Vragenlijsten: van ontwikkeling tot kwaliteitsverbetering

Fouad Rmila

Vragenlijsten:
van ontwikkeling tot kwaliteitsverbetering

Fouad Rmila

BWI werkstuk

Faculteit der Exacte Wetenschappen
Studierichting Bedrijfswiskunde en Informatica
De Boelelaan 1081a
1081 HV Amsterdam

Begeleider:
Dennis Roubos

mei 2008

Voorwoord

Het BWI werkstuk is een verplicht onderdeel van de studie Bedrijfswiskunde & Informatica (BWI). Het doel van het werkstuk is om op een heldere wijze een probleem te beschrijven voor een deskundige manager. Bij het schrijven van het werkstuk dient rekening te worden gehouden met het bedrijfsgerichte aspect van de studie en zal ook het wiskunde en/of het informatica aspect verwerkt moeten worden.

Tijdens een zomerstage bij Capgemini ben ik op het onderwerp van dit werkstuk gekomen. Gedurende deze stage is onderzoek gedaan naar het meten van de effectiviteit van architectuurafdelingen van bedrijven. Deze effectiviteit kan gemeten worden door middel van vragenlijsten, die door architecten en stakeholders van de architectuurafdeling worden ingevuld. Tijdens en na deze stage ben ik me verder gaan verdiepen in vragenlijstontwikkeling en heb ik ook gekeken naar wiskundige technieken om de kwaliteit van dergelijke vragenlijsten te beoordelen. Aangezien er nog geen dataverzameling had plaatsgevonden tijdens mijn stage, heb ik ervoor gekozen om zelf een klein onderzoek uit te voeren om de beschreven theorie in dit werkstuk toe te passen in een praktijksituatie.

Tijdens mijn stage heb ik gemerkt dat vragenlijsten een effectief meetinstrument kunnen zijn om bepaalde concepten te kwantificeren. Steeds meer bedrijven maken ook gebruik van dit middel. Het gekozen onderwerp houdt zodoende rekening met de bedrijfsgerichte component van de BWI opleiding. De wiskunde component komt terug in de technieken, die gebruikt kunnen worden, om de kwaliteit van vragenlijsten te beoordelen. Ook is de informatica component in dit werkstuk verwerkt door de resultaten van een zelfontwikkelde vragenlijst met het statistisch software pakket R te analyseren.

Tot slot wil ik graag Dennis Roubos bedanken voor zijn inzet en feedback tijdens de begeleiding van dit werkstuk.

Fouad Rmila
Amsterdam, mei 2008

Management samenvatting

In de psychologie, maar ook in andere wetenschappen, worden regelmatig constructen (zoals tevredenheid, vertrouwen, et cetera) bestudeerd die men niet rechtstreeks kan observeren of meten. Via observeerbare gedragingen, die op één of andere manier gerelateerd zijn aan het construct, kan getracht worden deze meetbaar te maken. Inzicht in deze observeerbare gedragingen verkrijgt men bijvoorbeeld via responsen op een vragenlijst. Om tot een kwalitatief goede vragenlijst te komen, dient de constructie van dit meetinstrument gestructureerd plaats te vinden. Hierbij is het vooral belangrijk dat het doel van de vragenlijst van te voren helder en goed afgebakend is. Wat betreft de formulering van de vragen dient uiteraard dubbelzinnigheid voorkomen te worden. Slecht geformuleerde vragen leiden namelijk tot invalide onderzoeksresultaten. Ook is het van belang dat een vragenlijst uitvoerig getest wordt alvorens deze aan de doelgroep wordt uitgereikt.

In dit werkstuk zijn Likertschalen beschouwd om constructen meetbaar te maken. Bij dit type antwoordschaal worden de responsen op verscheidene items gecombineerd tot een enkele score voor het onderzochte construct. De verzameling items, die gebruikt worden om het construct te meten, vormt een (meet)schaal. De items in een schaal dienen een sterke samenhang te vertonen, aangezien aan de hand van een schaal doorgaans een enkel algemeen kenmerk wordt gemeten. Om te onderzoeken in hoeverre de items in een vragenlijst hetzelfde concept meten, kan Cronbach‟s α worden berekend. Deze coëfficiënt, die af te leiden is met behulp van resultaten uit de klassieke testtheorie, geeft tevens een indicatie van de betrouwbaarheid van een Likertschaal vragenlijst.

In de ontwikkeling van een Likertschaal vragenlijst is het raadzaam om de betrouwbaarheid (mate waarin een schaal dezelfde resultaten geeft bij herhaalde toepassing onder dezelfde voorwaarden) en validiteit (mate waarin de schaal meet wat het behoort te meten) ervan te evalueren aan de hand van de antwoorden van proefrespondenten. Cronbach‟s α betrouwbaarheidscoëfficiënt wordt bij vragenlijstonderzoek ook veelvuldig gerapporteerd, ondanks dat het vermelden van deze puntschatter alleen misleidend kan zijn. In dit werkstuk is daarom ook onderzocht hoe een betrouwbaarheidsinterval voor Cronbach‟s α kan worden afgeleid. In tegenstelling tot betrouwbaarheid is het analyseren van de validiteit van een vragenlijst niet eenvoudig.

De responsen van de proefrespondenten kunnen ook gebruikt worden om de kwaliteit van de Likertschaal vragenlijst te verbeteren. Tijdens deze item-analyse wordt bijvoorbeeld onderzocht wat de invloed is van het verwijderen van een item uit de schaal op Cronbach‟s α. Een andere techniek kijkt naar de samenhang tussen een afzonderlijk item en de gehele schaal, gebaseerd op de overige items. Items die laag scoren op deze technieken kunnen vervolgens uit de vragenlijst worden verwijderd.

De besproken theorie is ten slotte toegepast in een praktijksituatie. Hiervoor is een vragenlijst geconstrueerd om de tevredenheid van studenten aan de Faculteit der Exacte Wetenschappen van de Vrije Universiteit te meten. Uit de Cronbach‟s α en het betrouwbaarheidsinterval van deze schatter bleek de originele vragenlijst een relatief goede betrouwbaarheid te hebben. Aan de hand van een item-analyse zijn desalniettemin zes kwalitatief ongunstige items uit de vragenlijst verwijderd. Op basis van de antwoorden van de respondenten zijn ook de resultaten van de vragenlijst geanalyseerd. Middels de Mann Whitney-toets is gebleken dat er geen wezenlijk verschil te ontdekken is tussen de algemene tevredenheid van bachelor en master studenten en tussen die van mannelijke en vrouwelijke studenten. Wel is aangetoond dat studenten die relatief lang staan ingeschreven voor een studie over het algemeen minder tevreden zijn hierover dan studenten die relatief kort staan ingeschreven.

Executive summary

In psychology, but also in other sciences, constructs (like satisfaction, confidence, et cetera) that one cannot directly observe nor measure are regularly being studied. Via observable conduct that is somehow related to the construct a researcher can try to measure it. Understanding of this observable conduct is obtained through responses on a questionnaire for example. To create a questionnaire of good quality the construction of this measuring instrument has to be done in a structured way. During this process it is especially important that the purpose of the questionnaire is clear and correctly defined beforehand. When the questions are formulated one has to avoid ambiguity of course, because poor formulated questions will lead to invalid research results. It is also important that the questionnaire is extensively tested before it is handed out to the target group.

In this thesis Likert scales are considered which can be used to make constructs measurable. A characteristic of this type of response scale is that several item responses may be combined to create a single score for the construct examined. The set of items, that is used to measure the construct, is also called a (measurement) scale. The items in a scale need to be strong related to each other, since a scale is often used to measure a single general concept. To examine to what extent the items in a questionnaire measure the same concept Cronbach‟s α can be calculated. This coefficient, which can be derived from results of the classical test theory, provides an indication of the reliability of a Likert scale questionnaire as well.

During the construction of a Likert scale questionnaire it is advisable to evaluate its reliability (extent to which a scale yields the same results each time it is used under the same conditions) and validity (extent to which the scale measures what it is supposed to measure) on the basis of responses of a test group. Therefore Cronbach‟s α reliability coefficient is reported frequently in survey researches, despite the fact that only applying this point estimator to the data can be misleading. That is why in this thesis we also explored how a confidence interval for Cronbach‟s α can be derived. In contrast to reliability, the analysis of validity of a questionnaire is much harder.

The responses of the test group can also be used to improve the quality of the Likert scale questionnaire. During this item analysis a researcher investigates for example what the effect is on Cronbach‟s α when an item is removed from the scale. Another technique looks at the correlations between separate items and the entire scale based on the remaining items. Subsequently the items with a low score on these techniques can be removed from the questionnaire.

Finally the theory discussed has been put into practice. For this end a questionnaire to measure the satisfaction of students at the Faculty of Sciences of the VU University Amsterdam has been developed. Based on Cronbach‟s α and the confidence interval of this estimator the original questionnaire appeared to have a relatively nice reliability. After carrying out an item analysis nevertheless six items of bad quality have been removed from the questionnaire. Based on the answers of the respondents we also have analysed the results of the questionnaire. With the Mann Whitney-test we have showed that in general there is no significant difference between the satisfaction of bachelor and master students and between the satisfaction of male and female students. On the other hand we have showed that students that are registered relatively long for their study are in general less satisfied about it than students that are registered relatively short.

Inhoudsopgave

Inleiding

In de sociale wetenschappen worden veelvuldig vragenlijsten gebruikt als meetinstrument in een onderzoek. Een vragenlijst is ook één van de mogelijkheden bij psychologisch onderzoek. Hierbij kan het bijvoorbeeld gaan om een zelfbeoordelingvragenlijst of een vragenlijst om iemand anders te beoordelen. Binnen het bedrijfsleven maakt men evenzo actief gebruik van vragenlijsten om bepaalde data te verzamelen. Hierbij kan men bijvoorbeeld denken aan kwantitatief marketing onderzoek of simpelweg aan tevredenheidonderzoeken onder het personeel. Tegenwoordig zien we dat vragenlijsten enorm populair zijn, mede omdat ze typische en originele informatie beschikbaar maken die niet via andere bronnen te verkrijgen is.

In dit werkstuk richten we ons vooral op vragenlijsten, die door de respondenten individueel ingevuld kunnen worden. Deze vragenlijsten worden ook wel *zelfgeadministreerde vragenlijsten* (self-administered surveys) genoemd. Een maatschappelijke trend is dat ieder jaar het aantal onderzoeken uitgevoerd met behulp van zelfgeadministreerde vragenlijsten, in het bijzonder via de mail, het aantal interviewonderzoeken overtreft. De redenen hiervoor zijn volgens [2, p.7] de lagere kosten van zelfgeadministreerde vragenlijsten en het feit dat organisaties dergelijke onderzoeken zelf kunnen uitvoeren, zonder een professionele organisatie hiervoor in te schakelen. Bij interviewonderzoeken, die telefonisch of face-to-face plaatsvinden, wordt meestal wel de hulp van een derde partij ingeroepen.

Middels een literatuurstudie zal in dit BWI werkstuk onderzoek worden gedaan naar hoe een zelfgeadministreerde vragenlijst op een gedegen manier kan worden geconstrueerd en hoe de kwaliteit ervan beoordeeld kan worden. Tevens zullen technieken besproken worden om de kwaliteit van het meetinstrument te verbeteren. In hoofdstuk 1 komen een aantal spelregels aan bod die bij de ontwikkeling van een willekeurige vragenlijst aandacht verdienen. In hoofdstuk 2 zal een veel voorkomende antwoordschaal, de *Likertschaal*, worden besproken. Hoofdstuk 3 staat in het teken van wiskundige technieken om de kwaliteit van Likertschaal vragenlijsten te beoordelen. De wiskundige afleiding van één van de technieken komt in Appendix A uitgebreid aan bod. Ook wordt in hoofdstuk 3 bekeken welke methoden aangewend kunnen worden om een vragenlijst te verbeteren. Ten slotte is de besproken theorie uit de hoofdstukken 1 tot en met 3 toegepast in een praktijksituatie. Een Likertschaal vragenlijst om de tevredenheid van studenten aan de Faculteit der Exacte Wetenschappen te onderzoeken, is ingevuld door studenten en vervolgens zijn de kwaliteit en de resultaten van deze vragenlijst geanalyseerd. Een uitgebreid verslag hierover is te lezen in hoofdstuk 4.

