‘Best-worst scaling’ toegepast op de marketing van 3D-televisies zonder bril

Bachelorscriptie Econometrie & Besliskunde (FEB 23100)

Erasmus School of Economics

Erasmus Universiteit Rotterdam

Door: C.N. Edelenbosch

Begeleider: Dr. A.J. Koning

Meelezer: Prof.Dr. P.H.B.F. Franses

Abstract

Veel marktonderzoekers gebruiken schalen om consumentenvoorkeuren te begrijpen. Het hanteren hiervan gaat gepaard met diverse problemen, welke in bepaalde mate verbeterd kunnen worden door over te gaan op een nieuwe techniek gebaseerd op rankings. In deze scriptie wordt de methode van 'best-worst scaling' (kortweg: BWS) besproken. Bij deze methode worden respondenten gedwongen om een keuze te maken uit diverse te bestuderen samples, waarbij zowel de beste als slechtste optie in een beschikbare (deel)verzameling van samples gekozen dienen te worden.

In een empirisch onderzoek met betrekking tot de consumentenvoorkeuren naar een toekomstig product, te weten 3D-televisies zonder bril, zal duidelijk worden gemaakt dat de in praktijk meest gehanteerde methode van 'line scaling' en de methode van 'best-worst scaling' leiden tot vergelijkbare resultaten. Echter, er zal bewijs worden geleverd door middel van een procedure met meervoudige vergelijkingen, waaruit te concluderen valt dat BWS leidt tot een betere discriminatie in voorkeuren. De voor- en nadelen van het gebruik van BWS zullen worden besproken en tegen elkaar worden afgewogen.

Trefwoorden: 3D-televisies, 'best-worst scaling', consumentengedrag, gebalanceerd incompleet blokontwerp en rankings.

Inhoudsopgave

Hoofdstuk 1

Introductie

1.1 Motivatie

Bij marktonderzoek wordt vaak gebruik gemaakt van het schalen van verschillende meervoudige eenheden. Zo kan men bijvoorbeeld een stelling beantwoorden met de antwoordmogelijkheden helemaal mee eens/ mee eens/ neutraal/ mee oneens of helemaal mee oneens. Het is ook mogelijk om de mate waarin men het eens is weer te geven met een cijfer van 1 tot en met 10, waarbij 1 staat voor helemaal mee oneens en 10 voor helemaal mee eens. De laatstgenoemde correspondeert met een zogenaamde 'line scale'. Deze schalen worden in de praktijk standaard toegepast en men kijkt zelden naar alternatieve methoden om de voorkeuren te kunnen beschrijven.

Marktonderzoekers zijn voortdurend bezig met het meten en voorspellen van het belang en de voorkeuren van diverse productattributen om vervolgens nog beter aan de wensen en de vraag van consumenten te kunnen voldoen. Een groot nadeel bij het hanteren van 'line scales' is dat deze methode vaak leidt tot een zwakke discriminatie (Hein *et al*., 2008). Er kan dan geen goed onderscheid worden gemaakt tussen het belang van verschillende attributen. Om te komen tot een sterke discriminatie kan men overgaan op de methode van 'best-worst scaling' (kortweg: BWS) (Finn en Louviere, 1992). De respondent moet dan steeds per deelverzameling van attributen kiezen welke de meest geprefereerde en de minst geprefereerde is. De onderzoeker is dan vervolgens in staat om een volledige ranking aan te maken van alle attributen.

Om te demonstreren hoe de methode van 'best-worst scaling' werkt zal er gebruik worden gemaakt van data met betrekking tot de mogelijke aanschaf van 3D-televisies zonder bril. Daarbij is er als doel gesteld om te bestuderen naar welk productconcept de voorkeur uitgaat. Ieder productconcept is verschillend en bestaat uit een combinatie van vijf verschillende attributen. Middels een gebalanceerd incompleet blokontwerp worden er specifieke deelverzamelingen van productconcepten aangemaakt. Per deelverzameling dient de meest geprefereerde en minst geprefereerde optie gekozen te worden. Vervolgens kunnen op twee aan elkaar verwante manieren volledige rankings van productconcepten worden verkregen. Om aan te kunnen tonen of de voorkeur naar een bepaald productconcept significant verschillend is dan die naar een ander productconcept worden er meervoudige vergelijkingsprocedures toegepast.

De scriptie zal gaan over de voor- en nadelen van het gebruik van 'best-worst scaling' ten opzichte van 'line scaling' en de verschillen tussen beide methoden, voornamelijk ten aanzien van de discriminatie tussen de verschillende productconcepten, toegepast op de markt van 3D-televisies zonder bril.

1.2 Opbouw scriptie

De volgende onderdelen van de scriptie zijn als volgt gestructureerd. Als eerste zal er in hoofdstuk 2 theorie verschaft worden over de verschillende schaalmethoden die veelvuldig in de praktijk worden toegepast, om daarmee te bepalen welke attributen een belangrijke bijdrage leveren aan hetgeen wat men onderzoekt. Daaropvolgend zal de methode van 'best-worst scaling' behandeld worden. Daarbij zal tevens aandacht worden besteedt aan het ontwerp van een benodigd onderzoek. Vervolgens zal aangegeven worden welke stappen ondernomen moeten worden om van de verzameling aan data te komen tot een complete ranking van attributen. In hoofdstuk 3 volgt een empirisch onderzoek met betrekking tot de consumentenvoorkeuren naar een toekomstig product, te weten 3D-televisies zonder bril aan bod komen. In hoofdstuk 4 zullen de resultaten van het empirisch onderzoek op overzichtelijke wijze beschreven worden om vervolgens een uitspraak te kunnen doen over de belangrijkste verschillen tussen de methoden. In het laatste hoofdstuk staan de conclusies, de beperkingen waarop ik stuitte (gedurende het empirisch onderzoek) en de richtlijnen voor verder onderzoek.

Hoofdstuk 2

Methodologie van ‘best-worst scaling’

2.1 Type data

Marktonderzoekers bestuderen hoe er nog beter aan de wensen van de consumenten kan worden voldaan om het aanbod op de vraag af te stemmen. Daarvoor moet men het belang en de voorkeuren van diverse productattributen meten en voorspellen. Om dit te kunnen uitvoeren is er data nodig. Hiervan zijn er twee typen bekend: ‘stated preference’ data en ‘revealed preference’ data. De eerstgenoemde betreft data met betrekking tot wat respondenten zeggen te gaan doen en de ander correspondeert met wat men daadwerkelijk doet/ heeft gedaan.

Men zou verwachten dat er geen verschil zou bestaan tussen de twee typen data, maar dat is in werkelijkheid niet het geval. Deze tegenstrijdigheid kan bijvoorbeeld worden veroorzaakt door het hanteren van een onheldere vragenlijst. De vraag die men zichzelf dan zou moeten stellen is met welk type data er idealiter gewerkt zou moeten worden. ‘Revealed preference’ data wordt in de regel vaak toegepast om te voorkeuren binnen bestaande markten te begrijpen. Dit kan, ter illustratie, worden gedaan door de verkoopcijfers te bestuderen. Een nadeel hiervan is dat men geen informatie kan verkrijgen over de toevoeging van nieuwe attributen of nieuwe gerelateerde producten. Een ander bijkomend nadeel is dat er vaak maar een beperkte hoeveelheid informatie beschikbaar is om marktsegmentatie toe te kunnen passen. Deze problemen kunnen verholpen worden door te werken met ‘stated preference’ data, waarbij er tevens inzichten verkregen kunnen worden in de toekomstige markten. Qua voorspellend vermogen bij veranderingen in gedrag kan men zodoende beter werken met ‘stated preference’ data. Aangezien marktonderzoekers op reguliere basis schattingen dienen te maken voor de effecten van nieuwe producten met nieuwe attributen of kenmerken, gebruikt men daardoor ‘stated preference’ data. In het vervolg van deze scriptie zal er naar dit type data worden verwezen.

2.2 Schaalmethoden

Marktonderzoekers geven in de praktijk voornamelijk de voorkeur aan een methode die zowel overzichtelijk is als een eenvoudige analyse toestaat. Men kiest dan veelal voor zogenaamde ‘line scales’. Hierbij dienen de respondenten hun mate van voldoening voor ieder attribuut bijvoorbeeld aan te geven op een intervalschaal van 1 t/m 10. Voor de respondenten is het niet moeilijk en tijdrovend om een cijfer toe te kennen aan een attribuut. Voor de marktonderzoekers is het op hun beurt weer eenvoudig om de data te

analyseren. Zij kunnen de verkregen gemiddelden voor de attributen met elkaar vergelijken door gebruik te maken van *t*-toetsen.

Zoals hierboven beschreven is maken marktonderzoekers in de meeste gevallen gebruik van 'line scales', met als motivatie dat het voor de respondenten en de marktonderzoekers eenvoudig is om hiermee te werken. Toch zijn er een aantal bezwaarlijke redenen te noemen om van de methode van 'line scales' af te wijken. Het is namelijk zo dat de beoordelingen van respondenten afhankelijk zijn van de manier waarop ze 'ratings' opvatten. Dit verschilt per individu, maar ook voor diverse segmenten waaruit de markt is opgebouwd. De afstand tussen opeenvolgende 'ratings' kunnen dus per individu verschillen (Crask en Fox, 1987). Daarnaast is het zo dat mensen hun beoordelingen soms beperken tot slechts een gedeelte van de intervalschaal (Couch en Keniston, 1960). Verder geldt dat de beoordelingen kunnen variëren tussen uiteenlopende culturen. De verschillen tussen de internationale markten zouden in bepaalde gevallen dan eerder verklaard kunnen worden door verschillen in schaalgebruik, dan dat er daadwerkelijk verschillen in voorkeuren aanwezig zouden zijn. De conclusies die men dan trekt uit de gehanteerde schalen kunnen daardoor onzuiver zijn (Cohen, 2003). Het zou in bepaalde gevallen logischer zijn om de intervalschaal te vervangen door een ordinale schaal om deze problemen (deels) te voorkomen.

Marktonderzoekers willen aan kunnen tonen of bepaalde attributen significant belangrijker zijn dan andere attributen. Als er gewerkt zal worden met een 'line scale', dan is het vaak lastig om hier een uitspraak over te kunnen doen. De reden daarvoor is dat respondenten soms ieder attribuut even (on)belangrijk vinden. Er zal dan een zwakke discriminatie tot stand komen tussen de verschillende attributen, doordat er sprake is van een hoge variantie in de resultaten (Hein *et al*., 2008). De marktonderzoeker kan dan geen betrouwbare conclusies trekken over het relatieve belang van bepaalde attributen, aangezien er geen mogelijkheid is om een juiste afweging te maken tussen de desbetreffende attributen.

De voorkeuren kunnen ook op een andere manieren worden gemeten. Een goede methode die in de marketing wordt toegepast zijn discrete keuze-experimenten (Louviere en Woodworth, 1983; Louviere *et al*., 2000). Er worden dan verschillende unieke combinaties van attributen gemaakt. Deze worden productconcepten genoemd. De respondenten moeten dan steeds per deelverzameling van productconcepten kiezen welke de meest geprefereerde en de minst geprefereerde is. De onderzoeker is dan vervolgens in staat om een volledige ranking aan te maken. Deze methode leidt tot een sterke discriminatie.

