item op het continuüm aangegeven met de eerste verticale lijn. We zien dat dit item vrij negatief is.



Figuur1: Items op hun continuüm

Wat ook opvalt is dat de roze grafiek er één is met een tamelijk spitse punt. Dit laatste geeft aan dat de respondenten die een vergelijkbare positie op het continuüm innemen als het item zelf, een grote kans hebben het met de stelling (het item) eens te zijn. Respondenten die een positie innemen die een stuk positiever (maar ook negatiever!) is veel minder snel geneigd zijn om het eens te zijn met de stelling. Er bestaat in het geval van dit item aldus een duidelijk verband tussen de positie op het continuüm van de respondent en het item. Dit verband nu noemen we de itemkarakteristiek, in figuur1 zijn deze karakteristieken afgebeeld als een roze en een groene lijn.
Voor alle duidelijkheid, voor het item “Het systeem loopt vrijwel nooit vast” bestaat een minder eenduidig verband voor de posities van de respondent en het item op het continuüm. Te zien is dat dit item een veel positievere houding heeft. Er zijn echter (vergeleken met het vorige item), veel meer respondenten die een afwijkende positie op het continuüm innemen, maar het toch eens zijn met de stelling. Hieruit valt de conclusie te trekken dat dit tweede item een minder een mindere scherprechter is met betrekking tot het bepalen van de positie van de respondent op het continuüm.

Bekende schaalmethoden

Er zijn verschillende soorten eendimensionale schaalmethoden. Hieronder volgt eerst een overzicht van de schaalmethoden die door de jaren heen de meeste bekendheid hebben gekregen. Vervolgens wordt kort ingegaan op de klassieke testtheorie, die tot op heden de meest gebruikte theorie is als het gaat om de ontwikkeling van gestandaardiseerde psychologische tests en toetsen.
Moderne schaalmethoden, die tot slot worden behandeld, zijn gebaseerd op de itemresponse theorie. Ze vormen een modern alternatief voor de schaalmethoden die van de klassieke testtheorie uitgaan.
In de loop der jaren zijn er een heleboel schaalmethoden ontwikkeld. Deze methoden hebben niet allemaal dezelfde functie.
Er zijn methoden die zich alleen richten op de constructie van een schaal, zoals de Thurstone-methode. In een vooronderzoek worden met behulp van beoordelaars en bepaalde analyses aan items getalswaarden toegekend. Hierna kan de schaal bij respondenten gebruikt te worden, zoals hieronder uitgebreider beschreven wordt in het gedeelte over Thurstone-schalen.
Andere methoden combineren het construeren van een schaal, het controleren of aan de criteria waaraan de schaal moet voldoen wordt voldaan, en het toekennen van schaalwaarden aan respondenten. De verschillende stappen gebeuren in een keer. Dit gebeurt bij Likert-, Guttman- en Coomb-schalen. De antwoorden van respondenten worden bij deze methoden voor meerdere doeleinden gebruikt. Ook deze schaalmethoden worden verderop nader belicht. Een uitgebreidere beschrijving van deze methoden is te vinden in (Dunn-Rankin, 2004), (Swanborn, 1993) en (Segers, 2002).

Thurstone-schalen

Representerend/indicerend meten	Representerend
Meetniveau	Verschilt
Onderscheid constructie- en toepassingsfase	Aparte fasen
Aantal stimuli	7 tot 10

Tabel 1: Overzicht karakteristieken Thurstone

Zoals eerder al vermeld bestaat de Thurstone methode uit verschillende fasen. In de eerste fase, wordt de schaal geconstrueerd voor een bepaald theoretisch kader. In de responsfase, die compleet gescheiden van de eerste fase, wordt respondenten daadwerkelijk gevraagd op de schaal aan te geven welke plek op de schaal (stimulus) het meest overeen komt met hun mening.
In 1927 ontwikkelde Thurstone zijn ‘Law of comparative judgement’. Het uitgangspunt is daarbij dat psychologische beoordelingen vergelijkbaar zijn met beoordelingen van fysische grootheden. Als het waargenomen verschil tussen twee stimuli groter is, dan is ook het percentage respondenten dat de stimuli verder uit elkaar plaatst op een continuüm groter. Uit de plaatsingen op dat continuüm, bijvoorbeeld kruisjes op een lijn, kan vervolgens een schaal worden afgeleid.
In de eerste fase wordt de respondenten voor die fase, de jury, gevraagd om op bepaalde stimuli te reageren door middel van een plaatsing op een theoretisch continuüm. Aan een groep studenten kan bijvoorbeeld gevraagd worden hoe groot ze het nut van een bepaald vak vinden. Dit kan dan variëren van totaal nutteloos tot heel erg nuttig, de twee uitersten van het continuüm. Alle studenten geven hun mening. Het gemiddelde van die meningen resulteert in de positie van een stimulus (in het voorbeeld: heel erg nuttig, redelijk nuttig, totaal nutteloos) op het continuüm.
De op die manier verkregen schaal kan in de responsfase gebruikt worden om de mening van de eigenlijke onderzoekselementen te toetsen.
Door het scheiden van de twee fases en doordat gebruik gemaakt wordt van een 100 koppige jury, is de Thurstone methode erg bewerkelijk.