Hoofdstuk 1

Vragenlijsten

1.1 Introductie

Een vragenlijst is een systeem voor het verzamelen van informatie om kennis, standpunten, en gedrag te beschrijven, te vergelijken of uit te leggen [1]. Deze informatie wordt verzameld door een groep mensen vragen te laten beantwoorden en/of stellingen te laten beoordelen. Het opstellen van deze vragen/stellingen dient op een gestructureerde wijze plaats te vinden. In paragraaf 1.2 komt een stappenplan aan bod, dat gebruikt kan worden om structuur te brengen in de vragenlijstontwikkeling.

Bij het opstellen van een vragenlijst staat een onderzoeker voor veel keuzes. Eén van deze keuzes behelst de vraagvormen die in de vragenlijst gehanteerd zullen worden. De gekozen vraagvorm heeft onder andere gevolgen voor de analysemogelijkheden van de antwoorden op de vragenlijst. In paragraaf 1.3 wordt aandacht besteed aan een aantal vraagvormen.

Het construeren van een vragenlijst is uiteraard een kritieke succesfactor binnen een onderzoeksproces, aangezien het natuurlijk van belang is dat potentiële respondenten de juiste vragen gesteld krijgen, de vragen eenduidig interpreteren, in staat zijn om accuraat te antwoorden en bereid zijn om de vragen te beantwoorden. Goed geformuleerde vragenlijsten zijn, volgens [3], dan ook makkelijk in te vullen door de respondenten, verzamelen informatie accuraat en consistent, en verschaffen data die geanalyseerd kan worden om antwoord te geven op onderzoeksvragen. In paragraaf 1.4 komen een aantal regels aan bod, die gevolgd kunnen worden, om tot goed geformuleerde vragen te komen.

Het is wel bekend dat het uitvoeren van een onderzoek middels een vragenlijst een waardevolle methode is om van een grote groep respondenten informatie over allerlei zaken te verzamelen. Toch dient men hierbij niet te vergeten dat aan de hand van vragenlijsten meestal perceptuele informatie wordt verkregen; deze informatie is moeilijker te interpreteren dan feitelijke informatie. Andere voor- en nadelen van vragenlijsten worden in paragraaf 1.5 besproken.

1.2 Vragenlijst constructie

Het construeren van een vragenlijst is meestal een onderdeel van een groter onderzoeksproces. Dit onderzoeksproces kan uit de volgende stappen bestaan [4]:

- *Vaststellen van de doelen van het project*
- *Bepalen van de doelgroep*
- *Kiezen van een interview methode*
- *Construeren van de vragenlijst*
- *Uittesten van de vragenlijst*
- *Vragenlijst laten invullen door steekproef*
- *Data analyse*

Het ontwikkelen van een vragenlijst is geen vanzelfsprekende opgave. Wanneer het probleem dat men wil bestuderen van te voren onvoldoende is gedefinieerd, kan de vragenlijst bijvoorbeeld leiden tot onbruikbare of triviale resultaten. Ook blijkt dat vragenlijsten die ondermaats ontworpen zijn, doorgaans onbevredigende data produceren. Dit heeft als gevolg dat de onderzoeksvraag niet beantwoord kan worden. Een ondermaatse vragenlijst hanteert bijvoorbeeld de verkeerde vragen, een onlogische ordening van de vragen, incorrecte antwoordschalen, of een slechte vormgeving.

Om ervoor te zorgen dat vragenlijsten op een gestructureerde manier worden opgesteld zijn er verschillende stappenplannen bekend (onder andere in [3], [4] en [5]). In [5] zijn de volgende acht stappen beschreven:

***Stap 1**: Bepaal welke informatie nodig is.*

Om deze informatie te bepalen zal de onderzoeker eerst het doel van de vragenlijst moeten weten. Deze eerste stap is zeer belangrijk en het is daarom van belang om de tijd te nemen om het doel helder te krijgen.

***Stap 2**: Bepaal welke populatie(s) te ondervragen.*

In deze stap zal de onderzoeker na moeten gaan wie het meest waarschijnlijk over de informatie beschikt waar hij/zij naar op zoek is.

***Stap 3**: Stel de steekproef vast.*

Hierbij is het van belang om de juiste mensen voor het onderzoek te selecteren. Van de doelgroep zal de onderzoeker na moeten gaan welk deel toegankelijk is voor de vragenlijst (toegankelijke populatie). Met behulp van een sampling strategie kan een steekproef van de toegankelijke populatie getrokken worden.

***Stap 4**: Stel de vragenlijst op.*

Bij het opstellen van de vragenlijst is het natuurlijk van belang dat de juiste vragen op een eenduidige manier geformuleerd worden. Om dit te bewerkstelligen kunnen onder andere bepaalde regels, met betrekking tot de formulering van de vragen, gevolgd worden. Een aantal van deze regels zullen besproken worden in paragraaf 1.4.

***Stap 5**: Plan hoe de vragenlijst wordt verdeeld, geretourneerd en eventuele vervolgstappen.*

Tegenwoordig zijn er veel manieren om een vragenlijst onder de steekproef uit te zetten, bijvoorbeeld in een persoonlijk gesprek via de telefoon of face-to-face. Tegenwoordig worden steeds meer onderzoeken ook elektronisch (zoals email, websites) verspreid.

***Stap 6**: Schrijf een goede begeleidende brief.*

Deze brief of introductie over de vragenlijst heeft invloed op het aantal respondenten dat meedoet aan het onderzoek. In de brief is het van belang om informatie te verschaffen over de reden waarom de respondent de vragenlijst heeft ontvangen, wat het doel is van de vragenlijst, waarom hij/zij de vragenlijst zou moeten invullen, hoe en wanneer te antwoorden, en of de antwoorden anoniem worden verwerkt. Tevens is het raadzaam om waardering uit te spreken

voor de medewerking van de respondent, eventuele opdrachtgevers te noemen en om stimulansen te geven. In [2, p. 153] valt te lezen dat het geven van een stimulans nog voordat de vragenlijst is ingevuld een zeer positief effect heeft op het aantal reacties.

***Stap** 7: Test de vragenlijst uit in een pilot.*

Het uittesten van een vragenlijst is een uitermate belangrijke stap in de ontwikkeling van een vragenlijst, alhoewel verschillende onderzoekers er meestal op een verschillende manier naar kijken [2, p.140]. Het proces van uittesten is in [2] onderverdeeld in vier vervolgfases, te weten:

- *Revisie door kundige collega"s en analisten*

In deze fase wordt nagegaan of alle noodzakelijke vragen in de vragenlijst zijn inbegrepen, of bepaalde vragen verwijderd kunnen worden, en of de vragenlijst effectieve antwoordcategorieën bevat.
Afhankelijk van de grootte en het belang van het onderzoek kunnen mensen van verschillende expertises in deze fase betrokken worden. Het doel van deze fase van uittesten is om de wezenlijke inhoud van de vragenlijst definitief te maken. Nadat de vragenlijst hierop is aangepast kan de tweede fase van uittesten uitgevoerd worden.

- *Interviews om de cognitieve- en motivatiekwaliteiten te evalueren*

Deze fase is bedoeld om te onderzoeken of alle respondenten de woorden in de vragenlijst begrijpen, of alle respondenten de vragen op dezelfde manier interpreteren, of alle vragen een antwoordmogelijk hebben die de respondent kan kiezen, of het aannemelijk is dat iedere respondent alle vragen leest en beantwoordt, en ten slotte, of de verzending (envelop, begeleidende brief, en vragenlijst) een positieve indruk achterlaat bij de respondent.
In de laatste jaren is een techniek, bekend als *cognitief interviewen*, ontwikkeld, waarmee onderzocht kan worden of respondenten de vragen op dezelfde manier interpreteren als de onderzoeker in gedachte heeft en of de vragen accuraat beantwoord kunnen worden (Forsyth en Lessler, 1991). Deze techniek staat ook wel bekend als de *hardop denkmethode*. In deze methode wordt een aantal respondenten individueel gevraagd om de vragenlijst in te vullen en daarbij hardop na te denken. Door de respondent alles te laten vertellen wat ze denken, probeert de interviewer te begrijpen hoe iedere vraag wordt geïnterpreteerd en of de bedoeling van iedere vraag goed overkomt.
Naast de hardop denkmethode is de *retrospectieve techniek* bekend, waarin een aantal respondenten wordt geobserveerd bij het invullen van de vragenlijst. Hierbij wordt goed gelet op gezichtsuitdrukkingen, twijfelingen en andere houdingen als gevolg van de begrijpelijkheid van de vragenlijst. Na het invullen worden al deze problemen besproken met de betreffende respondent.
Beide technieken kunnen gezamenlijk toegepast worden om een vragenlijst te evalueren. Op de voor- en nadelen van de technieken zal verder niet ingegaan worden.

- *Kleine pilot studie*

Sommige vragen blijven nog onbeantwoord nadat de bovenstaande activiteiten, om de vragenlijst uit te testen, zijn uitgevoerd. Aan de hand van een kleine pilot studie wordt nagegaan of de antwoordcategorieën van schaalvragen (zie paragraaf 1.3.2) zo opgesteld zijn dat mensen henzelf verdelen over de categorieën in plaats van dat zij zich concentreren op alleen één of twee daarvan. Ook wordt gekeken of bepaalde stellingen (items), waarmee de onderzoeker een schaal wil vormen, correlatie vertonen en zodoende in dezelfde schaal

samengetrokken kunnen worden, of dat bij hoge correlatie juist één van de items weggelaten kan worden. Meer informatie over deze zogenaamde multiple item schalen is te lezen in hoofdstuk 2. In de pilot studie wordt ook onderzocht of bepaalde vragen in hoge mate onbeantwoord blijven.

- *Laatste check*

Deze laatste check is bedoeld om simpelweg te achterhalen of in de vragenlijst misschien iets stoms is gedaan. Dit wordt uitgevoerd door een kleine groep mensen, dat niets te maken heeft met de ontwikkeling of verbetering van de vragenlijst, deze volledig te laten invullen. Mensen die werken van de ene verbeterde versie naar de andere versie zien namelijk soms duidelijke fouten over het hoofd. Hierbij kan men denken aan taalfouten, antwoordcategorieën die dubbel voorkomen, et cetera.

***Stap 8**: Beoordeel de technische geschiktheid (adequacy) van de vragenlijst.*

Deze stap bestaat uit drie onderdelen waarin de *interne validiteit* (internal validity), *betrouwbaarheid* (reliability) en *externe validiteit* (external validity) van de vragenlijst wordt onderzocht.

De *interne validiteit* van de vragenlijst is de mate waarin een verkregen resultaat waar en onbevooroordeeld is. Interne validiteit wordt bereikt wanneer de resultaten van de vragenlijst accurate indicaties zijn van de standpunten of kennis van de respondent. Indien de resultaten, echter, tot stand zijn gekomen door de wijze waarop de vragenlijst is uitgevoerd, kan gesteld worden dat de resultaten intern invalide zijn. Een manier om interne validiteit te onderzoeken is door de antwoorden te vergelijken met die van gelijke respondenten (zoals iemands partner) [5].
De *externe validiteit* van de vragenlijst is de betrouwbaarheid waarmee de gevonden resultaten gelden voor andere populaties. Dit type validiteit zegt dus in hoeverre de resultaten van een vragenlijst generaliseerbaar zijn.
Kortom, de *validiteit* van een vragenlijst is de mate waarin de vragenlijst meet wat het behoort te meten. Onderzoek hiernaar is belangrijk omdat het de bruikbaarheid van de vragenlijst bepaalt. Dit geldt niet alleen voor opinieonderzoeken maar ook voor feitelijke onderzoeken (factual surveys), bijvoorbeeld als de respondenten de vragen verschillend interpreteren. Meer over validiteit is te lezen in paragraaf 3.4.