De taken die de respondenten bij discrete keuzemodellen dienen uit te voeren zijn niet moeilijk als het aantal te combineren attributen beperkt is. Als dit aantal toeneemt, dan wordt het lastiger en vermoeiender voor de respondent en dan zou de respondent taakcomplexiteit kunnen ervaren (DeShazo en Fermo, 2002). Een te groot aantal keuzes kan zelfs demotiverend werken (Iyengar en Lepper, 2000). Om taakcomplexiteit te omzeilen kan men overgaan tot partiële rankings, waarbij er rankings worden bepaald voor deelverzamelingen van verschillende productconcepten.

2.3 'Best-worst scaling'

Er zijn diverse manieren om te komen tot rankings. Verreweg de meest eenvoudige en meest betrouwbare methode van ranken wordt bereikt door gepaarde vergelijkingen te gebruiken (Thurstone, 1927). De respondent moet in dat geval steeds per deelverzameling, bestaande uit twee items, kiezen welke het belangrijkste is. Als er *n* verschillende items zijn, dan volgt het aantal paren met behulp van de onderstaande formule:

$$Aantal\ paren(n) = n * (n-1)/2 \qquad (1)$$

Het nadeel van gepaarde vergelijkingen is dat het aantal te beschouwen paren aanzienlijk toeneemt als het aantal verschillende items toeneemt. Dat is te zien als we steeds hogere waarden invullen voor *n*. Zo zullen bijvoorbeeld 10 items leiden tot 45 paren, 15 items tot 105 paren en 20 items tot 190 paren.

Er is een manier om te voorkomen dat er teveel deelverzamelingen aanwezig zullen zijn waarbinnen aangegeven dient te worden wat het belangrijkste item is. Dit kan bereikt worden door meer items in een deelverzameling te plaatsen en te vragen naar een ranking van alle items binnen iedere deelverzameling. Aangezien het aantal items per deelverzameling is toegenomen, heeft dit als direct gevolg dat er een afname is van het aantal benodigde deelverzamelingen.

Een alternatief voor het ranken van alle items per deelverzameling is dat men alleen kiest welk item het meest geprefereerd en het minst geprefereerd wordt. In dat geval kom je tot de 'best-worst'-methode, welke in feite een uitbreiding is van de methode van gepaarde vergelijkingen. Deze methode modelleert het proces waarbij respondenten twee items kiezen met de voor hen maximale verschillen in preferentie of mate van belang op basis van een onderliggende verborgen dimensie (K.Y. Lam, 2011). Dat een volledige ranking per deelverzameling niet nodig is komt omdat het voor respondenten lastig is om de niet-extreme items te ranken (Ben-Akiva *et al*., 1992).

Het hanteren van de methode van 'best-worst scaling' heeft een aantal voordelen als men een vergelijking maakt met het gebruik van de 'line scales'. Elke respondent kan per deelverzameling alleen kiezen welk productconcept de meeste en de minste voorkeur geniet, waardoor er telkens direct een afweging zal worden gemaakt tussen alle productconcepten binnen de deelverzameling (Cohen, 2003). De beoordelingen van respondenten zullen dan niet afhankelijk zijn van de manier waarop de 'ratings' worden opgevat. De mogelijke verschillen tussen uiteenlopende culturen kunnen bij het gebruik van 'line scales' leiden tot onzuiverheden, maar dat zal worden voorkomen als men overgaat op de methode van 'best-worst scaling'.

Een groot voordeel als men data gaat bestuderen met behulp van 'best-worst scaling' is dat er een toegenomen discriminatie tussen de verschillende productconcepten van kracht is. Als men gebruik maakt van 'line scales' om de discriminatie in termen van preferenties te bestuderen, dan schiet men hier tekort. De zwakke discriminatie was dan ook

een belangrijke factor voor het ontwikkelen van deze nieuwe schaalmethode, die tegelijkertijd eenvoudig in gebruik moest zijn (Cohen en Neira, 2003; Cohen en Orme, 2004; Finn en Louviere, 1992). Men is zich van de voordelen bewust en zodoende wordt 'best-worst scaling' vanaf het moment van introductie al in vele verschillende gebieden gebruikt, variërend van de voedselindustrie tot de gezondheidszorg.

2.3.1 Ontwerp onderzoek

Zoals hierboven beschreven staat geldt dat bij het gebruiken van de 'best-worst scaling' er minimaal drie verschillende productconcepten in een deelverzameling geplaatst dienen te worden. Des te hoger het gekozen aantal verschillende productconcepten per deelverzameling, des te kleiner het benodigde totale aantal deelverzamelingen. Echter, het is niet eenvoudig om zomaar te bepalen welke productconcepten in een bepaalde deelverzameling horen als het aantal te bestuderen productconcepten groot is.

Het benodigde aantal deelverzamelingen en de specifieke productconcepten voor iedere deelverzameling dienen op een zorgvuldige manier te worden gekozen. Daarvoor kan een gebalanceerd incompleet blokontwerp (kortweg: BIBD) worden gebruikt. Een voordeel van het gebruik van een BIBD is dat een groot aantal productconcepten kan worden bestudeerd, om een volledige ranking te krijgen van alle productconcepten, door een relatief klein aantal deelverzamelingen en productconcepten per deelverzameling. Het BIBD beheert het aantal keren dat ieder paar productconcepten met elkaar vergeleken wordt en door dit aantal te verhogen zal het aantal deelverzamelingen en/of het aantal productconcepten per deelverzameling moeten stijgen. Als het aantal keer dat een productconcept met ieder ander productconcept wordt vergeleken toeneemt, dan zal dit ten goede komen aan de interne geldigheid van het onderzoek tegen de prijs dat het tijdrovender en ingewikkelder zal worden voor de respondent om deel te nemen.

Het BIBD kan worden afgeleidt uit een het ontwerp van een Latijns vierkant. Dat is in feite op te vatten als een matrix met n rijen en n kolommen met n productconcepten, waarbij elk productconcept exact een keer voorkomt per rij en kolom. Iedere rij kan dan worden beschouwd als een blok of een verzameling. Een Latijns vierkant is een gebalanceerd compleet blokontwerp, aangezien iedere rij alle productconcepten bevat en de totale mate van optreden van een productconcept voor ieder productconcept hetzelfde is (Weller en Romney, 1988). Als het aantal kolommen wordt gereduceerd door k kolommen te verwijderen, dan is het resultaat een ontwerp met $n-k$ productconcepten per rij. Een dergelijk ontwerp is een BIBD, mits ieder productconcept even vaak voorkomt in alle deelverzamelingen. Een van de vereisten van een BIBD is dat het aantal deelverzamelingen dan minimaal even groot is als het aantal productconcepten.

Een BIBD voor n productconcepten wordt doorgaans aangegeven met BIBD(n, b, r, p, λ), waarbij n dus staat voor het aantal productconcepten, b voor het aantal blokken/ deelverzamelingen, r voor het aantal herhalingen van een productconcept over alle blokken, p voor het aantal productconcepten per deelverzameling en λ voor het aantal

keren dat een paar productconcepten voorkomt. Al deze waarden zijn niet onafhankelijk van elkaar.
Om te controleren of er sprake is van een gebalanceerd incompleet blokontwerp, dienen de vergelijkingen (2) en (3) te kloppen:

$$n * r = b * p = totale\ aantal\ \ eenheden\ in\ het\ onderzoek \quad (2)$$

$$r * (p-1) = \lambda(n-1) \quad (3)$$

Een voorbeeld van een gebalanceerd incompleet blokontwerp is een BIBD(7,7,3,3,1). Hier geldt dat er 7 verschillende productconcepten zijn, waarbij elke van de 7 deelverzamelingen bestaat uit 3 van de 7 verschillende productconcepten. Ieder productconcept komt dan 3 keer voor om ervoor te zorgen dat ieder paar aan productconcepten slechts eenmalig optreedt in het ontwerp. Het ontwerp staat beschreven in de onderstaande tabel.

	Productconcept 1	**Productconcept 2**	**Productconcept 3**
1	2	3	6
2	1	3	5
3	3	4	7
4	1	2	7
5	5	6	7
6	1	4	6
7	2	4	5

Tabel 1 BIBD(7,7,3,3,1)

Marktonderzoekers die willen werken met data voor BWS dienen een correct BIBD te hanteren. Als er een verkeerde variant wordt gebruikt, dan kan dit leiden tot onzuivere resultaten. Dit heeft als consequentie dat er een mogelijke verkeerde ranking van productconcepten tot stand komt.

De eerste stap bij het bepalen welke BIBD geschikt is, is het beslissen van het aantal te vergelijken productconcepten. Er zal een afweging gemaakt dienen te worden tussen het aantal productconcepten per deelverzameling en het totale aantal deelverzamelingen. Uit de praktijk blijkt dat het optimaal is om vier tot zes productconcepten per deelverzameling te gebruiken, om ervoor te zorgen dat het niet te complex wordt voor een respondent om het meest geprefereerde en minst geprefereerde productconcept te kiezen. Daarnaast dient het aantal deelverzamelingen niet al te groot gekozen te worden, aangezien de verveling dan een rol gaat spelen. Dat kan ertoe leiden dat de respondent het onderzoek niet helemaal zal voltooien.

2.4 Analyse rankings

De 'best-worst scaling'-methode kan pas plaatsvinden nadat er eerst zorgvuldig gekeken is naar een bijpassend gebalanceerd incompleet blokontwerp om de verschillende deelverzamelingen van productconcepten mee aan te maken. Als dit in orde is, dan kan men vervolgens overgaan op verschillende processen om uiteindelijk te komen tot de gewenste output: de rankings van de productconcepten, waarbij aan te tonen valt welke productconcepten significant belangrijker zijn dan andere productconcepten. De verschillende stappen die ondernomen dienen te worden om tot de rankings te komen zullen in deze paragraaf aan bod komen. Daarbij zullen er twee verschillende methoden worden besproken.

Allereerst moet de keuze die een respondent in een bepaalde deelverzameling maakt worden getransformeerd naar een keuze voor het desbetreffende productconcept. Zodoende weet men wat respondent i per deelverzameling aangegeven heeft als productconcept die het meest resp. het minst worden geprefereerd. De productconcepten zijn genummerd met j , j = 1,…, J .

Je wilt uitkomen op twee matrices, B en W , die voor iedere respondent aangeven hoe vaak een bepaald productconcept in totaal als beste en als slechtste worden gekozen. Door deze matrices van elkaar af te halen wordt de matrix S verkregen met hierin alle 'best-worst'-scores.

$$b_{i,j} = \sum_{k=1}^{K} I[beste\ productconcept\ in\ deelverzameling\ k = j] \qquad (4)$$

$$w_{i,j} = \sum_{k=1}^{K} I[slechtste\ productconcept\ in\ deelverzameling\ k = j] \qquad (5)$$

$$s_{i,j} = b_{i,j} - w_{i,j} \qquad (6)$$

$$i = 1,\ldots, I \quad j = 1,\ldots J$$

Er is bij de bepaling van de eenheden in de matrices B en W gebruik gemaakt van een indicatorfunctie I [...]. Deze neemt de waarde 1 aan als de bewering tussen haakjes van toepassing is en 0 elders. Er is hier gewerkt met in totaal K deelverzamelingen.