Likert-schalen

Representerend/indicerend meten	Indicerend
Meetniveau	N.v.t.
Onderscheid constructie- en toepassingsfase	Geen onderscheid
Aantal stimuli	20 tot 25

Tabel 2: Overzicht karakteristieken Likert

De methode die door Likert is bedacht heeft geen aparte beoordelings- en responsfase. Er worden door de onderzoeker bedachte uitspraken aan de respondent voorgelegd. Deze kan hierbij aangeven in hoeverre hij of zij het met deze uitspraken eens is. De respondent kan kiezen uit een ordinale reeks, bestaande uit 3 tot 7 meningen ten aanzien vrij extreem geformuleerde stellingen (bijvoorbeeld: eens, eens noch oneens, oneens).
Om respondenten scherp te houden worden de stellingen over eenzelfde onderwerp afwisselend positief en negatief geformuleerd. Respondenten vallen hierdoor minder snel in een soort automatisme. De mening van elke respondent wordt echter verondersteld hetzelfde te blijven. Als de extremiteit van de vraagstelling dan verandert, zal een respondent ook voor een andere antwoordcategorie kiezen. Voor iedere uitspraak zijn daarom aparte schaalwaarden nodig. Er wordt vanuit gegaan dat de antwoorden van de respondenten normaal verdeeld zijn. Door de antwoorden op de stellingen te normaliseren, en door te sommeren over alle genormaliseerde antwoorden over eenzelfde onderwerp, wordt een persoonsscore van een respondent verkregen. Deze persoonsscore geeft aan hoe positief of negatief deze tegenover dat onderwerp staat.
Het is belangrijk dat uitspraken sterk geformuleerd worden. Het gebruik van zwakke of neutrale uitspraken kan problemen opleveren. Een vraag als “Heb je een deel van de colleges voor vak X gevolgd?” geeft bij een negatief antwoord twee mogelijkheden. “Nee, ik heb álle colleges gevolgd” en “Nee, ik heb geen énkel college gevolgd”. Het antwoord “nee” levert bij dergelijke vragen ambigue informatie op.

Guttman-schalen

Representerend/indicerend meten	Representerend
Meetniveau	Ordinaal niveau
Onderscheid constructie- en toepassingsfase	Geen onderscheid
Aantal stimuli	4 tot 7

Tabel 3: Overzicht karakteristieken Guttman

Guttman heeft bij het ontwikkelen van zijn schaalmethode verondersteld dat een respondent instemt met alle uitspraken die minder extreem zijn dan zijn eigen standpunt en niet instemt met extremere standpunten. Dit wordt een cumulatieve schaal genoemd. Om het standpunt van een respondent te bepalen worden die persoon tien tot twintig uitspraken voorgelegd die betrekking hebben op hetzelfde onderwerp. De respondenten kunnen kiezen tussen eens en oneens.
Vervolgens worden de standpunten gerangschikt op basis van het percentage respondenten dat het met dat standpunt eens was. Op die manier wordt een lijst verkregen die van minst extreem naar meest extreem loopt (volgens de respondenten). In het geval van een ideale cumulatieve schaal zijn zullen respondenten die het eens zijn met een standpunt het ook eens zijn met alle minder extreme standpunten.
In tabel 1, een voorbeeld uit Segers, J. (2002), is zo’n lijst met 6 stimuli opgenomen.

Items	% mee eens	schaalscore
Het allerergste dat men iemand kan aandoen is dat men hem/haar in zijn eer aantast	79%	1
Gehoorzaamheid en eerbied voor het gezag zijn de belangrijkste dingen die men kinderen moet leren	68%	2
Er zijn twee soorten mensen: sterken en zwakken	59%	3
De jeugd heeft vooral strene zelftucht nodig	55%	4
Er is maar een manier om werkelijk iets goed te doen	42%	5
Onze sociale problemen zouden grotendeels opgelost zijn als de misdadigere en associale elemeneten uit de samenleving zouden worden verwijderd	27%	6

Tabel 4: Voorbeeld van geordende lijst van standpunten

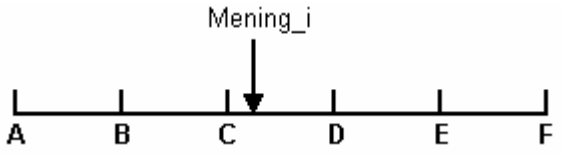
Iemand die vindt dat de jeugd meer zelftucht nodig heeft wordt dus ook verondersteld het met de daarboven gestelde standpunten eens te zijn. Als iemand het met drie standpunten eens is, is dat met de eerste drie. Antwoordpatronen als (1,1,1,0,0,0) zijn dus mogelijk, maar (0,1,1,0,0,1) bijvoorbeeld niet.