De *betrouwbaarheid* van de vragenlijst is de mate waarin de vragenlijst, iedere keer dat het wordt gebruikt, consistente informatie levert. Ofwel de kwaliteit van de vragenlijst om tot dezelfde resultaten te komen wanneer het ingevuld wordt door mensen met dezelfde denkbeelden in dezelfde omstandigheden. De betrouwbaarheid wordt meestal uitgedrukt in een getal tussen 0 (zeer onbetrouwbaar) en 1 (zeer betrouwbaar). Meer over de kwantificering van betrouwbaarheid komt aan bod in paragraaf 3.2. De betrouwbaarheid kan ook worden onderzocht door bijvoorbeeld feitelijke vragen te herhalen, respondenten opnieuw te interviewen of te controleren in hoeverre respondenten consistent zijn in hun antwoorden.

1.3 Vraagvormen

Vragen, die in een vragenlijst voor kunnen komen, zijn uiteraard op verschillende manieren in te delen. In deze paragraaf is dezelfde indeling als in [6] gehanteerd, waarbij vragen globaal zijn geclassificeerd in *gestructureerde* en *ongestructureerde* vragen. Een gestructureerde

vraag is meestal een keuzevraag waarbij de respondent kan kiezen uit een aantal van tevoren vastgestelde antwoordcategorieën. Een ongestructureerde vraag is meestal een open vraag waarbij de respondent de mogelijkheid krijgt om in eigen woorden zijn/haar antwoord op te schrijven. De mogelijke antwoorden op een ongestructureerde vraag liggen niet van tevoren vast.

Ondanks dat in ieder onderzoek wel ongestructureerde vragen voorkomen, zullen we ons in de rest van het werkstuk alleen maar concentreren op gestructureerde vragen, aangezien deze meer mogelijkheden voor statistische analyse geven. In de rest van deze paragraaf wordt ingegaan op een aantal soorten gestructureerde vragen.

1.3.1 Twee-keuze vragen

Voor dit type gestructureerde vragen kan de respondent uit twee antwoordmogelijkheden kiezen. Zoals “Ja” of “Nee”, “Mee eens” of “Mee oneens”, enzovoorts. Deze antwoordmogelijkheden zijn complementair aan elkaar. Twee-keuze vragen zijn over het algemeen makkelijk in te vullen door de respondent, wat resulteert in snellere beantwoording van de vragenlijst. Echter zijn deze vragen kwetsbaar voor meetfouten, aangezien de alternatieven zijn gepolariseerd. Dit wil zeggen dat de brede scala van mogelijke antwoordkeuzes tussen de twee polen is weggelaten. Het kiezen van de juiste formulering van de vraag is daarom in dit geval cruciaal om accurate antwoorden te verkrijgen. Een ander nadeel van twee-keuze vragen is dat de antwoorden meestal enige vorm van intensiteit of gevoel van de respondent missen. Er zijn namelijk gevallen waarin de ene respondent sterkere gevoelens heeft bij een bepaalde kwestie dan een andere respondent. De intensiteit gaat in dit geval verloren als de respondent maar uit twee antwoordmogelijkheden kan kiezen. De multiple choice vraag kan in dit geval uitkomst bieden.

1.3.2 Vragen gebaseerd op meetniveau

Vragen kunnen ook geclassificeerd worden op basis van hun meetniveau. Om bijvoorbeeld iemands beroep te meten kan een *nominale* vraag gehanteerd worden, waarbij de verschillende beroepen, waaruit gekozen kan worden, zijn genummerd. In dit geval heeft het nummer bij ieder antwoord geen betekenis behalve dat het de plek van de antwoordcategorie aangeeft.

Een ander type meetniveau dat gebruikt kan worden in vragenlijsten is het *ordinale* niveau. Een voorbeeld van een ordinale vraag is om de respondent een aantal automerken in een lijst te laten rangschikken in de volgorde van zijn/haar voorkeur. Als er vanuit wordt gegaan dat de lijst uit N automerken bestaat, is het de bedoeling dat de respondent aan het meest favoriete automerk het getalletje 1 toekent en aan het minst favoriete automerk het getalletje N.

Ook zijn er enquêtevragen waarvan de antwoordmogelijkheden op een *interval* meetniveau zijn gebaseerd. Eén van de meest populaire vraagsoorten van dit type is de *Likertschaal* vraag. Door middel van een meerkeuze antwoordmodel kan de respondent bijvoorbeeld de mate van instemming met een bepaalde uitspraak aangeven. Het antwoordmodel hanteert gewoonlijk als extremen de mate van instemming met, dan wel afwijzing van, de uitspraak. Een populaire variant van extremen is “volledig mee eens” en “volledig mee oneens”.

Bij deze schaal dient de kanttekening gemaakt te worden dat, technisch gezien, de antwoorden niet interval geschaald zijn, maar juist ordinaal. Binnen de statistische wereld is er daarom ook een discussie gaande over de toepasbaarheid van de Likertschalen. Hoofdstuk 2 is gewijd aan deze antwoordschaal.

Een ander type interval vraag is de *semantische differentiaal vraag*. Bij een dergelijk type vraag worden van een bepaald onderwerp, begrip of object twee tegenovergestelde polen als extremen gegeven. Deze extremen hebben een significant verschillende betekenis. De respondent wordt dan gevraagd aan te geven waar zijn/haar mening of voorkeur ligt ten aanzien van het onderwerp. Een voorbeeld van een semantische differentiaal vraag is te zien in figuur 1.1.

Probeer aan de hand van deze lijst jezelf te karakteriseren, door een kruisje te zetten bij de eigenschap die jou het best beschrijft:

	Helemaal	Een beetje	Neutraal	Een beetje	Helemaal	
Systematisch						Chaotisch
Actief						Passief
Amicaal						Formeel
Sociaal						Eenzelvig

Figuur 1.1: Semantische differentiaal vraag

Interval metingen kan men ook verkrijgen door een *cumulatieve* of *Guttmanschaal* te gebruiken. De antwoordschaal bestaat uit een aantal stellingen (items), waarbij de respondent dient te kiezen voor de stelling die zijn mening het beste weergeeft. Tussen de items van een Guttmanschaal komt een zeker hiërarchisch verband voor, dit wil zeggen dat wanneer de respondent het met een item eens is, dit dan waarschijnlijk ook geldt voor de items lager in de hiërarchie. Een voorbeeld van een Guttmanschaal, zoals die in [6] voorkomt, is als volgt:

Geef aan met welke uitdrukkingen u het eens bent:

__Bent u welwillend om immigranten toe te staan in uw land te wonen?
__Bent u welwillend om immigranten toe te staan binnen uw gemeenschap te wonen?
__Bent u welwillend om immigranten toe te staan in uw buurt te wonen?
__Bent u welwillend om een immigrant als buurman te hebben?
__Zou u uw kind laten trouwen met een immigrant?

Figuur 1.2: Guttmanschaal vraag

1.4 Schrijven van vragen

Uit de praktijk blijkt dat het formuleren van adequate vragen voor in een vragenlijst geen eenvoudige opgave is. Zoals in de introductie al ter sprake is gekomen, is het opstellen van een vragenlijst een kritieke succesfactor in een onderzoeksproces. Om te voorkomen dat in vragenlijsten bijvoorbeeld verkeerd woordgebruik gehanteerd wordt, een verkeerde structuur aangehouden wordt of vragen voorkomen die niet beantwoord kunnen worden, zijn er regels (onder andere in [1], [2] en [6]) die bij de constructie van de vragenlijst gevolgd kunnen worden. Met deze regels tracht de onderzoeker de juiste vragen te formuleren om tot valide onderzoeksresultaten te komen. In deze paragraaf zullen een aantal van deze regels besproken worden. Deze regels hebben met name betrekking op Likertschaal stellingen in zelfgeadministreerde vragenlijsten.

Een aantal regels met betrekking tot stellingen/vragen zijn als volgt:

- Een stelling dient één enkele kwestie te omvatten. Zorg ervoor dat er telkens één vraag gesteld wordt.

- Probeer een balans te vinden in de volgorde van de stellingen en het gebruik van positief en negatief geformuleerde stellingen.
- Houd rekening met de volgorde of „natuurlijke" ordening van stellingen, aangezien voorgaande stellingen de antwoorden op latere stellingen kunnen beïnvloeden.
- Vermijd negatieve woorden in stellingen. In combinatie met een aantal antwoordcategorieën kan dit namelijk leiden tot dubbele negatieven.
- Het taalgebruik in de stellingen dient simpel gehouden te worden, dat betekent geen technische, vage of dubbelzinnige woorden. Streef naar een taalgebruik van hetzelfde niveau als dat van de respondent.
- De schrijfstijl zal zo bondig en accuraat mogelijk moeten zijn.
- Vermijd hypothetische stellingen.
- De vragen en antwoordcategorieën dienen zo neutraal mogelijk geformuleerd te worden.

Wat betreft de antwoordcategorieën zijn de volgende regels belangrijk:

- Zorg ervoor dat de antwoordcategorieën elkaar uitsluiten en dus geen overlap vertonen.
- Zet de antwoorden in een logische volgorde. Dit stelt de respondent in staat om mogelijke antwoorden beter met elkaar te vergelijken.
- Zorg ervoor dat alle mogelijke antwoordcategorieën zijn inbegrepen. Met de antwoordcategorie „Anders" of „Specificeer" kan de respondent eventueel een eigen antwoord toevoegen.
- Wees voorzichtig met het gebruik van de „weet niet" antwoordoptie. Respondenten kunnen namelijk deze categorie zowel gebruiken als ze de vraag niet begrijpen, maar ook als ze geen mening over de stelling hebben.

Het volgen van deze regels kan de vragenlijstontwerper goed op weg helpen, maar alleen het volgen van deze regels is natuurlijk geen garantie voor succes. Zoals in het stappenplan van paragraaf 1.2 al naar voren is gekomen, zal de geconstrueerde vragenlijst onder andere uitgetest moeten worden in een pilot (stap 7) en ook zal de technische geschiktheid van de vragenlijst beoordeeld moeten worden (stap 8) om tot een adequaat meetinstrument te komen. Het beoordelen van de technische geschiktheid van een Likertschaal vragenlijst wordt in hoofdstuk 3 nader besproken.

1.5 Voor- en nadelen van vragenlijsten

Voor elke onderzoeksmethode zijn wel voor- en nadelen te benoemen, zo ook voor zelfgeadministreerde vragenlijsten ([2], [3], [5] en [8]). Hieronder zullen een aantal van deze voor- en nadelen worden benoemd. Aan het einde van deze paragraaf komen ook een aantal biases, die zich voor kunnen doen bij het invullen van vragenlijsten, aan bod.

De voordelen zijn deels aan bod gekomen, omdat zij normaalgesproken de drijfveer vormen om een vragenlijst op te stellen. Zo is een vragenlijst een waardevolle methode om van een grote groep respondenten informatie over allerlei zaken te verzamelen. Doordat de steekproef voldoende groot gekozen kan worden, is het mogelijk om met behulp van statistische methoden te onderzoeken of bepaalde metingen significant zijn. Ook kan met de verkregen dataset betrouwbare uitspraken worden gedaan over de doelgroep.
In tegenstelling tot andere onderzoeksmethoden kan met een zelfgeadministreerde vragenlijst uiteenlopende informatie worden ingewonnen. Zo kan de vragenlijst worden gebruikt om bijvoorbeeld overtuigingen, normen, waarden en gedrag te onderzoeken. Een ander voordeel van vragenlijsten is de mogelijkheid om de vragenlijst te standaardiseren. Dat betekent dat de vragen in een vragenlijst voor meerdere onderzoeken relevant kunnen zijn. Bovendien kan

een gestandaardiseerde vragenlijst na een bepaalde periode weer aan de respondenten worden voorgelegd, waarna de resultaten eenvoudig met elkaar kunnen worden vergeleken.
Andere pluspunten van vragenlijsten zijn dat ze relatief goedkoop zijn in vergelijking met andere onderzoeksmethoden, de respondent in staat wordt gesteld om over de antwoorden na te denken en dat ze eenvoudig en sneller in te vullen en te analyseren zijn. Overigens zorgt de anonimiteit van de respondent vaak voor eerlijkere antwoorden en hebben we bij zelfgeadministreerde vragenlijsten geen last van beïnvloeding door de interviewer (interviewer bias).