In de matrix S staat nu per individu de 'best-worst'-score voor ieder productconcept vermeldt, oftewel het aantal keren dat een bepaald productconcept als beste is gekozen minus het aantal keren dat dit productconcept als slechtste is gekozen in alle deelverzamelingen. Een positieve waarde voor een element in de matrix S wil zeggen dat productconcept j voor respondent i vaker als beste is gekozen dan als de slechtste.

Om te komen tot een algehele ranking van de productconcepten dienen de gegevens van alle respondenten samen bekeken te worden. Het is dan de bedoeling om een sommatie

te nemen over alle respondenten om te kijken wat de totale 'best-worst'-scores zijn per productconcept. De resulterende waarden kunnen dan in de vector *totaalscore* worden geplaatst. Zie vergelijking (7).

$$totaalscore_j = \sum_{i=1}^{I} s_{i,j} \qquad (7)$$

$$j = 1,...J$$

Er zijn verschillende manieren om te controleren of de waarden van de elementen in *totaalscore* juist lijken. Daarvoor zou men onder andere kunnen kijken naar de matrix S. Er moet voor alle elementen van deze matrix gelden dat $-r \leq s_{i,j} \leq r$. Immers, ieder productconcept verschijnt in totaal r keer in de BIBD en kan dan maximaal r keer worden gekozen als beste en 0 keer als slechtste of andersom. Verder moeten de onderstaande vergelijkingen (8) en (9) gelden. Daarbij geldt dat de 'best-worst'-scores per respondent voor alle productconcepten gelijk moeten zijn aan 0. Dit impliceert weer dat de sommatie van alle elementen van de matrix S gelijk moeten zijn aan 0. Vergelijking (9) volgt dus in feite uit vergelijking (8).

$$\sum_{j=1}^{J} s_{i,j} = 0 \qquad (8)$$

$$i = 1,...I$$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} s_{i,j} = 0 \qquad (9)$$

Nadat er geen fouten zijn opgespoord kan eenvoudig de totale ranking worden bepaald. De hoogste waarde van *totaalscore* correspondeert met het productconcept dat voor alle respondenten gemiddeld genomen de hoogste voorkeur geniet. Deze krijgt dan de rank 1. Daarna wordt er gezocht naar de een-na-hoogste waarde en het productconcept dat hierbij hoort krijgt rank 2. Dit wordt vervolgens gedaan totdat uiteindelijk rank *J* ook toegekend is.

Vaak worden er nog andere manieren gebruikt om, voorafgaande aan de bepaling van de rankings, te komen tot waarden die men kan interpreteren om vervolgens de ranking hierop te laten baseren. Men vermeldt naast de totale 'best-worst'-scores vaak ook de gemiddelde 'best-worst'-scores. Deze gemiddelden kunnen worden verkregen door de 'best-worst'-scores te delen door het aantal respondenten en het aantal keren dat een productconcept voorkwam in de BIBD, aangegeven met de waarde r. Een manier om de mate van voorkeur tussen verschillende productconcepten te vergelijken is door het bepalen van de relatieve ratioscores. Om deze te verkrijgen moet de wortel worden genomen van het quotiënt van het totaal aantal 'best'-scores en 'worst'-scores voor iedere respondent.

Om ervoor te zorgen dat er niet door 0 gedeeld zal worden, dient er 0,5 bij het totaal aantal 'worst'-scores te worden opgeteld indien nodig. Dit zal voor iedere respondent worden gedaan en dan zal er een gemiddelde worden bepaald per productconcept. Vervolgens wordt er geschaald met een bepaalde factor, wat ertoe leidt dat de score voor het belangrijkste productconcept gelijk zal zijn aan 100. Alle productconcepten kunnen dan paarsgewijs worden vergeleken middels het relatieve ratio.

Er is nog een andere manier om tot de uiteindelijke ranking te komen en dat is een niet-parametrische manier door gebruik te maken van individuele rankings (K.Y. Lam, 2011). Per deelverzameling dient de respondent dan aan te geven welk productconcept het meest wordt geprefereerd en dat productconcept krijgt de rank 1. Er moet ook een productconcept worden gekozen die het minst wordt geprefereerd en deze krijgt de rank p, dat gelijk is aan het aantal productconcepten per deelverzameling. De overige productconcepten krijgen eenzelfde rank (p +1)/2. Indien een productconcept niet voorkomt in een deelverzameling, dan wordt voor het gemak de rank 0 toegekend om later geen problemen op te leveren bij de bepaling van gemiddelde rankings voor de verschillende productconcepten. Men zou er ook voor kunnen kiezen om de overgebleven productconcepten van iedere deelverzameling te ranken, maar verschillende studies hebben reeds aangetoond dat er dan inconsistentie kan optreden (e.g. Ben-Akiva et al.,1992), vandaar dat er gekozen wordt voor eenzelfde rank. De waargenomen gemiddelde ranking van respondent i voor productconcept j wordt aangegeven met $x_{i,j}$. Een waarde kleiner dan (p +1)/2 wil zeggen dat productconcept j voor respondent i vaker als beste is gekozen dan als de slechtste.

$$x_{i,j} = \frac{1}{r}\sum_{k=1}^{K} rank\ van\ productconcept\ \ j\ in\ deelverzameling\ \ k \qquad (10)$$

$$i = 1,\ldots, I$$

Door het gemiddelde te nemen van de individuele rankings van een productconcept wordt een gemiddelde ranking bepaald die te vergelijken valt met die van andere productconcepten.

$$rankgemiddelde_{j} = \frac{1}{I}\sum_{i=1}^{I} x_{i,j} \qquad (11)$$

$$j = 1,\ldots, J$$

Naast de ranks wordt er doorgaans ook gekeken naar de aangepaste ranks. Die kunnen worden verkregen door de verwachte rank, (p +1)/2, van de zojuist bepaalde rank af te halen. De aangepaste rank van respondent i voor productconcept j wordt aangegeven met $a_{i,j}$, zie hiervoor vergelijking (12).

$$a_{i,j} = x_{i,j} - \frac{p+1}{2} \tag{12}$$

$$i = 1,\ldots, I \quad j = 1,\ldots J$$

De aangepaste ranks kunnen worden samengevat in een vector $aplus$, door allereerst de aangepaste ranks voor ieder productconcept voor alle respondenten te bepalen en vervolgens een sommatie te nemen, zie (13).

$$aplus_j = \sum_{i=1}^{I} a_{i,j} \tag{13}$$

$$j = 1,\ldots, J$$

Een controlemiddel om te kijken of de rankgemiddelden kloppen kan worden gedaan door na te gaan of het gemiddelde van alle rankgemiddelden gelijk is aan de rank die een productconcept toegekend krijgt als het van een willekeurige deelverzameling niet als beste of slechtste wordt geselecteerd, terwijl het productconcept wel in de deelverzameling zit. Vergelijking (14) moet dan gelden.

$$\frac{1}{J} \sum_{j=1}^{J} rankgemiddelde_j = \frac{p+1}{2} \tag{14}$$

Als alles lijkt te kloppen, dan kan eenvoudig de totale ranking worden bepaald. De laagste waarde uit de vector $rankgemiddelde$ correspondeert met het productconcept dat voor alle respondenten gemiddeld genomen de hoogste voorkeur geniet. Deze krijgt dan de uiteindelijke rank 1. Daarna wordt er gezocht naar de een-na-laagste waarde en het productconcept dat hierbij hoort krijgt rank 2. Dit wordt vervolgens gedaan totdat uiteindelijk rank *J* ook toegekend is.

Aangezien er bij de data die geanalyseerd zal worden bij ‘best-worst scaling’ gewerkt is met rankings lijkt het handig om dit ook toe te passen bij de data behorende bij de ‘line scale’. De respondenten kunnen dan voor ieder productconcept een getal van 1 tot en met 10 toekennen. Een hoger cijfer geeft dan een hogere mate van voorkeur aan. Om te vergelijken welk productconcept gemiddeld genomen de meeste voorkeur geniet wordt het gemiddelde genomen over alle cijfers die aan een productconcept zijn gegeven. Vervolgens volgt een methode die identiek is aan wat hierboven al tweemaal is toegepast, waarbij het hoogste gemiddelde correspondeert met rank 1 en de laagste met rank J.

2.4.1 Toetsen op discriminatie

Als de totale rankings zijn bepaald, dan is het vervolgens belangrijk om na te gaan welke productconcepten significant van elkaar verschillen in mate van preferentie of belang. Er zijn reeds drie verschillende methoden besproken om te komen tot de totale rankings, te weten de methode van 'line scales' en twee aan elkaar verwante 'best-worst scaling'-methoden. In deze paragraaf zal worden toegelicht hoe er per methode een discriminatie in productconcepten tot stand komt.

Bij het toetsen op discriminatie dienen er eerst een nulhypothese en een alternatieve hypothese te worden opgesteld.

> H_0: Er zijn geen verschillen in de voorkeuren tussen de productconcepten. Ieder productconcept is even gewild.
> H_a: Minstens 1 productconcept heeft een hogere voorkeur dan minstens 1 ander productconcept.

Als er voldoende significant bewijs is om de nulhypothese te verwerpen, dan dient er vervolgens gekeken te worden om welke productconcepten het gaat. Er zal gebruik worden gemaakt van meervoudige vergelijkingsprocedures om conclusies te kunnen trekken over de verschillen in paren van verschillende productconcepten.

Bij het gebruiken van de 'line scales' verkrijgt men gemiddelde cijfers voor alle productconcepten. Om deze onderling te kunnen vergelijken kan er gebruik worden gemaakt van 'one-way analysis of variance' ('one-way ANOVA') en van de Tukey-B-toets. De Tukey-B-toets is een *post hoc* toets die krachtig genoeg is en algemeen geaccepteerd wordt voor dergelijke procedures. Het is een parametrische toets die veronderstelt dat de populatievarianties gelijk zijn. Het veronderstelt ook dat de samplegroottes gelijk zijn. Door een significantieniveau α te kiezen kan worden geanalyseerd of er significante verschillen in voorkeuren aanwezig zijn tussen diverse paren van productconcepten. Om de resultaten goed met elkaar te kunnen vergelijken, kunnen de betrouwbaarheidsintervallen voor de gemiddelden in een figuur worden geplaatst. Als de betrouwbaarheidsintervallen van twee verschillende productconcepten niet overlappen, dan geldt dat er een significant verschil bestaat tussen de desbetreffende productconcepten.