Coomb-schalen

Representerend/indicerend meten	Representerend
Meetniveau	Ordinaal niveau
Onderscheid constructie- en toepassingsfase	Geen onderscheid
Aantal stimuli	4 tot 10

Tabel 5: Overzicht karakteristieken Coomb

Het Coombs-model gaat uit van voorkeursdata. Stimuli (uitstpraken) moeten door respondenten gerangschikt worden van minst geprefereerd naar meest geprefereerd. De respondent krijgt alle stimuli in een keer te zien en kan met de stimuli gaan schuiven tot hij ze in de juiste volgorde heeft staan. Dit in tegenstelling tot de eerder genoemde methoden, waar de respondenten geacht worden achtereenvolgens op de verschillende stimuli te reageren. De voorkeursvolgorde van een respondent wordt zijn Individual-scale (I-scale) genoemd. Uit de verschillende Individual-scales wordt vervolgens geprobeerd een Joint-scale (J-scale) voor alle respondenten af te leiden. De mening van een respondent wordt op een lijn geplaats waarop ook de verschillende stimuli zijn aangegeven (Figuur 1), het zogenaamde ideaalpunt van de rsepondent.



Figuur 2: ideaalpunt respondent i t.o.v stimuli A t/m F

De voorkeursvolgorde van de respondent wordt verkregen door de stimuli op te sommen in volgorde van de afstand tot de mening. Voor respondent i uit het voorbeeld wordt de I-scale dus: CDBEAF.
Uit deze I-scales kan de J-scale worden afgeleid. A en F zijn de extreme stimuli. Respondenten met het ideaalpunt in een van die extremen hebben dus respectievelijk I-scale ABCDEF en FEDCBA. Hieruit blijkt de volgorde van de stimuli op de J-scale.
N.B. Dat de volgorde ABCDEF is lijkt triviaal vanwege de alfabetische volgorde van de letters. Bedenk echter dat de letters voor verschillende stimuli staan die waarschijnlijk niet alfabetisch geordend zijn.
Bij andere I-scales staat altijd stimulus A of F op de laatste plaats, omdat altijd een van de extremen op het eind staat. Als ook de volgorde van de stimuli en de onderlinge afstand bekend zijn, zijn er nog maar een beperkt aantal antwoordmogelijkheden.

Klassieke testtheorie

Een testtheorie is, zoals het woord al aangeeft, een theorie op basis waarvan tests ontwikkeld en gecontroleerd kunnen worden. Er bestaan verschillende testtheorieën. Een van de bekendste theorieën, waarop veel onderzoek is gebaseerd en waarvan nog steeds veel gebruik wordt gemaakt, is de klassieke testtheorie. Een andere veelgebruikte theorie is de wat modernere item respons theorie. Deze wordt belicht in het volgende hoofdstuk. Het sterke punt van de klassieke testtheorie is dat het de betrouwbaarheid van tests beoordeeld.
Als bij een respondent meerdere malen een tests wordt gedaan is het niet waarschijnlijk dat het resultaat iedere keer hetzelfde is. De testscore is dus niet een foutloze afbeelding van de kennis of vaardigheid van de geteste persoon. De klassieke testtheorie gaat ervan uit dat een score is opgebouwd uit een meetfout (E) en een ware score (W):

$X = W + E$

Hierbij worden twee aannames gedaan over de fout:

- De gemiddelde/verwachte fout moet gelijk zijn aan nul. Positieve en negatieve fouten heffen elkaar op.
- De waargenomen meetfouten hangen niet samen.

Het gaat natuurlijk om het achterhalen van de ware score van een respondent. Deze ware score is in theorie te achterhalen door de test een zeer groot aantal keren bij dezelfde persoon af te nemen onder exact dezelfde omstandigheden en van al de geobserveerde scores het gemiddelde te bepalen. In de praktijk daarentegen is dit niet uitvoerbaar. Het is al onmogelijk iemand twee keer onder exact dezelfde omstandigheden een test te laten doen, laat staan een groot aantal keren. In de klassieke testtheorie zijn er echter slimme manieren gevonden om met andere, wel uitvoerbare procedures, iets zinnigs te kunnen zeggen over de grootte van de meetfout.

Test-hertest methode

Bij deze methode wordt de test herhaald. Er wordt ervan uitgegaan dat als een respondent twee maal dezelfde test aflegt, dit tweemaal hetzelfde resultaat oplevert. Verschillen tussen de eerste en de tweede test worden dan als toevalsfluctuaties beschouwd.
Een nadeel van deze methode is alleen dat een proefpersoon in de tijd tussen de twee tests zijn te meten houding ten opzichte van iets aangepast kan hebben. Verschillen tussen de eerste en de tweede test zijn dan niet meer toe te schrijven aan het toeval (onbetrouwbaarheid), maar berusten op een verandering in het ware deel van de score.