Vragenlijsten hebben ook hun beperkingen als onderzoeksinstrument. Zo zijn de resultaten afhankelijk van de eerlijkheid van de respondent. Respondenten geven namelijk meestal een sociaal en cultureel geaccepteerd antwoord op gevoelige items of een gunstig antwoord op persoonlijke vragen. Aangezien vragen vaak ook betrekking hebben op het verleden, zijn de resultaten tevens afhankelijk van het geheugen van de respondent. Bovendien heeft de respondent, bij zelfgeadministreerde vragenlijsten, de mogelijkheid om van te voren alle vragen te bekijken. Dit kan invloed hebben op zijn antwoordgedrag.
Een ander nadeel van vragenlijsten is dat de verkregen data afhankelijk is van de mate waarin respondenten de items begrijpen. Vandaar dat men de vragenlijst meestal uittest in een pilot en dat eenvoudige taalgebruik in stellingen wordt aangeraden (zie paragraaf 1.4). Een ander probleem is de bereidheid van respondenten om de vragenlijst in te vullen. Een lage bereidheid zorgt voor een lage response percentage, wat natuurlijk de bruikbaarheid van het onderzoek verkleint. Wat de minimale response percentage moet zijn, hangt af van het doel en aard van de studie.

Vragenlijsten hebben normaalgesproken ook met biases te maken. In een vroeg stadium van de vragenlijstontwikkeling zal men met deze biases rekening moeten houden. Als de biases erg groot zijn kan de geldigheid van de vragenlijst in twijfel worden getrokken en om die reden probeert men meestal ook methoden te vinden om de impact ervan te minimaliseren. Een aantal biases die zich bij vragenlijstonderzoek voor kunnen doen zijn de volgende:

Hawthorne effect
Het Hawthorne effect is een bias die zich vaak voordoet bij vragenlijstonderzoek. Deze bias houdt in dat respondenten anders gaan antwoorden vanwege het feit dat ze geselecteerd zijn voor het onderzoek. Door deze speciale aandacht neigen de respondenten om antwoorden te geven, die volgens hen, de onderzoeker het meeste zal bevallen. Om deze bias te minimaliseren dient de vragenlijstontwerper de vragen zo neutraal mogelijk te presenteren.
Het effect schijnt ook voor te komen bij productieprocessen, die door aandacht van het management, beter gaan lopen. Toch zijn er ook mensen die hun twijfels hebben over dit effect.

De 'Gewoonte' bias (The 'Habit' bias)
Wanneer de respondent een reeks van ongeveer dezelfde vragen voorgeschoteld krijgt, zal hij/zij in de gewoonte vallen om ze op dezelfde manier te beantwoorden, zonder bij iedere vraag na te denken over de betekenis ervan. Deze bias kan geminimaliseerd worden door verschillende vraagtypen in de vragenlijst op te nemen.

Niet-respondent bias (Non-respondent bias)
Dit type bias komt vaak terug bij vragenlijstonderzoek en is gebaseerd op de aanname dat individuen die niet meegedaan hebben aan het onderzoek dezelfde gevoelens en meningen

hebben als degenen die de vragenlijst wel hebben ingevuld. Uit studies blijkt echter dat niet-respondenten over het algemeen een negatievere opvatting hebben.

Het Halo effect (The Halo effect)
Het Halo effect is een antwoordbias die voor kan komen bij antwoordschalen. Respondenten geven bijvoorbeeld mensen die zij waarderen of respecteren een hoge score op alle items. Deze hoge beoordeling is ongeacht de daadwerkelijke prestatie van de persoon. De aanwezigheid van een bepaalde kwaliteit bij een persoon geeft de respondent dus de suggestie dat deze persoon ook andere kwaliteiten heeft. Deze bias komt voornamelijk voor wanneer mensen worden beoordeeld.

Het Hooivork effect (The Pitchfork effect)
Het Hooivork effect is het tegenovergestelde van het Halo effect. Deze antwoordbias kan zich voordoen wanneer respondenten bijvoorbeeld producten moeten beoordelen. In dit geval zullen zij meestal de subtiele kenmerken van het product oppikken en negatiever antwoorden dan normaal. Om het hooivork effect te voorkomen zal de vragenlijst niet moeten sturen naar negatieve antwoorden.

Hoofdstuk 2

Likertschaal vragen

2.1 Introductie

De Likertschaal is één van de meest gebruikte antwoordschalen in vragenlijsten. Met name in de psychologie en sociale wetenschappen is het gebruik van deze schaal enorm populair. In de rest van dit werkstuk zal ook de nadruk gelegd worden op vragenlijsten met een Likert antwoordschaal. De schaal is vernoemd naar de Amerikaanse onderwijzer en organisatiepsycholoog Rensis Likert (1903-1981), die het gebruik van de schaal beschreef in zijn proefschrift[1].


R. Likert

De schaal, die Likert ontwierp, was bedoeld om standpunten te meten en hij liet tevens zien dat deze schaal meer informatie opving dan de bestaande methoden. De Likertschaal is een antwoordschaal, die respondenten kunnen gebruiken, om de mate van instemming met een bepaalde stelling te specificeren. De extremen van het antwoordmodel geven de mate van instemming met, dan wel afwijzing van, de uitspraak aan. Een voorbeeld van een stelling met een Likert antwoordschaal is in onderstaand figuur te zien:

		Volledig mee oneens	Oneens	Neutraal	Eens	Volledig mee eens
1	Ik voel me thuis op het werk	1	2	3	4	5

Figuur 2.1: Likertschaal vraag

Een stelling wordt in dit verband ook wel een item genoemd. De antwoordextremen in het voorbeeld zijn “volledig mee eens” en “volledig mee oneens”. Daarnaast kunnen ook andere extremen tot de mogelijkheden behoren. Tevens is in het voorbeeld gebruik gemaakt van een 5-punts Likertschaal. Dit is een willekeurige keuze geweest. Afhankelijk van de gewenste precisie van de onderzoeker kunnen ook andere schaallengtes gehanteerd worden. De vragenlijstontwerper dient de schaallengte zo te kiezen dat de respondenten onderscheid kunnen blijven maken tussen de verschillende antwoordcategorieën. Bij een oneven aantal schalen, zoals in bovenstaand voorbeeld, heeft de respondent de keuze uit een middelste waarde “neutraal” of “noch mee eens, noch mee oneens”. Om te voorkomen dat veel respondenten voor deze antwoordcategorie kiezen, kan een gedwongen keuze antwoordschaal gebruikt worden. Deze schaal heeft een even aantal antwoordcategorieën, waarbij de neutrale positie niet meer kan worden ingenomen. De respondent wordt zodoende bij ieder item gedwongen te beslissen of hij/zij meer neigt naar de “mee eens” of “mee oneens” einde van de schaal.

[1] Likert, Rensis (1932), "A Technique for the Measurement of Attitudes", Archives of Psychology 140: p 1-55

In paragraaf 2.2 komt ter sprake hoe Likertschaal vragenlijsten gebruikt kunnen worden om constructen meetbaar te maken. Paragraaf 2.3 staat in het teken van het meetniveau van Likertschalen. Er bestaat namelijk veel discussie over welk meetniveau geldt voor Likertschalen. Het meetniveau van een schaal bepaalt onder andere welke methoden van data-analyse toegepast kunnen worden op de resultaten. Ten slotte wordt in paragraaf 2.4 een aantal voor- en nadelen van Likertschalen belicht.

2.2 Meten van constructen

Likertschalen zijn een uiterst handig hulpmiddel om constructen, die men niet zomaar rechtstreeks kan observeren of meten, toch te kunnen kwantificeren. Hierbij kan men denken aan constructen als vertrouwen, tevredenheid, attitude ten aanzien van politici, et cetera. Het construct, ook wel latente variabele genoemd, wordt dan bijvoorbeeld gemeten aan de hand van antwoorden op stellingen. Het doel van deze items is om verschillende aspecten van het te meten verschijnsel weer te geven. Uiteindelijk worden de antwoorden op de verscheidene items gecombineerd tot een enkele score voor het betreffende construct. Gezamenlijk vormen deze items een (meet)schaal. De eindscore van een respondent op een schaal is meestal de som van de scores van de individuele items. De Likertschaal wordt daarom ook wel de *gesommeerde ratingschaal* (summated rating scale) genoemd. Sommige items in een schaal kunnen in tegenovergestelde richting gesteld zijn dan andere items in de schaal. De scores van deze „tegenovergestelde" items zullen dan omgewisseld moeten worden, voordat de scores worden opgeteld. Voor een 5-punts Likertschaal met antwoordcategorieën 1 tot en met 5 zal een dergelijk item met score 1, een score 5 krijgen, enzovoorts.

In het onderzoeksproces zal de onderzoeker de relevante stellingen moeten selecteren, die samen de schaal gaan vormen, die het beoogde construct meet. Deze activiteit heeft invloed op de validiteit en betrouwbaarheid van de meting. In [7] wordt dit proces van het selecteren van de relevante stellingen met „domein sampling" vergeleken. Het domein is in dit geval de theoretische populatie van alle mogelijke items, die relevant zijn voor het meten van een bepaald construct. Zo geeft de verzameling van alle antwoorden op alle items in het domein de „werkelijke" attitude van de respondent ten opzichte van het construct. Natuurlijk kunnen niet alle items worden verzameld, maar op deze manier kan meten beschouwd worden als het trekken van een steekproef van items uit het domein. De antwoorden van de items in de steekproef kunnen dan geïnterpreteerd worden als een schatting van de attitude van de respondent. De validiteit van de meting met behulp van de geselecteerde Likert items is dan analoog aan de representativiteit van de steekproef en de betrouwbaarheid van de meting komt dan overeen met de verwachte fout van de schatting ten opzichte van de echte waarde.
Aangezien het in de praktijk voor een onderzoeker onmogelijk is om het domein van items volledig te bedenken, wordt in plaats daarvan een beperkt aantal items geformuleerd, die het construct zoveel mogelijk meten. De gemaakte analogie heeft echter wel belangrijke implicaties, namelijk dat schalen met meer items tot op zekere hoogte betrouwbaarder zijn, want grotere steekproeven hebben een kleinere standaardfout. Ook de keuze van welke items in een schaal te gebruiken is belangrijk, evenals dat dat het geval is voor de samenstelling van een steekproef. Zo resulteren meer soortgelijke items in een hogere betrouwbaarheid, maar de keerzijde is dat deze items dan mogelijk slechts een beperkte inhoud van het construct vatten.

Tot slot gaan we in op de dimensionaliteit van constructen. De dimensionaliteit van een construct wordt bepaald door de relevante onderliggende deelaspecten of dimensies. Een construct wordt ééndimensionaal genoemd wanneer het construct een representatie is van een

enkel algemeen kenmerk. De variabelen die gebruikt worden om een ééndimensionaal construct te meten, oftewel de verzameling items, dienen een sterke samenhang te vertonen. Indien tussen bepaalde variabelen, die een construct meten, nauwelijks verbanden worden gevonden is het mogelijk dat we te maken hebben met een meerdimensionaal construct. Een voorbeeld van een meerdimensionaal construct is de attitude tegenover een regering. Deze attitude is namelijk op te splitsen in attituden op verschillende dimensies, zoals belastingen, openbare werken, economisch beleid, sociaal beleid, et cetera. De attitude op deze deelaspecten kunnen beschouwd worden als ééndimensionale constructen.
Om te onderzoeken of een ééndimensionaal construct slechts een enkel kenmerk vertegenwoordigt, kan de interne consistentie van de multiple item schaal, nadat deze is ingevuld door een pilotgroep, worden onderzocht. Meer hierover komt aan bod in paragraaf 3.2.