Indien de eerste beschreven methode met betrekking tot 'best-worst scaling' wordt toegepast, dan kan er ook gebruik worden gemaakt van de Tukey-B-toets. In dit geval worden voor alle productconcepten de totale (gemiddelde) best-worst scores gebruikt in plaats van de gemiddelde cijfers die eraan zijn toegekend. Voor de rest is er qua procedure geen verschil bij bestudering van de discriminatie. Ook hier zullen betrouwbaarheidsintervallen worden aangemaakt die toestaan dat de significante verschillen tussen verschillende productconcepten eenvoudig zijn op te sporen.
Als de tweede methode wordt gebruikt die van toepassing is op de 'best-worst'-data, dan zal er het een en ander veranderen. De Tukey-B-toets is een parametrische toets en deze zal

hier plaatsmaken voor een niet-parametrische toets, gerelateerd aan de Friedman-toets. De toetsingsgrootheid die dan gebruikt wordt is W. Om tot hiertoe te komen dient allereerst de individuele covariantiematrix V_i te worden opgesteld, zie (15). Deze symmetrische matrix beschikt over J rijen en J kolommen.

$$V_i = \begin{bmatrix} \left(\frac{(p-1)^2}{2p}\right) & \left(\frac{1-p}{2p}\right) & \left(\frac{1-p}{2p}\right) & \cdots & \left(\frac{1-p}{2p}\right) \\ \left(\frac{1-p}{2p}\right) & \left(\frac{(p-1)^2}{2p}\right) & \left(\frac{1-p}{2p}\right) & \cdots & \vdots \\ \left(\frac{1-p}{2p}\right) & \left(\frac{1-p}{2p}\right) & \left(\frac{(p-1)^2}{2p}\right) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \left(\frac{1-p}{2p}\right) \\ \left(\frac{1-p}{2p}\right) & \cdots & \cdots & \left(\frac{1-p}{2p}\right) & \left(\frac{(p-1)^2}{2p}\right) \end{bmatrix} \qquad (15)$$

$$i = 1,\ldots, I$$

De individuele covariantiematrix is, zoals hierboven te zien valt, alleen afhankelijk van p. Om te komen tot de covariantiematrix voor alle respondenten, V^+, worden alle elementen vermenigvuldigd met het aantal respondenten. De toetsingsgrootheid wordt berekend door de rijvector $aplus$ te vermenigvuldigen met de gegeneraliseerde inverse van V^+ en vervolgens met de getransponeerde van $aplus$, zie (16).

$$W = aplus * (V^+)^- * aplus' \qquad (16)$$

Deze toetsingsgrootheid is $\chi^2(p-1)$ verdeeld. Als de bijbehorende p-waarde kleiner is dan de gekozen waarde α, dan dient de nulhypothese te worden verworpen. Om vervolgens te zien welke productconcepten een significant verschil in voorkeur hebben, als de nulhypothese is verworpen, wordt er wederom gebruik gemaakt van meervoudige vergelijkingsprocedures. De betrouwbaarheidsintervallen voor de ranks van de productconcepten kunnen dan weer met elkaar worden vergeleken.

Waar ik op hoop is dat de discriminatie van de productconcepten sterker is als men gebruik maakt van de methoden van ‘best-worst scaling’, dan dat men werkt met ‘line scaling’. Verder kan er gekeken worden of er nog opmerkelijke verschillen zijn tussen de twee verschillende methoden die worden gebruikt bij de ‘best-worst scaling’. Dit zal in hoofdstuk 4 worden getoond aan de hand van de resultaten van een empirisch onderzoek voor 3D-televies zonder bril.

Hoofdstuk 3

Toepassing marketing van 3D-televisies zonder bril

3.1 Algemene databeschrijving

Marktonderzoekers bestuderen vaak hoe goed een product met een nieuwe attribuut of een geheel nieuw product het zal gaan doen op de markt. In deze scriptie zal het te bestuderen product een 3D-televisie zonder bril zijn. Er zijn verschillende redenen waarom er voor dit product is gekozen. Een van de redenen is dat het een nieuw product is waar veel mensen zich al wat bij voor kunnen stellen, deels door het aanbod van 3D-films in de bioscopen. Een andere reden is dat het product vooral een relatief jonge en innovatieve doelgroep aanspreekt, waarvan ik van tevoren verwachtte dat hier op korte termijn voldoende relevante informatie van in te winnen valt. Verder betreft het een product dat al over enkele jaren op de markt zal verschijnen en naar verwachting binnen vijf tot tien jaar de huidige generatie televisies gaat vervangen. Het is bijzonder interessant om te weten waar de consument op zal gaan letten bij de mogelijke aanschaf ervan. In deze scriptie zal worden onderzocht aan welke combinatie van kenmerken de consument de voorkeur geeft en welke combinaties niet interessant zijn.

3.1.1 Keuze voor productconcepten

Na gesprekken gevoerd te hebben met inkopers en verkopers van diverse bedrijven in Rotterdam e.o. die televisies verkopen, is er informatie verkregen over de kenmerken waar men waarschijnlijk het meeste op zal letten bij de mogelijke aanschaf van 3D-televisies zonder bril. Hieruit kwam naar voren dat de belangrijkste kenmerken, in willekeurige volgorde, als volgt zijn:

1. Merk
2. Schermdiagonaal
3. Het hebben van een PC- en internetverbinding
4. Prijs
5. Resolutie

Per kenmerk kunnen er splitsingen worden doorgevoerd om te komen tot verschillende klassen. Zo zijn er A-merken zoals Samsung en B-merken zoals Daewoo, waarbij B-merken vaak pas later op de markt komen dan A-merken. De schermdiagonaal kan relatief klein of groot zijn. Om dit beter te splitsen, wordt er gewerkt met een schermdiagonaal tot 40 inch (=102 cm.) en een vanaf 40 inch. Een break bij een schermdiagonaal van 102 cm. lijkt

misschien aan de hoge kant, maar om optimaal van 3D-televisie te kunnen genieten is een groter scherm vereist in vergelijking met die van de huidige generatie televisies. De televisie kan wel of niet beschikken over een PC- en internetverbinding. Deze twee verbindingen worden samengevoegd, want als er een PC-verbinding is dan zal er ook een internetverbinding zijn en andersom. Aangezien het onderzoek zal worden uitgevoerd voor de introductiefase van het product zal de gemiddelde prijs behoorlijk hoog zijn. Tegenwoordig is er soms voor nog geen €500 al een zeer degelijke televisie te koop, maar dat is slechts 20% van de verwachte gemiddelde verkoopprijs van 3D-televisies zonder bril die de 'early adaptors' ervoor zullen moeten betalen. De grenswaarde van €2500 wordt dan ook gehanteerd als ondergrens voor een hoge prijs. Tenslotte kan er een lage of hoge resolutie zijn. Een hoge resolutie is gekoppeld aan 'Full HD 3D Ready' en een lage resolutie aan 'Half HD 3D Ready'.

De verschillende kenmerken worden gecombineerd met als resultaat unieke productconcepten bestaande uit vijf kenmerken. Voor ieder van de vijf kenmerken was er een splitsing in twee disjuncte klassen, waardoor het totale aantal unieke productconcepten gelijk is aan $2^5 = 32$. Er dient dan gekeken te worden hoe het aantal uit te voeren experimenten kan worden beperkt zonder dat er (teveel) relevante informatie verloren zal gaan. Om die reden is er gekozen om te werken met halffracties met ieder $2^{5-1} = 16$ productconcepten. Er zal gewerkt worden met de halffractie die beschreven staat in tabel (2). Er gaat dan wel informatie verloren over de mogelijke aanschaf van de resterende productconcepten, maar het zal minder moeite kosten om informatie in te winnen. Daarnaast ligt de essentie van de scriptie meer op de verschillen in discriminatie van productconcepten tussen diverse methoden en dat kan nog goed worden onderzocht.

Productconcept	Merk	Schermdiagonaal	PC en internet	Prijs	Resolutie
1	B-merk	Schermdiagonaal vanaf 40 inch	Wel	Hoog	Laag
2	A-merk	Schermdiagonaal vanaf 40 inch	Niet	Hoog	Laag
3	A-merk	Schermdiagonaal tot 40 inch	Wel	Laag	Hoog
4	B-merk	Schermdiagonaal vanaf 40 inch	Wel	Laag	Hoog
5	A-merk	Schermdiagonaal vanaf 40 inch	Niet	Laag	Hoog
6	B-merk	Schermdiagonaal tot 40 inch	Wel	Hoog	Hoog
7	B-merk	Schermdiagonaal tot 40 inch	Niet	Hoog	Laag
8	A-merk	Schermdiagonaal tot 40 inch	Niet	Hoog	Hoog
9	A-merk	Schermdiagonaal tot 40 inch	Wel	Hoog	Laag
10	A-merk	Schermdiagonaal tot 40 inch	Niet	Laag	Laag
11	A-merk	Schermdiagonaal vanaf 40 inch	Wel	Hoog	Hoog
12	B-merk	Schermdiagonaal vanaf 40 inch	Niet	Laag	Laag
13	B-merk	Schermdiagonaal tot 40 inch	Wel	Laag	Laag
14	B-merk	Schermdiagonaal tot 40 inch	Niet	Laag	Hoog
15	A-merk	Schermdiagonaal vanaf 40 inch	Wel	Laag	Laag
16	B-merk	Schermdiagonaal vanaf 40 inch	Niet	Hoog	Hoog

Tabel 2 Gekozen productconcepten

3.2 Beschrijving enquête

Om aan de data te komen is een enquête afgenomen, wat in bruikbare informatie resulteerde van 137 respondenten. De enquête is grofweg te verdelen in drie delen. Het eerste deel bestaat uit de vraag wat voor cijfer van 1 tot en met 10 men zou toekennen aan alle productconcepten, op basis van mate van voldoening, als men dit productconcept zou aanschaffen. Een 1 staat daarbij voor de laagst haalbare voldoening en een 10 voor de maximale voldoening. De productconcepten moesten hierbij onafhankelijk van elkaar beoordeeld worden. Het tweede deel van de enquête bestaat uit zestien unieke deelverzamelingen van zes verschillende productconcepten waarbij telkens aangegeven moest worden welk productconcept het meest en welke het minst werd geprefereerd. Het derde deel van de enquête was bedoeld voor de statistieken om een eventuele marktsegmentatie mogelijk te maken. Voordat alle vragen beantwoordt moesten worden was er eerst een filmpje te zien over 3D-televisie en de toekomst ervan om de enquête leuker te maken voor de respondent. Echter, het was vooral bedoeld om de respondent een beter beeld te geven van de nieuwe vorm van televisie kijken in de hoop dat dit een gunstig effect zou hebben op de consistentie van de individuele data. Daarnaast heb ik er een geabstraheerd plaatje bijgeplaatst van een 3D-televisie om het productconcept duidelijker over te laten komen.