Split-half methode

Een andere methode om de ware score te benaderen is de split-half methode. De test wordt in dit geval verdeeld in twee parallel lopende delen. Ook hier geldt dat er nadelen aan de methode kleven.
Een test kan op vele manieren in twee delen gesplitst worden en exacte parallelliteit is bijna niet realiseerbaar.

Cronbach's alpha

Een veelgebruikte maatstaf voor de betrouwbaarheid van een test is Cronbach's alpha. Het is een ondergrens voor de betrouwbaarheid. Cronbach's alpha kan gezien worden als de gemiddelde betrouwbaarheid van alle verschillende split-half tests die er met de items van die test mogelijk zijn. Hoe meer items de test heeft hoe hoger Cronbach's alpha en hoe hoger de minimale betrouwbaarheid van de test.

Item-response theorie

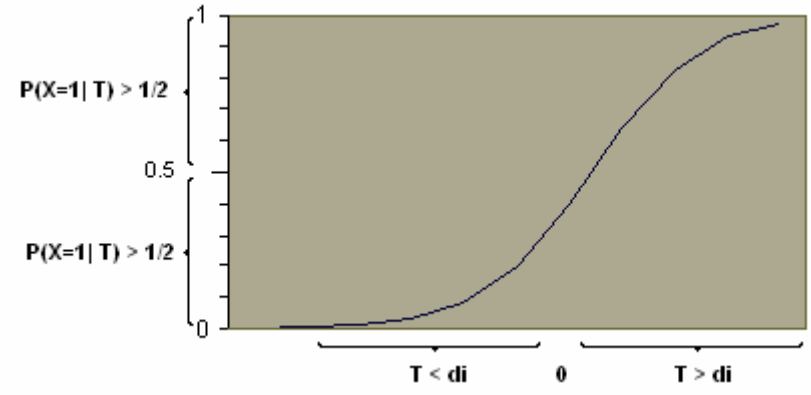
Waar de klassieke test theorie zich voornamelijk bezighoudt met de betrouwbaarheid en de standaardfout van de test, heeft de item-respons theorie (IRT) meer aandacht voor de niet zichtbare kenmerken van een respondent. De antwoorden van respondenten op de items van de test geven een beeld van een eigenschap van een persoon die niet direct te zien is. Op het eerste gezicht valt bijvoorbeeld niet te zien hoe intelligent iemand is, en dit is ook niet met directe vragen vast te stellen (“Ben jij intelligent?”). Met een IRT-test valt hier wel iets zinnigs over te zeggen, door de prestaties in bepaalde testsituaties te bekijken en deze als uitgangspunt te nemen om achterliggende eigenschappen te beschrijven.
Bij de analyse van gegevens wordt bij de item-response theorie gekeken naar het niveau van de items. De kans dat een respondent een goed antwoord geeft op een vraag is afhankelijk van de moeilijkheid van de vraag en de kennis van de proefpersoon. Een makkelijke vraag en een grote kennis bij de proefpersoon zorgen voor een hoge kans op een goed antwoord, een moeilijke vraag en lage kennis verhogen de kans op een fout antwoord. De items die voor het onderzoek gekozen worden, moeten derhalve in overeenstemming zijn met wat men van de proefpersonen wil meten. Als men de intelligentie van volwassenen wil meten is het vrijwel zinloos de proefpersonen te vragen een reeks makkelijke sommetjes te maken. Met deze items zou echter wel de rekenvaardigheid van kinderen op de lagere school gemeten kunnen worden. Het niveau van de items moet passen bij de verkregen testgegevens van de proefpersoon. Als dit niet het geval is moten de items verwijderd worden. Een bekend item-respons model is het model van Rasch. Dit wordt hieronder behandeld.

Model van Rasch

Volgens Rasch wordt het antwoord van de respondenten op een item bepaald door de bekwaamheid van de respondent en de moeilijkheid van het item. De moeilijkheid en de bekwaamheid worden respectievelijk uitgedrukt in de item-parameter en de bekwaamheidsparameter. De samenhang tussen deze parameters wordt gezien als een kansmodel. Hierdoor kan het ook voorkomen dat een er per ongeluk een fout antwoord wordt gegeven op een makkelijke vraag, bijvoorbeeld doordat de respondent afgeleid wordt, of dat er moeilijke vragen per ongeluk goed beantwoord worden doordat respondenten gokken. Er geldt:

- Persoonsbekwaamheid (T) < item-moeilijkheid (di) => P(correct antwoord) < 0.5
- Persoonsbekwaamheid (T) = item-moeilijkheid (di) => P(correct antwoord) = 0.5
- Persoonsbekwaamheid (T) > item-moeilijkheid (di) => P(correct antwoord) > 0.5

In figuur 3 is dit verband weergegeven.