2.3 Meetniveau van Likertschalen

De antwoorden op een enkele Likert item worden normaalgesproken als ordinale data behandeld. De aanname is hierbij dat respondenten de afstanden tussen twee opeenvolgende antwoordcategorieën niet als gelijk ervaren. Binnen de statistische wereld zijn er daarom ook mensen die kanttekeningen plaatsen bij het gebruik van bepaalde methoden voor data-analyse op Likertschaal data. Anderen stellen weer dat het meetniveau van een schaal een empirisch probleem is dat, net zoals de betrouwbaarheid en validiteit, onderzocht dient te worden. Men kan dit onderzoeken door de ontwikkelde schaal te vergelijken met andere schalen, die bijvoorbeeld wel een interval meetniveau hebben. Indien daarbij lineaire verbanden worden gevonden, kan de schaal als intervalschaal worden toegepast.
De statistici die vinden dat het meetniveau van een meetschaal empirisch vastgesteld moet worden, gaan er dus vanuit dat dit meetniveau niet door de schaalconstructie bepaald wordt. In [7] valt te lezen dat empirisch onderzoek van onder andere Dawes (1977) aantoont dat zelfs eenvoudige beoordelingsschalen soms data opleveren, die een lineair verband lijken te hebben met bepaalde fysische metingen, zoals lengte. In het onderzoek komt ook ter sprake dat multiple item schalen nog betere verbanden geven. De resultaten van deze onderzoekers geven volgens [7] een empirische en logische rechtvaardiging om statistische methoden voor data op intervalniveau toe te passen op Likertschaal data.

2.4 Voor- en nadelen van Likertschalen

De Likertschaal is, zoals eerder al vermeld, één van de meest gebruikte antwoordschalen in meetinstrumenten. Daar zijn verschillende redenen voor op te noemen, zo staat in [8, p.2] vermeld dat een goed ontwikkelde Likertschaal instrument een goede validiteit en betrouwbaarheid met zich mee kan brengen, dat het een relatief goedkoop en makkelijk te ontwikkelen instrument is en dat de schaal voor respondenten eenvoudig en snel in te vullen is. Het gebruik van Likertschalen heeft echter ook nadelen, namelijk dat deze schalen geen informatie geven over de mate van betrokkenheid van de respondent bij het construct. Een ander nadeel is ook dat gelijke eindscores op een Likertschaal via verschillende wegen kunnen zijn bereikt. Zo kan de ene respondent een gemiddelde score behalen door op alle items gemiddeld te scoren, terwijl de andere respondent een gemiddelde score behaalt door op ene helft van de items hoog te scoren en op de andere helft laag te scoren of omgekeerd.
Likertschalen kunnen door verschillende oorzaken ook een vertekend beeld geven. Zo hebben respondenten vaak de neiging om extreme antwoordcategorieën te vermijden, wat ook wel

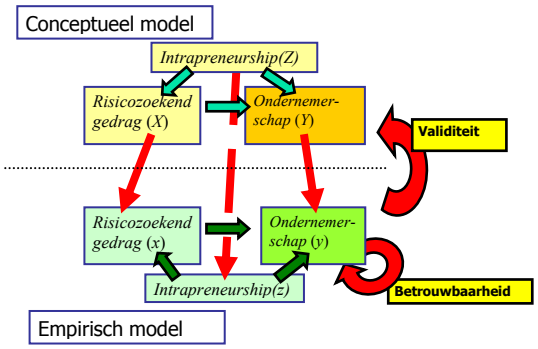
centrale tendens bias wordt genoemd. Tevens zijn respondenten het vaak eens met een stelling door de wijze waarop het geformuleerd is (*instemming bias*). Bovendien proberen respondenten zichzelf of hun organisatie vaak positiever af te schilderen dan werkelijk het geval is (*sociale wenselijkheid bias*). Naast deze biases spelen natuurlijk ook de biases van paragraaf 1.5 een rol bij Likertschalen.

Hoofdstuk 3

Analyse van Likertschalen

3.1 Introductie

De constructie van een Likertschaal vragenlijst verloopt zoals de constructie van andere vragenlijsten in een aantal stappen, zie ook paragraaf 1.2. Nadat een groot aantal items, dat relevant is voor het onderwerp van het construct, is verzameld, kan de kwaliteit van de schaal beoordeeld worden. Het beoordelen van de kwaliteit van de schaal kan helpen om de schaal verder te verbeteren door bijvoorbeeld bestaande items te verwijderen of door juist nieuwe items toe te voegen. Twee bekende aspecten van de kwaliteit van een schaal zijn *validiteit* en *betrouwbaarheid*. Deze twee criteria zijn in stap 8 van paragraaf 1.2 al kort aan bod gekomen en onderstaande figuur biedt een verdere verduidelijking:



Figuur 3.1: Conceptuele en empirische relatieschema van het begrip ondernemerschap (bron [9])

In figuur 3.1 is het conceptuele en empirische relatieschema van het begrip ondernemerschap weergegeven. Het conceptueel model (boven de stippellijn) brengt in kaart welke factoren verondersteld worden van invloed te zijn op iemands ondernemerschap. Het conceptueel model volgt meestal na een (theoretisch) vooronderzoek. Het empirische model (onder de stippellijn) beschrijft wat zich in de werkelijkheid (empirie) afspeelt. De betrouwbaarheid is de mate waarin een schaal (bijvoorbeeld om ondernemerschap te meten) dezelfde resultaten geeft bij herhaalde toepassing onder dezelfde voorwaarden. De betrouwbaarheid heeft alleen betrekking op het empirische model. Validiteit daarentegen, zegt iets over de relatie tussen het empirische en het conceptuele model. De validiteit is de mate waarin een schaal meet wat het zou moeten meten. Een belangrijke opmerking hierbij is dat de betrouwbaarheid en validiteit van een vragenlijst niet bewezen kunnen worden. In de volgende paragrafen zullen slechts methoden besproken worden, waarmee men validiteit en betrouwbaarheid kan proberen aan te tonen. Bovendien geldt dat wanneer betrouwbaarheid en validiteit is vastgesteld, we eigenlijk alleen de conclusies, die via het meetinstrument verkregen worden, valide en betrouwbaar kunnen noemen [6]. Dat betekent dus dat validiteit en betrouwbaarheid geen betrekking hebben op het meetinstrument zelf. In het vervolg zullen we gemakshalve toch spreken over de validiteit en betrouwbaarheid van een vragenlijst.

Paragraaf 3.2 staat in het teken van een aantal technieken om de betrouwbaarheid van een Likertschaal vragenlijst te schatten. Daarnaast wordt in dit hoofdstuk aandacht besteed aan een methode om de kwaliteit van een schaal als geheel te verbeteren, ook wel item-analyse genoemd (paragraaf 3.3). Paragraaf 3.4 gaat kort in op het begrip validiteit.

3.2 Betrouwbaarheid

Om de betrouwbaarheid van een Likertschaal vragenlijst te schatten kunnen verschillende methoden gehanteerd worden ([6], [9] en [10]). Dezelfde vragenlijst kan bijvoorbeeld op twee verschillende tijdstippen worden toegepast. De betrouwbaarheid van de vragenlijst kan dan worden gemeten door de samenhang tussen beide metingen te bepalen. Dit type betrouwbaarheid staat ook bekend als *„test-retest reliability"*. De samenhang tussen de metingen is te kwantificeren met behulp van de correlatiecoëfficiënt tussen de eindscores van de respondenten uit de twee metingen. Een nadeel van deze aanpak is dat de tijd tussen beide metingen een cruciale rol speelt. Wanneer hetzelfde wordt gemeten, heeft de tijd tussen de metingen immers invloed op de correlatie tussen de twee observaties.
Een andere manier om de betrouwbaarheid van een Likertschaal vragenlijst af te leiden is door twee verschillende, maar vergelijkbare, schalen onder de doelgroep uit te zetten. Deze schalen kunnen bijvoorbeeld verkregen worden door een grote verzameling items, die op hetzelfde construct betrekking hebben, in twee groepen te verdelen. De betrouwbaarheid kan dan geschat worden door de correlatie tussen beide parallelle metingen te bepalen. Een nadeel van deze zogenaamde *„parallel-forms reliability"* is dat het opstellen van een grote groep items geen eenvoudig karwei is.

In onderstaande figuur zijn test-retest reliability (links) en parallel-forms reliability (rechts) geïllustreerd.

Figuur 3.2: Test-retest reliability en parallel-forms reliability (bron [6])

In de praktijk is gebleken dat het erg moeilijk en tijdrovend is om een tweede vragenlijst te construeren, waarmee de originele vragenlijst op betrouwbaarheid kan worden getoetst. Dat heeft er volgens Cronbach [10] toe geleid dat de zogenaamde *halveringsmethode* (split-half method) is bedacht. Volgens deze aanpak is de betrouwbaarheid van een Likertschaal vragenlijst ook te schatten door alle items, nadat deze op één moment door de doelgroep is ingevuld, in twee willekeurige groepen te verdelen. De betrouwbaarheid wordt dan geschat op basis van de correlatie tussen de somscores op de twee halve schalen. In paragraaf 3.2.1 zal nader beschreven worden hoe de halveringsmethode op een Likertschaal vragenlijst kan worden toegepast.

Veelvuldig wordt een generalisatie van de hiervoor beschreven halveringsmethode toegepast om de betrouwbaarheid van een Likertschaal vragenlijst te schatten. Deze schatting van de betrouwbaarheid staat bekend onder de naam *Cronbach's alpha*. Paragraaf 3.2.2 gaat dieper in op de definitie van Cronbach's α en hoe deze coëfficiënt voor een Likertschaal vragenlijst kan worden berekend. In paragraaf 3.2.3 zullen we deze populaire coëfficiënt verder bestuderen en er een betrouwbaarheidsinterval voor afleiden.

3.2.1 Halveringsmethode

Om de betrouwbaarheid van een gesommeerde ratingschaal, zoals de Likertschaal, te schatten, kan de split-half betrouwbaarheid worden berekend. De oorspronkelijke vragenlijst wordt in twee vragenlijsten gesplitst, waarbij wordt gestreefd naar twee inhoudelijk homogene vragenlijsten. Door een eerlijke verdeling van de items qua inhoud en moeilijkheid probeert men dit te bewerkstelligen. Meestal wordt de vragenlijst simpelweg gesplitst door de vragen met een even nummer te scheiden van de vragen met een oneven nummer [13]. Wanneer de vragenlijst is ingevuld door de doelgroep, kan men de correlatie tussen de somscores op de twee halve schalen als volgt berekenen:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\,s_x s_y}. \qquad (3.1)$$

Waarbij:

n	Aantal respondenten, die de vragenlijst hebben ingevuld.
x_i	Somscore van respondent i op de eerste halve schaal.
$\bar{x}$	Gemiddelde van de somscores op de eerste halve schaal.
y_i	Somscore van respondent i op de tweede halve schaal.
$\bar{y}$	Gemiddelde van de somscores op de tweede halve schaal.
s_x	Standaarddeviatie somscores op de eerste halve schaal.
s_y	Standaarddeviatie somscores op de tweede halve schaal.

Wanneer de oorspronkelijke vragenlijst perfect betrouwbaar is, wordt verwacht dat de twee halve schalen perfect gecorreleerd zijn ($r_{xy} = 1$). Alle respondenten zouden namelijk in dat geval precies dezelfde somscores behalen op de twee halve schalen. Dat betekent dus dat hoe sterker de correlatie tussen de somscores op de twee halve schalen, hoe groter de betrouwbaarheid van de oorspronkelijke vragenlijst is.

De halveringsmethode geeft een schatting van de betrouwbaarheid van elk van de helften, maar niet van de samengevoegde vragenlijst. Om te corrigeren voor het aantal items in de twee halve schalen kan men in plaats van de correlatiecoëfficiënt ook gebruik maken van de Spearman-Brown formule. Met deze formule kan een schatting van de betrouwbaarheid van de somscore gebaseerd op alle items verkregen worden. De Spearman-Brown formule is als volgt:

$$\rho = \frac{2r_{xy}}{1 + r_{xy}}. \qquad (3.2)$$

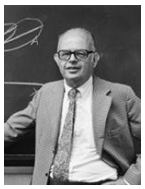
Waarbij ρ staat voor de *Spearman-Brown split half betrouwbaarheidscoëfficiënt* en r_{xy} voor de correlatiecoëfficiënt tussen de somscores van de twee halve schalen (zie formule (3.1)). In formule (3.2) zien we dat het corrigeren voor het aantal items in de halve schalen gebeurd door de correlatiecoëfficiënt te verdubbelen en vervolgens te normaliseren.