De volgorde van het stellen van de vragen was voor iedereen hetzelfde. Ik had dit graag random gemaakt, door de eerste twee delen van de enquête geen vaste plaats te geven. Helaas was dat niet mogelijk via de tool voor het online afnemen van de enquête. Het was ook niet mogelijk om de volgorde van de vragen waarbij een cijfer gegeven moest worden te laten variëren. Hetzelfde geldt voor de volgorde waarbij de deelverzamelingen voorkwamen waarbij de beste en slechtste productconcepten gekozen moesten worden. Als dit alles wel mogelijk was geweest, dan had dit voor het onderzoek nog beter geweest. Ondanks dat verwacht ik dat dit niet heeft geleidt tot zeer afwijkende resultaten.

Zoals hierboven al is vermeldt zijn er zestien unieke deelverzamelingen gebruikt die allemaal bestaan uit zes verschillende productconcepten. De deelverzamelingen zijn bepaald aan de hand van een geschikt gebalanceerd incompleet blokontwerp. Er is gekozen voor een BIBD(16,16,6,6,2), zie tabel (3). Ieder productconcept komt dus in totaal 6 keer voor in het gehele ontwerp en in totaal wordt twee keer hetzelfde paar aan productconcepten binnen een deelverzameling met elkaar vergeleken. Het feit dat een paar productconcepten meer dan eens met elkaar vergeleken wordt komt ten goede aan de interne geldigheid van het onderzoek. Door te werken met 16 productconcepten was dit het ontwerp dat heeft geleidt tot het kleinste aantal deelverzamelingen, waarbij de restrictie was opgelegd dat het aantal productconcepten per deelverzameling niet groter mag worden dan 6. Het bleek dat er 60 respondenten waren die halverwege het tweede gedeelte van de enquête zijn afgehaakt. Dat is zonde, want anders was er data bekend van 197 respondenten. Echter, het ontwerp had niet kleiner gekund als er gekozen is om te werken met vijf verschillende kenmerken. Ik verwacht niet dat het komt doordat de respondenten het moeilijk vonden om op de vragen

van het tweede gedeelte te antwoorden. Er is namelijk al eerder aangetoond dat de taken die de respondenten dienen uit te voeren eenvoudig worden bevonden en dat de vragen eenvoudiger zijn te beantwoorden dan de vragen uit het eerste gedeelte (Marley en Louviere, 2005).

	Product-concept 1	Product-concept 2	Product-concept 3	Product-concept 4	Product-concept 5	Product-concept 6
1	1	3	5	9	15	16
2	1	6	8	11	13	16
3	3	4	7	12	13	16
4	1	2	3	10	13	14
5	2	3	4	6	9	11
6	3	6	8	12	14	15
7	3	5	7	8	10	11
8	2	5	6	10	12	16
9	2	4	5	8	13	15
10	1	4	8	9	10	12
11	4	10	11	14	15	16
12	1	2	7	11	12	15
13	2	7	8	9	14	16
14	6	7	9	10	13	15
15	5	9	11	12	13	14
16	1	4	5	6	7	14

Tabel 3 BIBD(16,16,6,6,2)

Het derde deel van de enquête heeft als doel bepaalde demografische gegevens te verkrijgen om een eventuele marktsegmentatie toe te staan. De gegevens die beschikbaar zijn: het aantal uren dat men per week televisie kijkt, of men wel eens een 3D-film heeft gezien in de bioscoop, geslacht, leeftijd, hoogst voltooide opleiding, netto inkomen en de prijs die men bereid is om voor een nieuwe 3D-televisie zonder bril te betalen. Er kan dan bijvoorbeeld worden bestudeerd of er significante verschillen bestaan in de voorkeuren voor de productconcepten tussen mannen die veel televisie kijken en vrouwen die weinig televisie kijken. Het is nuttig dat deze extra informatie voorhanden is, maar er is hier in deze scriptie niets mee gedaan. Voor verder onderzoek zou er zeker wat mee gedaan kunnen worden. Kanttekening hierbij is dat het merendeel van de respondenten een universitaire opleiding volgt, waardoor vooral de leeftijd, hoogst voltooide opleiding en het netto inkomen niet representatief zijn voor de gemiddelde Nederlander. Echter, dit onderzoek is bedoeld voor de 'early adapters' en dan is het naar mijn mening minder erg, want daar vallen vooral veel relatief jongere mensen onder.

Er zijn enkele transformaties van de data doorgevoerd om er eenvoudig mee te kunnen werken. Zo zijn de keuzes binnen iedere deelverzameling omgezet naar de keuzes voor bepaalde productconcepten. De toegepaste analyse van de data staat beschreven in paragraaf 2.4. De resultaten zullen in hoofdstuk 4 uitgebreid aan bod komen.

Hoofdstuk 4

Resultaten

4.1 Resultaten analyse

In dit hoofdstuk zullen de resultaten besproken worden van de toepassing van de 'line scaling'-methode en de twee verschillende 'best-worst scaling'-methoden. Daaropvolgend zullen de verschillen in de resultaten van beide methoden worden besproken om een uitspraak te kunnen doen naar welke methode de voorkeur uitgaat.

4.1.1 Resultaten 'line scaling'

Bij de methode van 'line scaling' moesten de respondenten ieder productconcept een cijfer van 1 tot en met 10 geven. Een hoger cijfer betekent dat de voorkeur naar dit product groter is. Vervolgens werd toen de gemiddelde score voor ieder productconcept over alle 137 respondenten bepaald door het gemiddelde cijfer uit te rekenen. Door de gemiddelde scores te ordenen van hoog naar laag wordt duidelijk welk productconcept de meeste voorkeur tot aan de minste voorkeur geniet. Zie tabel (4).

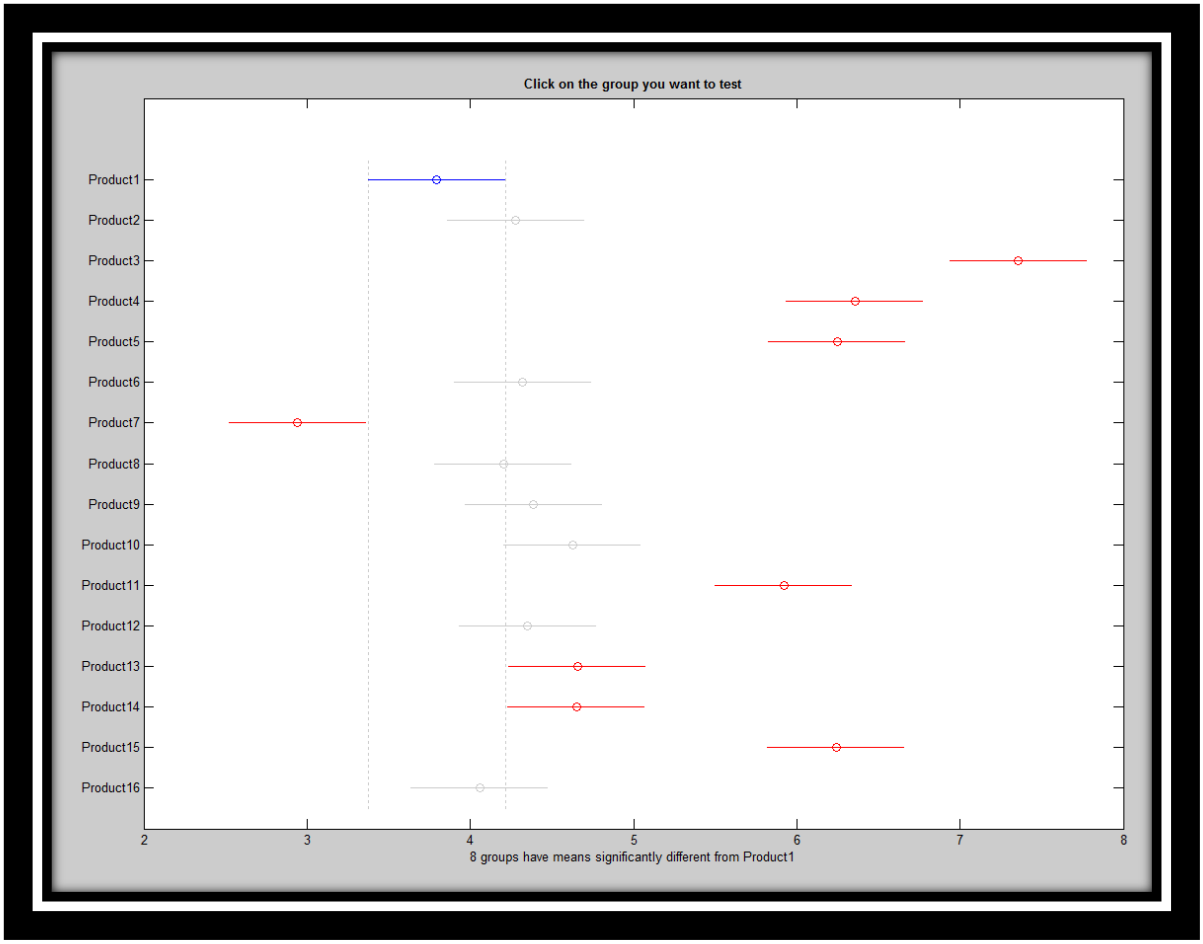
productconcept	gemiddelde score
3	7,36
4	6,36
5	6,25
15	6,24
11	5,92
13	4,66
14	4,65
10	4,63
9	4,39
12	4,35
6	4,32
2	4,28
8	4,20
16	4,06
1	3,80
7	2,94

Tabel (4) Resultaten 'line scaling'

Uit de bovenstaande tabel volgt dat productconcept 3 de hoogste score heeft en dus als beste uit de bus komt. Uit tabel (2) volgt dat het dan gaat om het productconcept met de

volgende kenmerken: A-merk, schermdiagonaal tot 40 inch, pc- en internetverbinding, lage prijs en hoge resolutie. Het is niet vreemd dat dit productconcept als beste is gekozen, want het lijkt te beschikken over vier van de vijf meest gunstige kenmerken. Productconcept 7 scoort het laagste. Dit was te verwachten, want dit productconcept heeft de volgende kenmerken: B-merk, schermdiagonaal tot 40 inch, geen pc- en internetverbinding, hoge prijs en lage resolutie. Verder kan men erover discussiëren of een relatief kleine schermdiagonaal juist gunstig of ongunstig is.

De nulhypothese van geen verschillen in voorkeur tussen de productconcepten moet worden verworpen, want uit de 'one way ANOVA' blijkt dat de toetsingsgrootheid F de waarde 45,49 aanneemt met een bijbehorende p-waarde van 0. De betrouwbaarheidsintervallen voor de gemiddelde scores per productconcept zijn weergegeven in figuur 1. Op de x-as staat de gemiddelde score en op de y-as staan onder elkaar de verschillende productconcepten. Als een betrouwbaarheidsinterval van een bepaald productconcept overlapt met die van een ander productconcept, dan geldt dat er geen significant verschil in voorkeuren bestaat. Voor productconcept 1, aangegeven met blauw, geldt dat dit het geval is voor productconcepten 2, 6, 8, 9, 10, 12 en 16. De bijbehorende betrouwbaarheidsintervallen van deze productconcepten worden dan grijs gemaakt in de figuur. Een rood betrouwbaarheidsinterval geeft aan dat er een significant verschil is. Dat is het geval voor de overige productconcepten. Deze meervoudige vergelijkingsprocedures kunnen voor ieder productconcept worden uitgevoerd.