Figuur 3: ideaalpunt respondent i t.o.v stimuli A t/m F

Uit de figuur blijkt ook dat gegeven de waarde van de vaardigheidsparameter, T, de kans om een moeilijker item goed te maken altijd kleiner is dan de kans een makkelijker item goed te maken.
Om te meten met volgens het model van Rasch moeten drie stappen worden doorlopen.

- Schatten van itemparameters
- Schatten van persoonsparameters
- Controleren van persoonsparameters

Het schatten van bovengenoemde parameters is niet eenvoudig. Het vergt veel rekenwerk en is niet praktisch uitvoerbaar zonder gebruik te maken van de rekenkracht van computers. Vroeger vormde dit een probleem. Door personal computers, nieuwe programmatuur en snellere processoren is dit probleem echter nu al achterhaald.
Naast het schatten van de parameters moet ook gecontroleerd worden of de waargenomen gegevens bij het model passen. Per item wordt gekeken of het bij het model past. Zo niet, dan wordt het, zoals eerder al gezegd, verwijderd.
Er zijn verschillende gebruiksmogelijkheden als er met de IRT gewerkt wordt:

- Zo kunnen de items verzameld worden in de zogenaamde itembanken. Dit zijn grote verzamelingen van items die dezelfde vaardigheid testen. Als iemand een onderzoek wil doen kan hij een aantal items uit een itembank met betrekking tot de vaardigheid van zijn onderzoek selecteren.
- Daarnaast is het mogelijk om computergestuurd adaptief te testen. Items kunnen zodanig geselecteerd worden dat ze dicht bij het vaardigheidsniveau van de respondent aansluiten.
- Er kan gekeken worden naar item-onzuiverheid. Hebben mensen met hetzelfde vaardigheidsniveau dezelfde antwoorden of zijn er verschillen tussen (groepen) mensen? Zijn er verschillen tussen bijvoorbeeld mannen en vrouwen of Engelstaligen en Nederlandstaligen.
- Tot slot is het gemakkelijk om afwijkende antwoorden te identificeren. Als iemand bijvoorbeeld veel fouten maakt onder zijn bekwaamheidsniveau valt dit direct op. Er kan dan onderzocht worden waar deze fouten vandaan komen.

Klassieke testtheorie tegenover IRT

Een groot verschil tussen de klassieke testtheorie en de item-response theorie is dat bij de klassieke testtheorie de score afhankelijk is van de extreemheid van het item, maar de extreemheid van het item niet afhankelijk van de persoon.
Daarnaast gaat het er bij de IRT vaak om eigenschappen meten en bij klassieke testtheorie vaak om een houding of mening ten opzichte van standpunten.
Doordat de items niet aangepast worden aan de bekwaamheden of standounten van de respondent is ook de gemiddelde meetfout groter bij de klassieke testtheorie. Bij de IRT kan dit bijgesteld worden doordat snel te detecteren is of een proefpersoon te makkelijke of te moeilijke vragen krijgt. Om iemands positie op een continuüm te bepalen kan er als het ware steeds meer ingezoomd worden op het stukje waar de proefpersoon zich bevindt. Dit kan gedaan worden door de items te veranderen.
Tot slot is bij de klassieke testtheorie het meetniveau niet empirisch verifieerbaar. Bij de IRT volgt het meetniveau uit de theorie. Het model van Rasch heeft metingen op tenminste interval niveau.

Nadelen

De IRT is echter wel complex, en het kost moeite om er vertrouwd mee te raken. Ook verwerpen IRT modellen meer items. Voor het schatten van de parameters zijn een heleboel respondenten nodig.