In [10] staat te lezen dat de *split-half Spearman-Brown* procedure veertig jaar lang gold als standaard methode om vragenlijsten te analyseren. Desondanks is deze aanpak ook regelmatig bekritiseerd. De halveringsmethode is onder andere door Brownell [19] en Kuder en

Richardson [20] bekritiseerd om het feit dat deze aanpak geen unieke waarde voor de betrouwbaarheid oplevert. De betrouwbaarheidscoëfficiënt van een vragenlijst is nu namelijk afhankelijk van welke items er gegroepeerd worden bij het splitsen van de items in twee groepen. Omdat verschillende splitsingen leiden tot verschillende uitkomsten voor de betrouwbaarheids-coëfficiënt, kunnen we niet met zekerheid uitspraken doen over de betrouwbaarheid van de vragenlijst. In de volgende paragraaf wordt een andere betrouwbaarheidscoëfficiënt besproken, die dit probleem niet heeft.

3.2.2 Cronbach's alpha

De items, die gebruikt worden om een schaal te vormen, dienen intern consistent te zijn. Grofweg betekent dit dat alle items hetzelfde concept behoren te meten. Om een indicatie te krijgen van de mate waarin de items in een vragenlijst hetzelfde concept meten wordt veelvuldig Cronbach"s α berekend. Deze maat is door de Amerikaanse psycholoog Lee J. Cronbach (1916 – 2001) geïntroduceerd in [10].



L.J. Cronbach

Cronbach"s α wordt berekend nadat de vragenlijst door een grote groep respondenten is ingevuld en geeft aan in hoeverre de antwoorden van deze groep respondenten op de items consistent zijn. De waarden, die Cronbach"s α kan aannemen, variëren van -∞ tot 1. In een later stadium zal blijken dat alleen positieve waarden voor deze coëfficiënt zinvol zijn. Hoe dichter Cronbach"s α bij 1 zit, des te hoger is de interne consistentie en dus de betrouwbaarheid van de schaal. Logischerwijs wordt een lage waarde van deze coëfficiënt beschouwd als ongunstige kwalificatie van de schaal.

De formule van Cronbach"s α is als volgt:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} s_{Y_i}^2}{s_X^2}\right). \qquad (3.3)$$

Waarbij:

K	Aantal items in de vragenlijst.
Y_i	Score op item i van één respondent, $i=1,\ldots,$K.
$s_{Y_i}^2$	Steekproefvariantie van de scores op item i over alle respondenten.
s_X^2	Steekproefvariantie van de somscores over alle respondenten.

De somscore X van een bepaalde respondent wordt verkregen door de som van de scores op alle items te nemen, oftewel $X = \sum_{i=1}^{K} Y_i$.

Om de uitdrukking van Cronbach"s α af te leiden, hebben we resultaten uit de *klassieke testtheorie* nodig. Klassieke testtheorie gaat uit van een zeer algemene relatie tussen de waargenomen en theoretische uitkomst van een schaal. Volgens de gedachtegang in deze theorie is de waarnemingsscore op te splitsen in een ware score (true score, T_i) en een

meetfout (error score, ε_i). Voor de waarnemingsscore (Y_i) van een willekeurige respondent op item i van een vragenlijst betekent dit:

$$Y_i = T_i + \varepsilon_i, \qquad voor\ i = 1,....,K. \tag{3.4}$$

De ware score is de waarde die een verschijnsel volgens de theorie heeft. De meetfout kan gezien worden als een willekeurige fout, die inherent is aan het meetproces. Factoren die invloed hebben op de meetfout zijn bijvoorbeeld de stemming van een respondent bij het invullen van de vragenlijst. Ook kan weinig motivatie of een lage intelligentie leiden tot meetfouten. Slecht geformuleerde vragen spelen tevens een rol. Verder kunnen fouten van de onderzoeker, zoals telfouten, gezien worden als meetfouten.

De meetfouten worden verondersteld onderling onafhankelijk verdeeld te zijn en ook onafhankelijk te zijn van de ware score. De verwachting van de meetfouten is bovendien gelijk aan 0. In wiskundige notatie zijn deze eigenschappen als volgt op te schrijven:

$$\begin{cases} E(\varepsilon_i) = 0, & voor\ i = 1,....,K, \\ Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j, \\ 0, & i \neq j, \end{cases} \\ Cov(T_i, \varepsilon_i) = 0, & voor\ i = 1,...,K. \end{cases} \tag{3.5}$$

Hoe uiteindelijk met de resultaten uit de klassieke testtheorie de formule van Cronbach‟s α is af te leiden en welke aannamen daarbij zijn gedaan, komt in Appendix A uitvoerig aan bod.

In de context van de klassieke testtheorie is een definitie voor de betrouwbaarheid eenvoudig te achterhalen. Een meting is namelijk betrouwbaar wanneer het voor het merendeel de ware score, in vergelijking met de meetfout, representeert. Een maat of statistiek voor de betrouwbaarheid kan zodoende verkregen worden door de mate van variabiliteit van de ware score onder de respondenten, relatief aan de waarnemingsscore variabiliteit te nemen:

$$betrouwbaarheid = \frac{Var(\text{ware score})}{Var(\text{waarnemingsscore})}. \tag{3.6}$$

Voor een schaal betekent dit dat de betrouwbaarheid gedefinieerd kan worden als de ratio van de „ware somscore‟ variantie en de variantie van de waargenomen somscores van de respondenten.
Uit vergelijking (3.6) blijkt dat een hoge betrouwbaarheid alleen bereikt kan worden wanneer de variantie van de waarnemingscores van respondenten niet te veel verschilt van de variantie van hun ware scores. Bij de afleiding van de formule van Cronbach‟s α wordt dit quotiënt ook verkregen (zie vergelijking (A.3) in Appendix A).

Vragenlijsten bestaan normaalgesproken uit meerdere items, omdat toevalsinvloeden, die bij de beantwoording van afzonderlijke items invloed hebben, tegen elkaar weg kunnen vallen. Als we teruggaan naar de formule van Cronbach‟s α zien we dat deze inderdaad ook afhangt van het aantal gebruikte items (K). Uit de formule is niet meteen duidelijk wat de invloed is van het verhogen van K op de betrouwbaarheidscoëfficiënt. Uit [17] volgt dat het toevoegen van items niet altijd resulteert in een grotere waarde van Cronbach‟s α. Het is namelijk afhankelijk van de lengte van de originele schaal en de correlaties tussen de items of het

toevoegen van een item een positief effect heeft op deze betrouwbaarheidscoëfficiënt. Des te hoger K en/of de itemcorrelaties zijn des te hoger de kwaliteit van het toegevoegde item moet zijn om Cronbach's α te laten toenemen. De kwaliteit van een item kan hierbij gezien worden als de correlatie van het item met de originele items.

Wat een goede waarde is voor Cronbach's α, verschilt per onderzoeksdiscipline. Zo valt in [12] te lezen dat een goede test van bijvoorbeeld iemands intelligentie een betrouwbaarheid van boven 0.90 heeft en dat een redelijk goede schaal voor de meting van bijvoorbeeld een attitude (zoals tevredenheid, risicozoekend gedrag, et cetera) een betrouwbaarheid van tenminste 0.70 heeft. Een Cronbach's α van tenminste 0.70 betekent dat de verzameling items meer gemeenschappelijks meten dan dat zij verschillende dingen meten. Wanneer de betrouwbaarheidscoëfficiënt een waarde van beneden 0.60 heeft, zal men na moeten gaan of de verzameling items wel een schaal vormt.

Wanneer de items in een vragenlijst geen ware score meten en we dus eigenlijk alleen maar meetfouten meten, zal de variantie van de som van de itemscores (s_X^2) gelijk zijn aan de som van de varianties van de individuele itemscores ($\sum_{i=1}^{K} s_{Y_i}^2$). We hebben hier namelijk te maken met meetfouten, die ongecorreleerd zijn voor alle respondenten. Dat betekent dat Cronbach's α gelijk is aan 0 als de items geen ware score meten. Als de waargenomen itemscores Y_i positief samenhangen, dan is de schaalvariantie s_X^2 groter dan de som van de itemvarianties en ligt de betrouwbaarheidscoëfficiënt dicht bij 1. Ten slotte is Cronbach's α gelijk aan 1 als alle items volkomen betrouwbaar zijn en allen hetzelfde concept (ware score) meten. In [11] wordt dit laatste gecontroleerd door bijvoorbeeld aan te nemen dat de K items in een schaal identiek zijn en om die reden perfect gecorreleerd. Aangezien alle $s_{Y_i}^2$ dan gelijk zijn aan elkaar en $s_X^2 = Var(K \cdot Y_i) = K^2 Var(Y_i) = K^2 s_{Y_i}^2$, is Cronbach's α gelijk aan:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} s_{Y_i}^2}{s_X^2}\right) = \frac{K}{K-1}\left(1 - \frac{K \cdot s_{Y_i}^2}{K^2 \cdot s_{Y_i}^2}\right) = \frac{K}{K-1}\left(1 - \frac{1}{K}\right) = 1.$$

Een negatieve Cronbach's α wordt verkregen wanneer de waargenomen itemscores Y_i negatief samenhangen. Dit gebeurt bijvoorbeeld als de scores op negatief gestelde items niet zijn gehercodeerd, zoals in paragraaf 2.2 ter sprake is gekomen. De schaalvariantie kan dan namelijk iets kleiner zijn dan de som van de itemvarianties, waardoor de betrouwbaarheids-coëfficiënt kleiner is dan 0. Normaalgesproken is een negatieve waarde voor Cronbach's α dus niet mogelijk, tenzij de onderzoeker een fout heeft gemaakt.

In de volgende paragraaf komen verschillende interpretaties van Cronbach's α aan bod en wordt ook gekeken naar de misinterpretaties van deze coëfficiënt. Aangezien Cronbach's α slechts een puntschatter is, zal ook beschreven worden hoe een betrouwbaarheidsinterval voor deze schatter kan worden berekend.

3.2.3 Cronbach's alpha nader bestudeerd

Een directe interpretatie van Cronbach's α, zoals uit [10] en [11] volgt, heeft te maken met de veronderstelling dat de items in een vragenlijst eigenlijk alleen maar een verzameling zijn uit

een grote groep mogelijke items. Wanneer we twee willekeurige verzamelingen bestaande uit K items uit deze mogelijke items kunnen vormen, zal de verwachte correlatie tussen de somscores van deze schalen gelijk zijn aan Cronbach‟s α. Een andere interpretatie, die gerelateerd is aan bovengenoemde interpretatie, is dat de waarde van Cronbach‟s α het gemiddelde is van alle split half coëfficiënten, die resulteren uit verschillende splitsingen van een vragenlijst [10]. Dit is de reden waarom de Cronbach‟s α methode gezien wordt als generalisatie van de halveringsmethode.

De maat Cronbach‟s α wordt, onder andere in de psychologie, zeer vaak toegepast om iets te kunnen zeggen over de betrouwbaarheid van een vragenlijst. Ondanks de populariteit van deze betrouwbaarheidscoëfficiënt zijn er ook kanttekeningen te plaatsen bij het gebruik ervan. Zo wordt in [14] vermeldt dat een hoge waarde voor Cronbach‟s α niet meteen betekent dat binnen de schaal niet meerdere subschalen aanwezig kunnen zijn. Cronbach‟s α wordt namelijk vaak onterecht gerapporteerd als maat voor de homogeniteit of ééndimensionaliteit van een vragenlijst. De interne consistentie van de items in een schaal, waar Cronbach‟s α betrekking op heeft, is natuurlijk wel een voorwaarde voor homogeniteit. Uit [15] blijkt tevens dat Cronbach‟s α geen robuuste schatter is voor de betrouwbaarheid. We zullen dit nader uitleggen door de formule van Cronbach‟s α als volgt te herschrijven:

$$\alpha = \frac{K}{K-1}\left(\frac{s_X^2 - \sum_{i=1}^{K} s_{Y_i}^2}{s_X^2}\right) = \frac{K}{K-1}\left(\frac{Var(\sum_{i=1}^{K} Y_i) - \sum_{i=1}^{K} Var(Y_i)}{Var(\sum_{i=1}^{K} Y_i)}\right)$$

$$= \frac{K}{K-1}\left(\frac{\sum\sum_{i \neq j} s_{Y_i Y_j}}{\sum\sum_{i,j} s_{Y_i Y_j}}\right) \qquad (3.7)$$

waarbij $s_{Y_i Y_j}$ de covariantie van het paar (Y_i, Y_j) voorstelt.