Figuur 1 Plot van de betrouwbaarheidsintervallen bij toepassing van de 'line scaling'-methode

4.1.2 Resultaten 'best-worst scaling' methode 1

Bij de eerste methode van 'best-worst scaling' moest iedere respondent voor elk van de zestien deelverzamelingen aangeven welk productconcept het beste en het slechtste was. Ieder productconcept kwam in totaal zes keer voor, dus per individu geldt dat een productconcept minimaal nul en maximaal zes keer als beste resp. slechtste gekozen kon worden. Dat resulteerde vervolgens per productconcept in een individuele 'best-worst'-score tussen -6 en 6. Vervolgens zijn de gegevens van alle 137 respondenten bijeen genomen om te komen tot de resultaten van de gehele sample. Door de 'best-worst'-scores te ordenen van hoog naar laag wordt duidelijk welk productconcept de meeste voorkeur tot aan de minste voorkeur geniet. De resultaten hiervan staan vermeldt in tabel (5).

Product-concept	Best scores	Worst scores	B-W scores	Gemiddelde B-W scores	B/W	Wortel (B/W)	Relatief belang
3	489	11	478	0,58	44,45	6,67	100,00
4	374	31	343	0,42	12,06	3,47	52,10
5	312	29	283	0,34	10,76	3,28	49,19
11	252	23	229	0,28	10,96	3,31	49,65
15	237	15	222	0,27	15,80	3,97	59,62
14	73	60	13	0,02	1,22	1,10	16,54
13	85	86	-1	0,00	0,99	0,99	14,91
10	42	90	-48	-0,06	0,47	0,68	10,25
6	86	138	-52	-0,06	0,62	0,79	11,84
9	70	158	-88	-0,11	0,44	0,67	9,98
12	25	125	-100	-0,12	0,20	0,45	6,71
8	53	163	-110	-0,13	0,33	0,57	8,55
1	18	173	-155	-0,19	0,10	0,32	4,84
16	23	199	-176	-0,21	0,12	0,34	5,10
2	32	265	-233	-0,28	0,12	0,35	5,21
7	21	626	-605	-0,74	0,03	0,18	2,75

Tabel 5 Resultaten 'best-worst scaling' methode 1

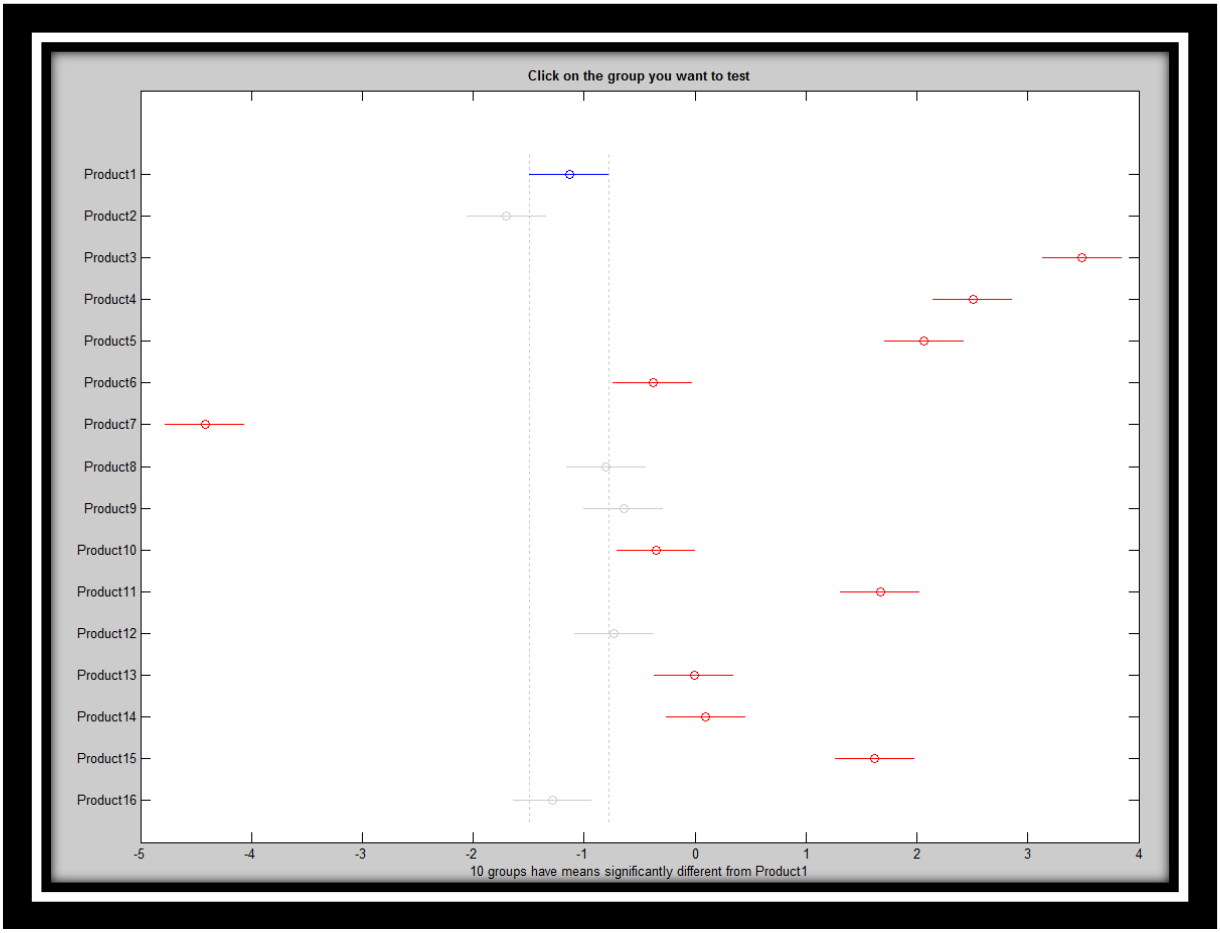
Wat opvalt bij de ordening van hoog naar laag op basis van de 'best-worst'-scores, die staan beschreven in de vierde kolom, is dat de ontstane rangordening van productconcepten niet hetzelfde is als de rangorde die zou ontstaan als men bijvoorbeeld alleen zou ordenen op basis van de 'best'-scores. De waarde die hoort bij productconcept 13 is namelijk hoger dan die van productconcept 14; 85<73. Voor de 'worst'-scores geldt juist dat de bijbehorende waarden van de productconcepten met de hoogste voorkeur afwijken, want daar geldt pas vanaf productconcept 8 een stijgend verloop in waarde.

Een alternatieve manier om de mate van voorkeur tussen verschillende productconcepten te vergelijken is door het bepalen van de relatieve ratioscores. Om deze te verkrijgen moet de wortel worden genomen van het quotiënt van het totaal aantal 'best'-scores en 'worst'-scores voor iedere respondent.

Aangezien de 'best'-scores en de 'worst'-scores niet monotoon dalend resp. stijgend zijn, geldt dat het quotiënt hiervan niet monotoon zal verlopen. Dat is te zien in de zesde kolom. Daar valt vooral op dat de waarde behorende bij productconcept 15 nog hoger is dan de waarde van productconcept 4. Dit heeft tot effect dat het relatieve belang van productconcept 15 groter is dan die van productconcept 4 en dat is iets wat je niet zou verwachten op basis van de verkregen ordening van de productconcepten.

Hier geldt wederom dat de meeste voorkeur uitgaat naar productconcept 3 en de minste voorkeur naar productconcept 7. Aan de hand van de 'best-worst'-scores valt op te merken dat er slechts zes productconcepten zijn waar een positieve waarde bij hoort en wat correspondeert met een productconcept dat men (relatief) graag zou willen hebben. Het gaat hierbij om productconcepten 3, 4, 5, 11, 15 en 14. Productconcept 13 kan worden opgemerkt als een soort grensgeval, aangezien de 'best-worst'-score dan praktisch 0 is.

De nulhypothese van geen verschillen in voorkeur tussen de productconcepten moet worden verworpen, want uit de 'one way ANOVA' blijkt dat de toetsingsgrootheid F de waarde 168,85 aanneemt met een bijbehorende p-waarde van 0. In figuur 2 worden de betrouwbaarheidsintervallen voor de gemiddelde individuele 'best-worst'-scores beschreven. Er valt op dat er nu minder overlapping is van de betrouwbaarheidsintervallen dan in figuur 1. Dat houdt in dat er meer significante verschillen in voorkeuren zijn tussen de verschillende productconcepten. In figuur 1 is bijvoorbeeld te zien dat er 8 productconcepten waren die significant van elkaar verschillen, maar dat zijn er nu 10.



Figuur 2 Plot van de betrouwbaarheidsintervallen bij toepassing van 'best-worst scaling'-methode 1

4.1.3 Resultaten 'best-worst scaling' methode 2

Bij de tweede methode van 'best-worst scaling' moest iedere respondent voor elk van de zestien deelverzamelingen aangeven welk productconcept het beste en het slechtste was. De beste kreeg als ranking de waarde 1 en de slechtste 6. De overige vier productconcepten kregen als ranking de waarde 3,5. Ieder productconcept kwam in totaal zes keer voor, dus per individu geldt dat een productconcept minimaal 6 en maximaal 36 als totale waarde kreeg. De gemiddelde individuele rank zit dan weer tussen 1 en 6 in. De gemiddelde rank voor ieder productconcept is bepaald door het gemiddelde te nemen over alle 137 respondenten. Door de waarde van de gemiddelde rank te ordenen van laag naar hoog wordt duidelijk welk productconcept de meeste voorkeur tot aan de minste voorkeur geniet. De resultaten hiervan staan vermeldt in de derde kolom van tabel (6). Door bij alle individuele rankings te corrigeren voor de gemiddelde rank, worden de aangepaste rankings verkregen. Door te sommeren over alle respondenten worden de waarden uit de tweede kolom verkregen. Wat uit de tabel duidelijk wordt is dat een lagere waarde in de tweede kolom correspondeert met een lagere gemiddelde rank en dus een hogere voorkeur.

productconcept	aplus	gemiddelde rank
3	-199,17	2,05
4	-142,92	2,46
5	-117,92	2,64
11	-95,42	2,80
15	-92,50	2,82
14	-5,42	3,46
13	0,42	3,50
10	20,00	3,65
6	21,67	3,66
9	36,67	3,77
12	41,67	3,80
8	45,83	3,83
1	64,58	3,97
16	73,33	4,04
2	97,08	4,21
7	252,08	5,34