Casus

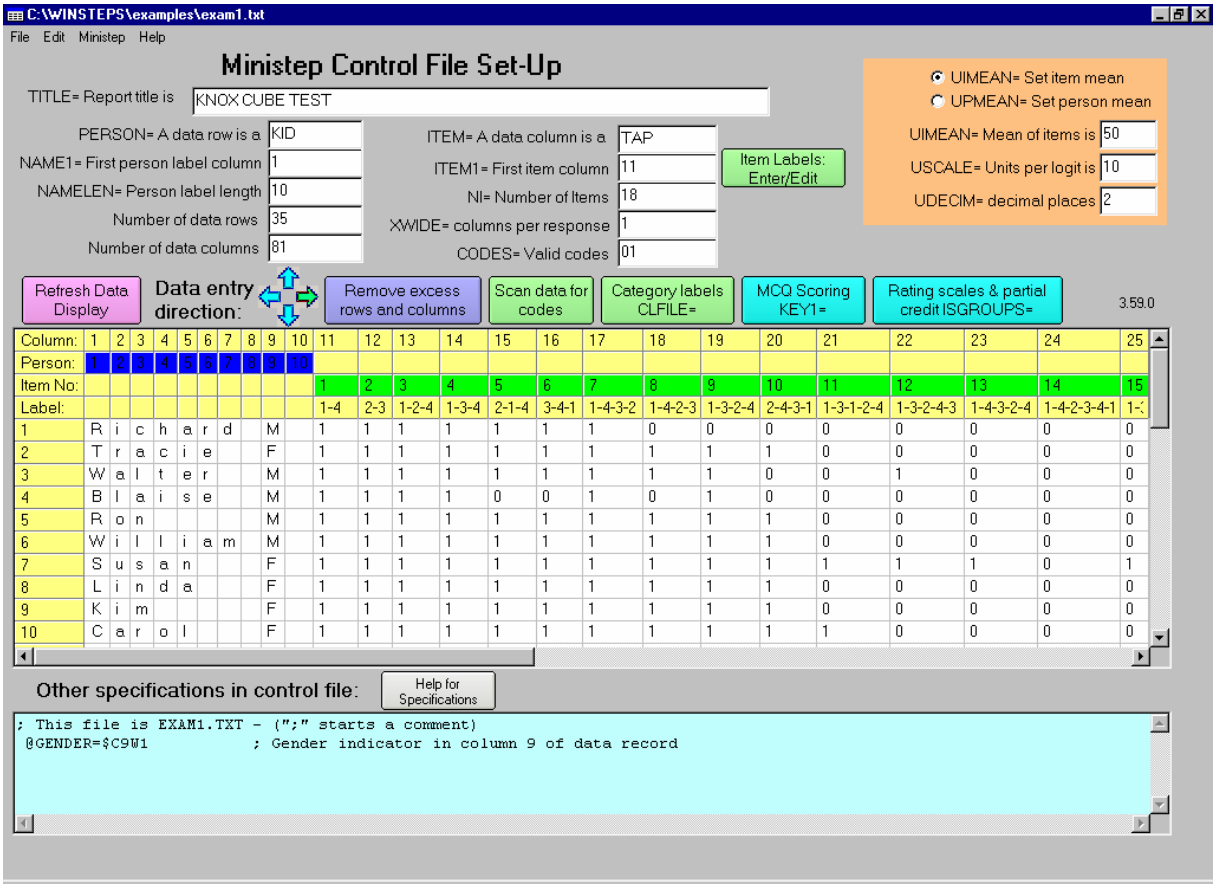
In deze casus willen wij een software product behandelen, waarin het werken met de Item Response theorie aan de orde komt. Na een zoektocht over het internet met behulp van Google, stuitten wij een op de theorie van Rasch gebaseerd software product, genaamd Winsteps. Dit is een professioneel product en derhalve niet vrij beschikbaar. Op de website http://www.winsteps.com vonden we echter al vrij snel een zogenaamd student evaluation version met de naam Ministep. Deze is te vinden en gratis down te loaden op http://www.winsteps.com/ministep.htm. Voor deze casus hebben wij derhalve gebruik gemaakt van deze gratis tool en bijgeleverde voorbeeld dataverzamelingen.

Winsteps

Zoals gezegd gebruikt Winsteps het model van Rasch iets te kunnen zeggen over de data sets die meestal bestaan uit personen en items. Er kunnen verschillende soorten scores per item gebruikt worden. Het programma kan omgaan met stimuli waar met goed/fout op gescoord kan worden. Maar ook op stimuli waarvoor multiple-choice antwoorden zijn opgesteld.
Er kunnen verschillende analyses en diagnoses op de data losgelaten worden. Ui t de dta kunnen grafieken gegenereerd worden, hetgeen de data inzichtelijker maakt en beter communiceerbaar naar eindgebruikers. Onverwachte data wordt ontdekt en aan de gebruiker aangegeven. Bijvoorbeeld wanneer iemand slecht scoort op peen makkelijke stimulus en moeilijkere weer wel goed heeft en er sprake is van een toevalsfoutje. Ook multidimesionaltiteit wordt ondersteunt. Winstep kan samenwerken met andere softwareproducten zoals MS Excel en SPSS.
Typische applicaties waarvoor Winsteps gebruikt kunnen worden zijn onderwijs toetsen, psychologische tests, assessments, onderzoek naar houdingen en meningen en het kalibreren van itembanken. Er kunnen tot 1,000,000 personen en tot 30,000 items verwerkt worden