Cronbach‟s α kan dus geschat worden door empirische varianties of covarianties in formule (3.7) te substitueren. Het is bekend dat klassieke schatters, zoals empirische varianties en covarianties, zeer gevoelig zijn voor slechts een paar foute observaties [15]. De schatting van Cronbach‟s α kan dus een vertekend beeld geven als er zich uitbijters in de data bevinden. In [15] wordt een robuuste schatter van de betrouwbaarheid geïntroduceerd. Deze schatter voor de betrouwbaarheid wordt verkregen door de covariantiematrix van de itemscores te schatten aan de hand van bestaande robuuste schatters voor de covariantie. De elementen van deze covariantiematrix worden daarna gebruikt in formule (3.7).

Een ander probleem, dat regelmatig voorkomt bij het rapporteren van Cronbach‟s α, is dat onderzoekers meestal de aanname doen dat een bepaald niveau van α gewenst of adequaat is. Meestal wordt dit niveau gelijk aan 0.70 genomen [14]. Aangezien puntschatters misleidend kunnen zijn, is het raadzaam om naast de waarde voor Cronbach‟s α ook een betrouwbaarheidsinterval voor deze schatter te rapporteren. Hoe een dergelijk interval af te leiden zal nu aan bod komen.

Betrouwbaarheidsinterval

In [16] wordt de aanname gedaan dat de itemscores $\boldsymbol{Y} = [Y_1 .... Y_K]'$ een multivariate normale verdeling hebben. Onder deze aanname is vergelijking (3.3) te beschouwen als de meest aannemelijke schatter (MAS) van Cronbach‟s α. Deze MAS duiden we aan met $\hat{\alpha}$ en uit [16] volgt dat als $n \to \infty$:

$\sqrt{n}(\hat{\alpha}-\alpha)$ normaal verdeeld is met verwachting 0 en variantie

$$Q=\left[\frac{2K^2}{(K-1)^2(\iota'\Phi\iota)^3}\right]\cdot\left[(\iota'\Phi\iota)(tr\Phi^2+tr^2\Phi)-2(tr\Phi)(\iota'\Phi^2\iota)\right] \qquad (3.8)$$

Waarbij *K* en *n* zoals hiervoor respectievelijk het aantal items en het aantal respondenten voorstellen. Φ is de *K x K* covariantiematrix van de itemscores en ι is een vector van enen ter lengte *K*. *tr* stelt de trace-operator voor die de elementen op de hoofddiagonaal van een matrix bij elkaar optelt.

Op basis van bovenstaande resultaten kunnen we een betrouwbaarheidsinterval voor de schatter $\hat{\alpha}$ afleiden. Uit de verdeling van $\sqrt{n}(\hat{\alpha}-\alpha)$ volgt dat:

$$\frac{\sqrt{n}(\hat{\alpha}-\alpha)}{\sqrt{Q}}\xrightarrow{d}N(0,1),\qquad als\ n\to\infty. \qquad (3.9)$$

Aangezien voor grote *n* de grootheid $\frac{\sqrt{n}(\hat{\alpha}-\alpha)}{\sqrt{Q}}$ bij benadering standaard normaal verdeeld is, kan deze zodoende als een "bijna pivot[2]" beschouwd worden. Vervolgens kunnen we met behulp van de "bijna pivot" en de definitie van het β-bovenkwantiel ξ_β[3] van de standaard normale verdeling het volgende afleiden:

$$\lim_{n\to\infty}P\left(\frac{\sqrt{n}\,|\hat{\alpha}-\alpha|}{\sqrt{Q}}\le\xi_{\beta/2}\right)=1-\beta. \qquad (3.10)$$

Dit is te herschrijven naar:

$$\lim_{n\to\infty}P\left(\hat{\alpha}-\xi_{\beta/2}\sqrt{\tfrac{Q}{n}}\le\alpha\le\hat{\alpha}+\xi_{\beta/2}\sqrt{\tfrac{Q}{n}}\right)=1-\beta. \qquad (3.11)$$

Uit bovenstaande uitdrukking volgt dat $[\hat{\alpha}-\xi_{\beta/2}\sqrt{\frac{Q}{n}},\hat{\alpha}+\xi_{\beta/2}\sqrt{\frac{Q}{n}}]$ een benaderend betrouwbaarheidsinterval is voor α met onbetrouwbaarheid β. Voor het gemak wordt de uitdrukking *Q* in het betrouwbaarheidsinterval vaak vervangen door een schatter $\hat{Q}$, die we als volgt definiëren:

$$\hat{Q}=\left[\frac{2K^2}{(K-1)^2(\iota'S\iota)^3}\right]\cdot\left[(\iota'S\iota)(trS^2+tr^2S)-2(trS)(\iota'S^2\iota)\right], \qquad (3.12)$$

waarbij *S* een schatter van de covariantiematrix voorstelt.

Een benaderend 95% betrouwbaarheidsinterval voor α is nu als onderstaand symmetrisch interval te schrijven:

---

[2] Een pivot is een functie $T(X,\theta)$ van de waarneming en de parameter θ (dus geen statistiek) zodanig dat de kansverdeling van $T(X,\theta)$ onder de aanname dat θ de ware parameter is een vaste verdeling bezit, niet afhankelijk van θ of andere onbekenden [18].

[3] We noteren met ξ_β het getal zodanig dat $1-\Phi(\xi_\beta)=\beta$.

$$[\hat{\alpha} - 1.96\sqrt{\tfrac{\hat{Q}}{n}},\ \hat{\alpha} + 1.96\sqrt{\tfrac{\hat{Q}}{n}}]. \qquad (3.13)$$

Uit het betrouwbaarheidsinterval kunnen we ook de standaardfout (standard error) van de schatter aflezen. De standaardfout van $\hat{\alpha}$ is gelijk aan $\sqrt{\frac{\hat{Q}}{n}}$.

In [17] wordt geadviseerd om bij de berekening van Cronbach‟s α ook een betrouwbaarheids-interval voor deze schatting te rapporteren. Een onderzoeker zou eventueel ook een toets kunnen uitvoeren om na te gaan of Cronbach‟s α significant verschilt van een van te voren vastgestelde drempelwaarde. Het probleem hierbij is dat het niet duidelijk is welke hypothese moet worden getoetst, oftewel met welke waarde de verkregen Cronbach‟s α vergeleken moet worden. In [17] wordt gesteld dat het voldoende is om naast de waarde van Cronbach‟s α alleen het betrouwbaarheidsinterval te vermelden. Een betrouwbaarheidsinterval geeft namelijk het domein weer van de meest aannemelijke waarden voor α en een waarde van α, die buiten dit interval valt, zal één van de waarden zijn die bij een nulhypothese vergelijking wordt verworpen.

3.3 Item-analyse

Na het toepassen van een Likertschaal vragenlijst onder een groep proefrespondenten kan, zoals in paragraaf 3.2.2 staat beschreven, de betrouwbaarheid van de vragenlijst worden geschat met behulp van Cronbach‟s α. Tijdens deze fase kan een onderzoeker ook een *item-analyse* uitvoeren om de kwaliteit van de vragenlijst te verbeteren. Item-analyse is de techniek die standaard wordt gebruikt bij de analyse van een Likertschaal vragenlijst en komt neer op het verwijderen van de items, die niet blijken bij te dragen aan de kwaliteit van de schaal. De analyse wordt herhaald totdat de schaal niet verder verbeterd kan worden.

In [9] valt te lezen dat het behouden of verwijderen van items onder andere kan worden bepaald door de *item-rest correlatie* of α *indien item verwijderd* te berekenen. Deze twee technieken zullen hieronder kort beschreven worden.

Item-rest correlatie

Deze techniek heeft als doel om de samenhang tussen een afzonderlijk item en een schaal met de overige items te bepalen. Elk item in een schaal moet namelijk duidelijk positief correleren met alle andere items samen. Om de item-rest correlatie te berekenen wordt het betreffende item bij de berekening uit het totaal weggelaten, om te voorkomen dat de correlatie hoog uitvalt. Zo krijgen we een indicatie voor de mate waarin het item hetzelfde meet als de andere items. De item-rest correlatie r_{iR} voor item i kan dus als volgt berekend worden:

$$r_{iR} = \rho(Y_i, R_i), \qquad (3.14)$$

waarbij $\rho(X,Y)$ de correlatiecoëfficiënt tussen de stochastische variabelen X en Y is en $R_i = \sum_{k=1:k\neq i}^{K} Y_k$ de schaalscore van een respondent voorstelt, waarbij de score op item i is weg gelaten.

De item-rest correlatie voor een bepaald item *i* moet, volgens [12], tenminste 0.35 zijn. Wel wordt de kanttekening gemaakt dat een item met een $r_{iR} < 0.35$ soms wordt gehandhaafd als het item inhoudelijk niet kan worden gemist. Om te voorkomen dat het begrip te veel wordt versmald kan men zodoende een minimum van 0.20 voor r_{iR} nemen [12].

α indien item verwijderd

Een andere techniek om items, die niet blijken bij te dragen aan de kwaliteit van de schaal, op te sporen, is om voor ieder item *α indien item verwijderd* te berekenen. *α indien item verwijderd* is gelijk aan de Cronbach"s α van de schaal zonder dat betreffende item. Stel dat de oorspronkelijke vragenlijst een Cronbach"s α heeft gelijk aan *a* dan is het item waarvoor geldt dat *α indien item verwijderd* groter is dan *a*, een kandidaat om uit de vragenlijst verwijderd te worden. Op deze manier worden items, die een negatieve invloed hebben op Cronbach"s α, verwijderd uit de vragenlijst. Natuurlijk is het tijdens deze analyse verstandig om naast *α indien item verwijderd* ook een betrouwbaarheidsinterval voor deze schatting van *α* te berekenen, zodat met meer zekerheid items al dan niet verwijderd kunnen worden.

Naast *item-rest correlatie* en *α indien item verwijderd* kunnen ook andere maatstaven beschouwd worden om inzicht te krijgen in de kwaliteit van de items in een vragenlijst. Uit [17] volgt dat een gebruikelijke stap in de ontwikkeling van een schaal het berekenen van de *item-totaal correlaties* is. Deze correlatie tussen de item- en totaalscores dient als eerste inzicht om items op te sporen die ongunstige indicatoren zijn van het construct. Vervolgens kunnen de item-rest correlaties berekend worden om te corrigeren voor de invloed van de itemscore op de totaalscore. Een andere eerste analyse van een schaal is het bekijken van de *inter-itemcorrelaties* zoals in [6] gedaan wordt. In paragraaf 3.4.1 wordt uitgelegd dat deze analyse ook verricht kan worden om de convergente validiteit van een vragenlijst te onderzoeken.

Het uitvoeren van een item-analyse, nadat de vragenlijst is uitgezet onder een groep proefrespondenten, kan ertoe leiden dat de kwaliteit van de vragenlijst verder verbeterd wordt. Toch betekent dit niet dat de kwaliteit van een verzameling items enkel en alleen op grond van deze criteria beoordeeld moet worden [12]. De items moeten namelijk ook inhoudelijk geacht worden hetzelfde meetdoel te hebben.

3.4 Validiteit

Om een goede vragenlijst te verkrijgen dient naast de betrouwbaarheid evenzeer de validiteit, ook wel geldigheid, van de vragenlijst onderzocht te worden. Validiteit verwijst naar hoe goed de vragenlijst meet wat het behoort te meten. Een conditie voor een valide vragenlijst is dat de vragenlijst betrouwbaar is [13]. Een onbetrouwbare vragenlijst kan namelijk nooit valide zijn. Er zijn verschillende typen validiteit. Eén categorisatie van validiteit is in Stap 8 van paragraaf 1.2 al kort aan bod gekomen. In deze paragraaf zullen we ons bezighouden met andere typen validiteit, te weten *convergente* (paragraaf 3.4.1) en *discriminante validiteit* (paragraaf 3.4.2). Deze typen validiteit zijn subcategorieën van *construct validiteit*. Construct validiteit zegt iets over de mate waarin deelaspecten van een omvangrijk begrip het hele begrip dekken. Het onderzoeken van construct validiteit heeft als doel om te achterhalen of er misschien onvoldoende rekening is gehouden met andere variabelen die ook van invloed zijn

op het onderzochte begrip. De resultaten van een vragenlijst kunnen namelijk op het eerste gezicht perfect lijken aan te sluiten bij de theorie, terwijl er bepaalde aspecten, die aan een begrip te onderkennen zijn, ongemeten blijven.