Tabel 6 Resultaten 'best-worst scaling' methode 2

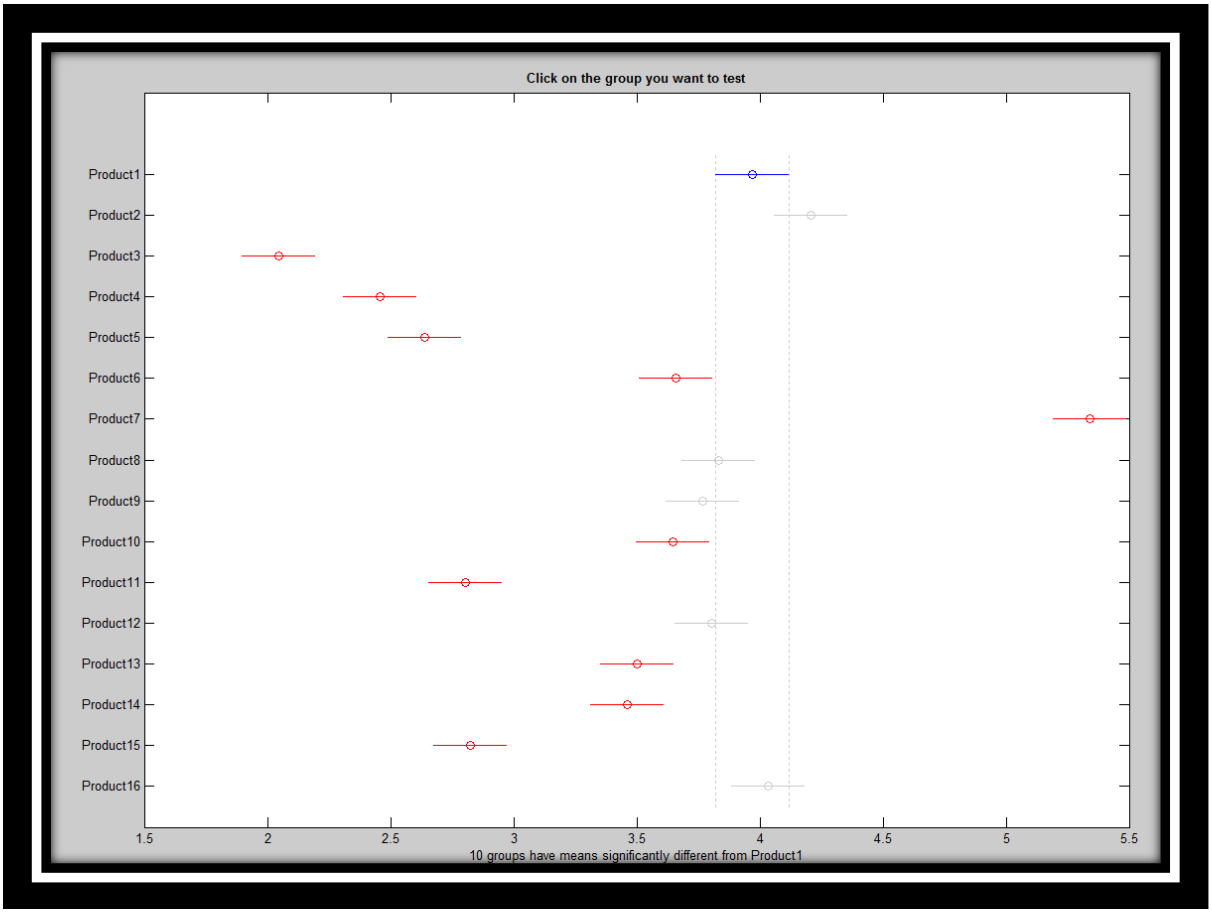
Hier geldt wederom dat de meeste voorkeur uitgaat naar productconcept 3 en de minste voorkeur naar productconcept 7. Aan de hand van de gemiddelde ranks valt op te merken dat er slechts zes productconcepten zijn waar een ranking bij hoort die lager is dan de gemiddelde ranking 3,5; wat correspondeert met een productconcept dat men (relatief) graag zou willen hebben. Het gaat hierbij om productconcepten 3, 4, 5, 11, 15 en 14. Dit zijn dezelfde productconcepten die in 4.1.2 een positieve 'best-worst'-score hebben.

De nulhypothese van geen verschillen in voorkeur tussen de productconcepten moet worden verworpen, want uit de 'one way ANOVA' blijkt dat de toetsingsgrootheid F de waarde 168,85 aanneemt met een bijbehorende p-waarde van 0. Dit is exact hetzelfde als de gevonden waarde voor F in 4.1.2. De reden daarvoor is dat de methoden aan elkaar verwant zijn. De individuele gemiddelde rank van een bepaald productconcept kan namelijk worden bepaald door vergelijking (10) te veranderen in het volgende:

$$x_{i,j} = \frac{1}{r}\Big(\sum_{k=1}^{K} I[\,productconcept\ \ j\ is\ de\ beste\ uit\ deelverzameling\ \ k\,] + r\sum_{k=1}^{K} I[\,productconcept\ \ j\ is\ de\ slechtste\ uit\ deelverzameling\ \ k\,] + \frac{r+1}{2} * \Big(r - \sum_{k=1}^{K} I[\,productconcept\ \ j\ is\ de\ beste\ uit\ deelverzameling\ \ k\,] - \sum_{k=1}^{K} I[\,productconcept\ \ j\ is\ de\ slechtste\ uit\ deelverzameling\ \ k\,]\Big)\Big) \qquad (17)$$

$$i = 1,\ldots,I \quad j = 1,\ldots J$$

Hier geldt r = 6 en K = 16, omdat de productconcepten in totaal 6 keer voorkomen in de 16 deelverzamelingen. Door te werken met de bovenstaande formule wordt in feite een koppeling gemaakt met de eerste methode van BWS. Dit leidt dan ook tot dezelfde F-waarde. De waarde voor W is gelijk aan 526,30. De significante verschillen in voorkeuren zijn hetzelfde als bij de eerste BWS-methode, maar de betrouwbaarheidsintervallen zijn anders (zie figuur 3). Dit model kan mogelijk worden opgevat als makkelijker qua interpretatie.



Figuur 3 Plot van de betrouwbaarheidsintervallen bij toepassing van 'best-worst scaling'-methode 2

4.2 Verschillen in discriminatie

Na het bepalen van de totale rankings is het belangrijk om na te gaan of productconcepten significant van elkaar verschillen in voorkeuren. In de vorige paragraaf werd duidelijk dat er productconcepten zijn die significant van elkaar verschillen, want de gebruikte toetsingsgrootheden werden allemaal verworpen. Door het bekijken van de betrouwbaarheidsintervallen van de productconcepten in de figuren 1 t/m 3 is duidelijk gemaakt welke productconcepten precies significant van elkaar verschillen. Om dit overzichtelijker te maken kunnen productconcepten worden ingedeeld in clusters. Binnen een cluster geldt dat de productconcepten niet significant van elkaar verschillen, maar daarbuiten wel. Het aantal clusters en het aantal productconcepten per cluster zijn een maat voor de discriminatie. Hoe meer clusters er zijn en/of des te kleiner het aantal productconcepten per cluster, des te sterker de discriminatie. In tabel (7) wordt de clustering beschreven als er gebruik is gemaakt van de methode van 'line scaling'.

Product-concept	cluster 1	cluster 2	cluster 3a	cluster 3b	cluster 4
3	x				
4		x			
5		x			
15		x			
11		x			
13			x		
14			x		
10			x	x	
9			x	x	
12			x	x	
6			x	x	
2			x	x	
8			x	x	
16			x	x	
1				x	
7					x

Tabel 7 Clustering bij 'line scaling'

De bovenstaande tabel dient als volgt te worden geïnterpreteerd: productconcepten met een kruisje in dezelfde kolom zijn niet significant verschillend qua voorkeur. Er is hierbij gewerkt met de Tukey-B-toets met een significantielevel $\alpha = 0{,}05$. Wat opvalt is dat er vier disjuncte clusters zijn. Cluster 1 bevat productconcept 3 en dat is degene waar de meeste voorkeur naar uitgaat. In cluster 2 bevinden zich productconcepten 4, 5, 11 en 15 en deze zijn onderling niet significant verschillend. De derde cluster is in twee subclusters gesplist. De reden hiervoor is dat productconcepten 13 en 14 niet significant verschillend zijn van de

productconcepten 10, 9, 12, 6, 2, 8 en 16 maar wel verschillen met productconcept 1. Productconcept 7 is het minst geprefereerde productconcept en zit in die vierde cluster.

Als je overgaat op de methode van 'best-worst scaling' om de clustervorming te bepalen, dan maakt het niet uit of er gebruik is gemaakt van de eerste of tweede methode. In tabel (8) zijn de resultaten beschreven:

Product-concept	cluster 1	cluster 2a	cluster 2b	cluster 3a	cluster 3b	cluster 3c	cluster 3d	cluster 3e	cluster 4
3	x								
4		x							
5		x	x						
11			x						
15			x						
14				x					
13				x	x				
10				x	x	x			
6				x	x	x			
9					x	x	x		
12						x	x		
8						x	x		
1							x	x	
16							x	x	
2								x	
7									x

Tabel 8 Clustering bij 'best-worst scaling'

Wat opvalt is dat er nu veel meer subclusters zijn. Cluster 1 en 4 blijven even groot en bevatten dezelfde productconcepten evenals de andere clusters maar daar is er wel het een en ander veranderd. Als men kijkt naar de volgorde in cluster 2, dan was dat in het eerste geval {4, 5, 15, 11} . Echter, dat is in het geval van BWS {4, 5, 11, 15} geworden. De laatste twee productconcepten zijn dus van plaats verwisseld. Door de vorming van twee subclusters in cluster 2 is op te merken dat de productconcepten {4, 11} en {4, 15} nu significant verschillend zijn qua voorkeur. In cluster 3 is er veel meer veranderd. In het geval van 'line scaling' valt op dat er slechts 2 subclusters zijn, waaruit af te leiden valt dat productconcept 1 significant verschillend is van productconcepten 13 en 14. Daarnaast is de volgorde van de productconcepten behoorlijk veranderd. Er zijn nu veel meer productconcepten significant verschillend in voorkeuren en dat zijn: {14, 9}, {14, 12}, {14, 8}, {14, 16}, {14, 2}, {13, 12}, {13, 8}, {13, 16}, {13, 2}, {10, 1}, {10, 16}, {10, 2}, {6, 1}, {6, 16}, {6, 2}, {9, 2}, {12, 2} en {8, 2}. In totaal leidt de toepassing van 'best-worst scaling' dus tot 20 extra paren aan productconcepten die significant verschillend zijn qua voorkeur. Er is zodoende, op de komende markt van 3D-televisies zonder bril, aangetoond dat de methode van 'best-worst scaling' leidt tot een sterkere discriminatie dan de methode van 'line scaling'.

Hoofdstuk 5

Conclusies

5.1 Conclusies

Marktonderzoekers bestuderen hoe er nog beter aan de wensen van de consumenten kan worden voldaan om het aanbod op de vraag af te stemmen. Daarvoor moet men het belang en de voorkeuren van diverse productattributen meten en voorspellen. Er wordt 'stated preference' data verzameld via aselecte steekproeven en er worden schalen gebruikt om de consumentenvoorkeuren te kunnen begrijpen. Het kiezen van de juiste schaal is hierbij cruciaal.