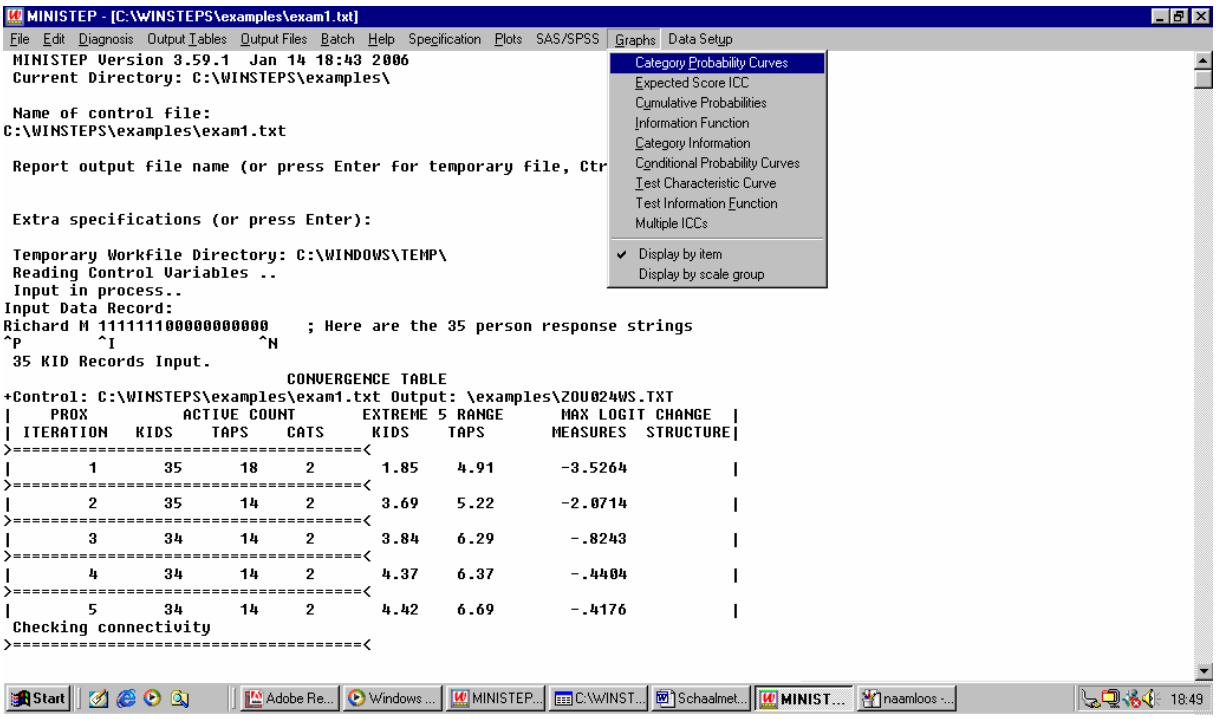
Voorbeeld dataset

De voorbeeld dataset is een standaard voorbeeld dat meegeleverd wordt met de demo software. De data in deze dataset bestaat uit respondenten, de stimuli en de scores van deze respondenten voor deze stimuli. In dit voorbeeld gaat het om de scores van respondenten voor de Knox cube test. Bij deze test krijgt iedere respondent een stimulus van een reeks van een bepaald aantal getallen en is het de bedoeling dat de respondent de getallenreeks foutloos natypt. Als de respondent de reeks foutloos reproduceert is de score “1”, anders “0”.
In figuur 3 is een screen-shot van het programma te zien. De eerste 10 velden van een rij zijn gereserveerd om een respondent weer te geven. Eerst zijn of haar naam, en vervolgens in kolom 9 het geslacht van de respondent. In het commentaar onderaan staat dit ook aangegeven bij de specificaties. In de kolommen 11 tot en met 81 (maar tot 24 zichtbaar) staan de verschillende scores voor de stimuli.
In de vierde rij van de eerste kolom staat “Label”. Hiermee kunnen de verschillende stimuli onderscheiden worden. In dit voorbeeld is in de label-rij voor iedere kolom opgenomen op welke getallenreeks er gereageerd moest worden. Goed te zien is dat de eerste stimuli, de makkelijke door iedereen foutloos worden nagebootst. Worden de stimuli echter moeilijker dan beginnen de eerste foutjes te komen. Ook de eerder beschreven “schoonheidsfoutjes” zijn terug te vinden in deze data. Sommige respondenten scoren niet op relatief makkelijke stimuli maar wel weer op moeilijkere. Zoals gezegd Kan dit door Winsteps aangegeven worden.