Veruit de bekendste schaal is de zogenaamde 'line scale'. De motivatie voor deze schaal is dat het voor de respondenten en de marktonderzoekers eenvoudig is om hiermee te werken. Bij de analyse worden dan de gemiddelde cijfers per attribuut berekend en met elkaar vergeleken. Ondanks dat zijn er toch een aantal bezwaarlijke redenen te noemen om van de methode van 'line scales' af te wijken. Het is namelijk zo dat de beoordelingen van respondenten afhankelijk zijn van de manier waarop ze 'ratings' opvatten. Dit verschilt per individu, maar ook voor diverse segmenten waaruit de markt is opgebouwd. Het is belangrijk om aan te kunnen tonen of bepaalde attributen significant belangrijker zijn dan andere attributen. Als er gewerkt zal worden met een 'line scale', dan is het vaak lastig om hier een uitspraak over te kunnen doen. De reden daarvoor is dat respondenten soms ieder attribuut even (on)belangrijk vinden. Er zal dan een zwakke discriminatie tot stand komen tussen de verschillende attributen. Op deze manier kan men dan geen betrouwbare conclusies trekken over het relatieve belang van bepaalde attributen, aangezien er geen mogelijkheid is om een juiste afweging te maken tussen de desbetreffende attributen.

Om de grootste problemen van hierboven te voorkomen kan men overgaan op de methode van 'best-worst scaling'. De verschillende attributen worden dan in unieke deelverzamelingen ingedeeld volgens een toereikend gebalanceerd incompleet blokontwerp. Er geldt dan dat ieder attribuut even vaak voorkomt en dat ieder paar aan attributen even vaak voorkomt. De respondent moet dan steeds per deelverzameling van attributen kiezen welke de meest geprefereerde en de minst geprefereerde is. Aangezien respondenten constant afwegingen moeten maken tussen verschillende attributen heeft dit tot effect dat de discriminatie sterker zal zijn. De methode van 'best-worst scaling' maakt het toegankelijk om op een eenvoudige manier een ranking per individu aan te maken. Door de rankings van de gehele sample te bestuderen kan een complete ranking van alle attributen worden verkregen.

In deze scriptie is een empirisch onderzoek gedaan naar de consumentenvoorkeuren van 3D-televisies zonder bril. Er is hiervoor data verzameld middels een enquête en dat is gebruikt om de verschillen tussen 'line scaling' en 'best-worst scaling' te bestuderen. Voor de 'best-worst scaling' zijn twee aan elkaar gerelateerde methoden toegepast. Men moest de voorkeuren kenbaar maken voor zestien verschillende productconcepten. Een productconcept is hier een unieke combinatie van vijf attributen: merk, schermdiagonaal, pc- en internetverbinding, prijsklasse en resolutie. Door het gebruik van een gebalanceerd incompleet blokontwerp werden de verschillende deelverzamelingen aangemaakt waarbij de respondenten moesten aangeven welk productconcepten het meeste en minste werden geprefereerd. Daarnaast moest aan ieder productconcept onafhankelijk een cijfer van 1 t/m 10 worden gegeven voor de mate van voorkeur. Na de analyse konden de verschillen tussen de diverse schaalmethoden worden onderzocht.

De ranking van alle productconcepten na toepassing van de twee verschillende schaalmethoden leek in veel opzichten op elkaar. Bij beiden kwam bijvoorbeeld productconcept 3 als beste naar voren en productconcept 7 als slechtste. De volgorde voor de tussenliggende productconcepten is niet helemaal hetzelfde, maar de verschillen zijn niet problematisch. Waar het vooral om gaat is of productconcepten significant van elkaar verschillen in mate van voorkeur. Dat wordt duidelijker als er wordt gewerkt met clustering. De productconcepten worden dan allemaal ingedeeld in clusters. Binnen een cluster geldt dat de voorkeuren niet significant verschillend zijn, maar dat is buiten de cluster wel het geval. Des te groter het aantal clusters en des te kleiner het aantal productconcepten per cluster, des te sterker de discriminatie. Bij de methode van 'line scaling' worden er vier clusters verkregen, waarbij cluster 1 uit productconcept 3 bestaat, cluster 4 uit productconcept 7, cluster 2 uit productconcepten 4, 5, 11 en 15 en cluster 3 uit de rest. Verder geldt dat binnen cluster 3 een splitsing is aangemaakt in twee subclusters. Bij de twee methoden van 'best-worst scaling' worden ook vier clusters verkregen met exact dezelfde productconcepten per cluster. Het enige en belangrijke verschil is dat er veel meer subclusters zijn. Cluster 2 bestaat nu uit twee subclusters en cluster 3 uit maar liefst vijf subclusters. In totaal leidt de toepassing van 'best-worst scaling' tot 20 extra paren aan productconcepten die significant verschillend zijn qua voorkeur.

Er is zodoende, op de komende markt van 3D-televisies zonder bril, aangetoond dat de methode van 'best-worst scaling' leidt tot een sterkere discriminatie van voorkeuren dan de methode van 'line scaling'.

5.2 Beperkingen

Er zijn enkele beperkingen aan het gebruiken van de methode van ‘best-worst scaling’. Uit de praktijk blijkt dat het optimaal is om het aantal deelverzamelingen niet al te groot te kiezen, aangezien de verveling dan een rol kan gaan spelen. Dit kan een verklaring zijn voor het feit dat 60 van de 197 mensen die deelnamen aan de uitgevoerde enquête het niet helemaal hebben afgerond. Het kan er ook toe leiden dat respondenten mogelijk proberen om extra op te letten of hun antwoorden consistent zijn. Om de methode van ‘best-worst scaling’ te gebruiken moet er dus op worden gelet dat het aantal deelverzamelingen niet te groot is. Bovendien dient het aantal productconcepten per deelverzameling niet al te groot te zijn, want anders kan het mogelijk te vermoeiend worden om de beste en slechtste optie per deelverzameling te kiezen.

In het empirische onderzoek zijn er nog enkele beperkingen bijgekomen door de indeling van de enquête. De volgorde van het stellen van de verschillende type vragen was voor iedereen hetzelfde. Ik had dit graag random gemaakt, maar dat niet helaas niet mogelijk via de tool voor het online afnemen van de enquête. Het was ook niet mogelijk om de volgorde van de vragen waarbij een cijfer gegeven moest worden te laten variëren. Hetzelfde geldt voor de volgorde waarbij de deelverzamelingen voorkwamen waarbij de beste en slechtste productconcepten gekozen moesten worden. Als dit alles wel mogelijk was geweest, dan had dit voor het onderzoek nog beter geweest.

5.3 Richtlijnen voor verder onderzoek

In deze scriptie is gewerkt met de methode van ‘best-worst scaling’. Er zijn andere rankmethoden voorhanden, waarbij niet alleen de beste en slechtste optie per deelverzamelingen worden gekozen. Je zou ook alle opties per deelverzameling kunnen ranken en op basis daarvan komen tot de uiteindelijke totale ranking van alle opties. Deze methode vereist meer afwegingen en kan mogelijk leiden tot een betere discriminatie, ook al is bekend dat respondenten er mogelijk moeite mee hebben om niet-extreme opties te ranken wat vervolgens kan leiden tot voortijdig afhaken en inconsistentie.

Om enkele beperkingen van de enquête te omzeilen zou het een idee zijn om deze op te splitsen in twee delen, waarbij een respondent random een van de twee versies krijgt. Ieder deel bevat dan dezelfde vragen, maar dan met de helft van het totale aantal productconcepten die onderling met elkaar vergeleken moeten worden. Je krijgt op die manier mogelijk meer data, omdat er minder respondenten afhaken. Dit gaat wel ten koste van de informatie die een respondent zou kunnen geven met betrekking tot de andere productconcepten. Het integreren van de twee delen kan ook lastig worden.

Wat men verder nog zou kunnen onderzoeken is welke productconcepten de voorkeur genieten bij bepaalde segmenten van de markt. Dat is interessant voor verschillende marketingcampagnes die zich richten op bepaalde doelgroepen. In de enquête zijn demografische gegevens gevraagd, waardoor analyse van rankings per marktsegment mogelijk kan worden gemaakt.

Referenties

Ben-Akiva, M., Morikawa, T., en Shiroishi, F. (1992). "Analysis of the reliability of preference ranking data". *Journal of Business Research,* Vol. 24 No. 2, pp. 149-164.

Cohen, S.H. (2003). "Maximum difference scaling: improved measures of importance and preference for segmentation." *Sawtooth Software Conference Proceedings, Sequiem, WA,* pp. 61-74.

Cohen, S.H. en Neira, L. (2003). "Measuring preference for product benefits across countries: overcoming scale usage bias with maximum difference scaling." *ESOMAR 2003 Latin America Conference Proceedings, Amsterdam,* pp. 1-22.

Cohen, S.H. en Orme, B. (2004). "What's your preference?" *Marketing Research,* Vol. 21 No. 1, pp. 50-63.

Couch, A. en Keniston, K. (1960). "Yeasayers and naysayers: agreeing response set as a personality variable." *Journal of Abnormal and Social Psychology,* Vol. 60, March, pp. 151-174.

Crask, M.R. en Fox, R.J. (1987). "An exploration of the interval properties of three commonly used marketing research studies: a magnitude estimation approach". *Journal of the Marketing Research Society,* Vol. 29 No. 3, pp. 317-339.

DeShazo, J.R. en Fermo, G. (2002). "Designing choice sets for stated preference methods: the effects of complexity on choice consistency." *Journal of Modelling in Management,* Vol. 5 No. 2, pp. 94-123.

Finn, A. en Louviere, J.J. (1992). "Determining the appropriate response to evidence of public concerns: the case of food safety." *Journal of Public Policy and Marketing,* Vol. 11 No. 1 pp. 12-25.

Hein, K.A., Jaeger, S.R., Carr, B.T. en Delahunty, C.M. (2008). "Comparison of five common acceptance and preference methods." *Food Quality and Preferences,* Vol. 19, pp. 579-588.

Iyengar, S.S. en Lepper, M.R. (2000). "When choice is de-motivating: can one desire too much of a good thing?" *Journal of Personality and Social Psychology,* Vol. 79, pp. 995-1006.

Lam, K.Y. (2011). "Reliability and rankings." *ERIM PhD Series in Research in Management 230*, pp. 51-65.

Louviere, J.J. en Woodworth, G. (1983). "Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data." *Journal of Marketing Research,* Vol. 20, pp. 350-367.

Louviere, J.J., Hensher, D.A. en Swait, J.D. (2000). "Stated choice methods: analysis and applications." *Cambridge University Press, Cambridge.*

Marley, A.A.J. en Louviere, J.J. (2005). “Some probabilistic models of best, worst, and best worst choices.” *Journal of Mathematical Psychology,* Vol. 49, pp. 464-480.

Thurstone, L.L. (1927). “A law comparative judgment.” *Psychological Review,* Vol. 34, pp. 273-286.

Weller, S.W. en Romney, A.K. (1988). “Systematic Data Collection.” *Sage, Newbury Park, CA.*