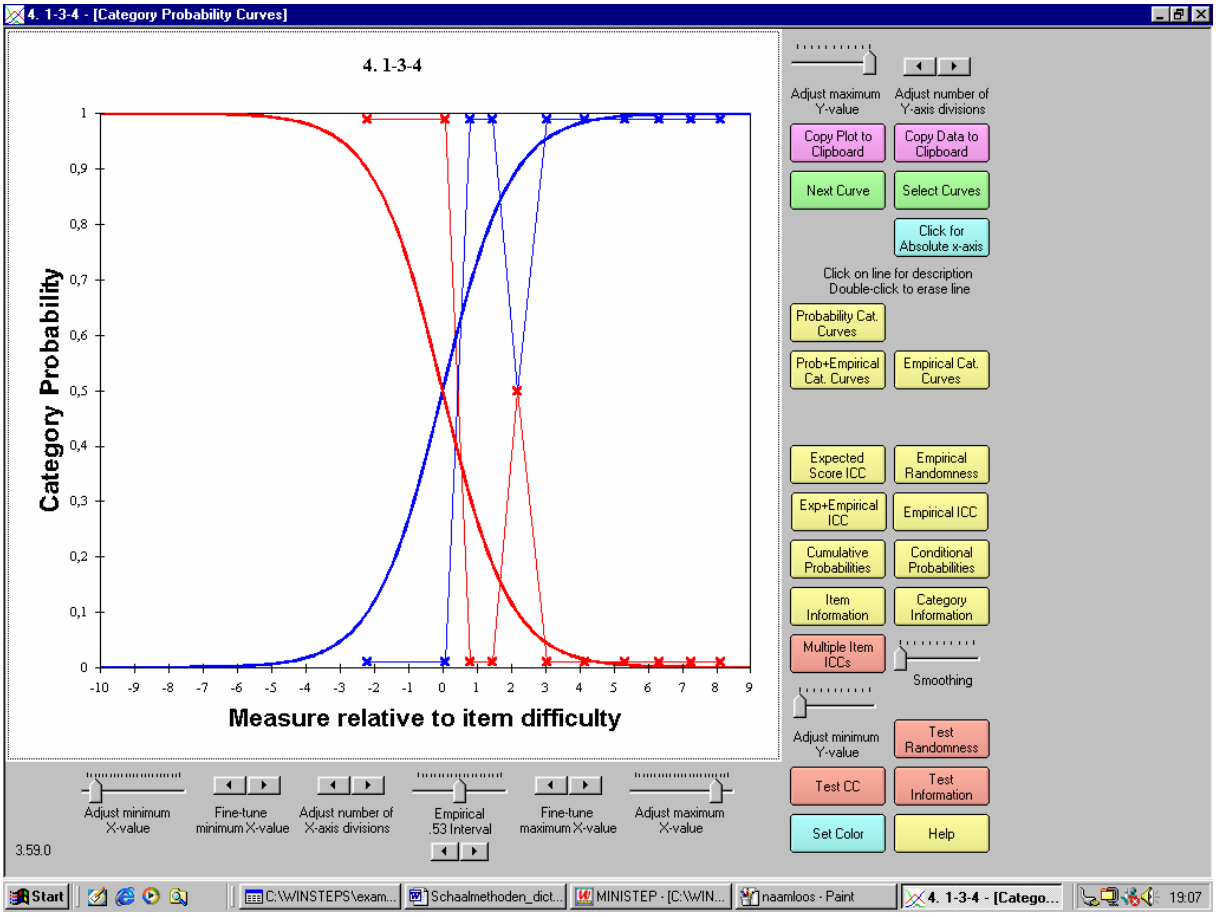
Figuur 4: Winstep met voorbeeld data

Het is natuurlijk de bedoeling om de data te analyseren en hier conclusies uit te trekken. Winsteps bevat hiervoor enorm veel mogelijkheden. Als in het File menu voor start winstep gekozen wordt begint het programma met analyseren. Zie figuur 5.



Figuur 5: Winsteps analyse van de data

In de taakbalk boven in figuur 5 kan uit een heleboel verschillende mogelijkheden van onder andere analyseren en representeren gekozen worden. In het voorbeeld wordt voor een grafiek gekozen. Deze is te zien in figuur 6.



Figuur 6: grafiek voor label 1-3-4

Boven in beeld is te zien dat dit de grafiek voor label 1-3-4 is. Door gebruik te maken van de knoppen naast de grafiek kan gekozen worden voor andere stimuli. Ook kan de manier van representeren aangepast worden. Verschillende grafieken kunnen bij elkaar in worden afgebeeld, de assen kunnen aangepast worden, de grafiek kan om de grafiek te editten en te exporteren naar andere bestanden.

Het gaat te ver om alle functies van Winsteps te behandelen. Dit blijkt alleen al uit de handleiding die een demotiverende 320 pagina's dik is. Bovendien kunnen met het programma veel handelingen verricht worden die in het dictaat niet aan de orde zijn gekomen.

Bronvermelding

[1] Dunn-Rankin, P. et al (2004, second edition)
Scaling methods
Laurens Erlbaum Associates, Mahwah

[2] Kruskal, J.B., W. Myron (1978)
Multidimensional scaling
Sage University Paper Series on Quantitative Applications in the Social Sciences
Newbury Park, CA

[3] Mokken, R.J. (1971)
A theory and procedure of scale analysis
Mouton, The Hague

[4] Segers, J. (2002)
Methoden voor de maatschappij wetenschappen
Van Gorcum, Assen
Hoofdstuk 6

[5] Swanborn, P.G. (1993, derde druk)
Schaaltechnieken: theorie en praktijk van acht eenvoudige procedures
Boom, Amsterdam