3.4.1 Convergente validiteit

Convergente validiteit heeft te maken met de veronderstelling dat metingen van een bepaald construct, die in theorie met elkaar verbonden zijn, in werkelijkheid ook met elkaar verbonden dienen te zijn. Om convergente validiteit te onderzoeken kan bijvoorbeeld gekeken worden naar de samenhang tussen de resultaten van het oorspronkelijke onderzoek en de resultaten van een gelijksoortig onderzoek. Als schatting van convergente validiteit van een vragenlijst wordt vaak de correlatiecoëfficiënt tussen de scores van de verschillende metingen genomen. In [6] worden de items in een Likertschaal vragenlijst gezien als verschillende metingen van éénzelfde construct. Dat betekent dat de items in theorie het construct meten en dat de correlatie tussen de verschillende itemscores een indicatie geeft van in hoeverre de verschillende items gerelateerd zijn aan hetzelfde construct. Indien de correlaties tussen de verschillende items hoog uitvallen, is het zeer aannemelijk dat de items hetzelfde construct meten, maar dat betekent niet dat je automatisch het gewenste construct meet.
Convergente validiteit kan ook onderzocht worden door te kijken naar de samenhang tussen de resultaten van een onderzoek en observeerbaar gedrag.

3.4.2 Discriminante validiteit

Een vragenlijst bezit voldoende discriminante validiteit indien metingen, die theoretisch geen verband hebben met elkaar, dat in werkelijkheid ook niet hebben. Om dit aan te tonen kan bijvoorbeeld gekeken worden naar de samenhang tussen de onderzoeksresultaten en de resultaten van een andersoortig onderzoek. Een lage correlatie tussen de resultaten maakt het aannemelijk dat we met beide onderzoeken verschillende constructen meten. In [6] wordt discriminante validiteit aangetoond door de correlatiecoëfficiënten tussen de itemscores behorende bij twee verschillende, maar verwante, constructen te berekenen. Indien de correlaties laag zijn is het zeer waarschijnlijk dat de items verschillende constructen meten.

Het moet gezegd worden dat er nog veel andere vormen van validiteit bestaan, die na de constructie van een vragenlijst onderzocht kunnen worden. In dit werkstuk zullen deze verder niet aan bod komen.

Hoofdstuk 4

Praktijkstudie: Likertschaal vragenlijst construeren, uitzetten en analyseren

4.1 Introductie

In dit laatste hoofdstuk zal de theorie, die in de voorgaande hoofdstukken is besproken, worden toegepast in een praktijksituatie. Om dit doel te bewerkstelligen is een vragenlijst ontwikkeld om de tevredenheid van studenten over hun studie te meten. Deze vragenlijst is verspreid onder studenten aan de Faculteit der Exacte Wetenschappen (FEW) van de Vrije Universiteit Amsterdam. Hoe de definitieve vragenlijst tot stand is gekomen, is te lezen in paragraaf 4.2. In paragraaf 4.3 worden de technieken uit hoofdstuk 3 toegepast om de vragenlijst te analyseren en verder te verbeteren. Het leek ons ook interessant om iets te zeggen over de resultaten van het onderzoek onder de studenten. In paragraaf 4.4 zullen de resultaten van de verbeterde vragenlijst geanalyseerd worden.

4.2 Opstellen van de vragenlijst

Bij het opstellen van de vragenlijst om de tevredenheid van FEW studenten over hun studie te meten, is getracht om zoveel mogelijk de stappen van paragraaf 1.2 door te lopen. Zo is eerst gekeken naar de belangrijkste aspecten die invloed hebben op de tevredenheid van studenten. In ons onderzoek is gekozen voor de volgende aspecten:

- *Inhoud van het studieprogramma*
- *Kwaliteit van de docenten*
- *Studielast*
- *Faciliteiten van de Faculteit der Exacte Wetenschappen*

Voor ieder categorie zijn er stellingen geformuleerd die de tevredenheid over dat deelaspect meten. Wat betreft de stellingen is getracht om deze op een eenduidige manier te formuleren. Vanwege de beperkte tijd voor het onderzoek is de vragenlijst niet uitgebreid getest in een pilot. Wel zijn opmerkingen van medestudenten en de begeleider van dit werkstuk verwerkt in de definitieve vragenlijst. Dit heeft geleid tot een vragenlijst bestaande uit 34 vragen. De eerste zes vragen zijn algemeen van aard en de andere 28 vragen zijn stellingen, die betrekking hebben op de vier bovenstaande aspecten. Als antwoordschaal voor de stellingen is gebruik gemaakt van de Likertschaal met een even aantal antwoordcategorieën. Bij een dergelijk aantal antwoordcategorieën wordt de respondent bij ieder item gedwongen een antwoord richting de “mee eens” of “mee oneens” einde van de schaal te kiezen. De antwoordcategorieën met de daaraan toegekende scores zijn in tabel 4.1 te vinden.

Antwoordcategorie	***Score***
Disagree strongly	*1*
Disagree	*2*
Tend to disagree	*3*
Tend to agree	*4*
Agree	*5*
Agree strongly	*6*

Tabel 4.1: Antwoordcategorieën met bijbehorende score

De definitieve vragenlijst, die aan de FEW studenten is voorgelegd, is opgenomen in Appendix B. Via de mail zijn alle studenten van de faculteit benaderd om aan het onderzoek mee te werken. In deze mail is informatie verschaft over het doel van het onderzoek en is getracht om de studenten over te halen de vragenlijst in te vullen. Studenten konden via een link in de mail naar de website met de vragenlijst gaan. Gebruik is gemaakt van de site www.esurveyspro.com om de vragenlijst op te stellen. Op deze website kunnen verschillende soorten vragenlijsten ontworpen en beheerd worden. Via een eigen account kan inzicht in de resultaten van de vragenlijst verkregen worden en kunnen de antwoorden van de respondenten op de vragen worden geëxporteerd naar bijvoorbeeld een csv file.

De vragenlijst heeft drie weken op de site gestaan en in totaal hebben 244 studenten de vragenlijst beantwoord. In de twee volgende paragrafen zullen analyses gedaan worden op de antwoorden van de respondenten.

4.3 Analyse van de vragenlijst

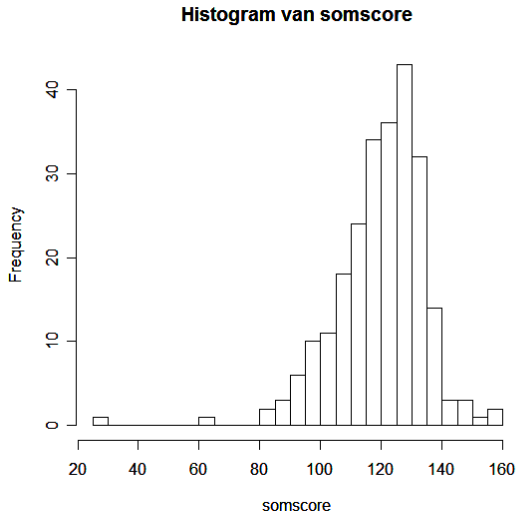
Het statistisch softwarepakket R is gebruikt om de vragenlijst te analyseren. De gecodeerde antwoorden van de respondenten zijn in een 244 *x* 34 matrix gestopt, waarbij in iedere rij de antwoorden van een respondent op de vragen is te lezen. Zodoende heeft iedere kolom van de matrix betrekking op een vraag uit de vragenlijst. Aangezien het berekenen van onder andere Cronbach‟s α en het uitvoeren van een item-analyse alleen op items uitgevoerd kan worden is een nieuwe matrix geconstrueerd, waarin alleen de itemscores van de respondenten in voorkomen. De antwoorden op de algemene vragen van de vragenlijst zijn hierbij dus gescheiden van de itemscores. Omdat de vragenlijst uit zes algemene vragen bestaat, houden we nu een matrix over met 244 (aantal respondenten) rijen en 28 (aantal items) kolommen. Vervolgens zijn de antwoorden op items, die negatief gesteld zijn, gehercodeerd. In de gebruikte vragenlijst (zie Appendix B) is alleen item 6 negatief gesteld. Het hercoderen van de itemscores kan simpelweg uitgevoerd worden door de volgende bewerking:

nieuwe itemscore = 7-oude itemscore.

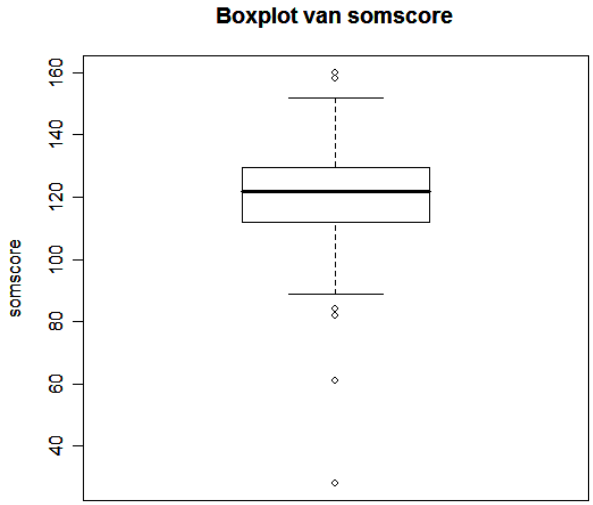
4.3.1 Verwijderen van uitbijters

Voorafgaand aan de analyse van de vragenlijst zal gekeken moeten worden of er zich misschien uitbijters in de data bevinden. Het kan natuurlijk voorkomen dat respondenten de vragenlijst hebben ingevuld zonder echt na te denken bij iedere vraag. Extreem grote of kleine waarnemingen kunnen dan een vertekend beeld geven en leiden tot verkeerde bevindingen. In paragraaf 3.2.3 is opgemerkt dat de Cronbach‟s α betrouwbaarheidscoëfficiënt geen robuuste schatter is. Het is dus van belang om uitbijters uit de data te verwijderen om juiste uitspraken over de betrouwbaarheid van de schaal te kunnen doen.

Inzicht in mogelijke uitbijters kan verkregen worden door een boxplot van de somscores van de respondenten te maken. In figuur 4.2 is deze boxplot afgebeeld. Ook het histogram van de somscores (figuur 4.1) biedt inzicht in de spreiding van de somscores.



Figuur 4.1: Histogram van de somscores



Figuur 4.2: Boxplot van de somscores

Uit de boxplot is af te leiden dat 6 somscores als extreme waarden beschouwd kunnen worden. Het bekijken van het antwoordpatroon van deze respondenten heeft ertoe geleid dat alleen de kleinste somscore als uitschieter is aangemerkt. Deze respondent heeft namelijk alle vragen met *"volledig mee oneens"* beantwoord. De vijf andere extreme waarden, die uit de boxplot volgen, liggen relatief dicht bij de maximale en minimale waarden die geen uitschieter zijn[4]. De maximale en minimale waarden die geen uitschieter zijn, worden in de boxplot met een horizontale lijn weergegeven. Ook als we naar het histogram kijken zien we dat het kleinste datapunt betrekkelijk ver ligt van de rest van de data.

4.3.2 Betrouwbaarheid en item-analyse

In R zijn functies (zie Appendix C) geprogrammeerd waarmee de *Spearman-Brown split half betrouwbaarheidscoëfficiënt* berekend kan worden voor verschillende splitsingen van een schaal in twee subschalen. Er is voor gekozen om deze betrouwbaarheidscoëfficiënt te berekenen voor een splitsing gebaseerd op itemnummer. Het splitsen gebeurt in dit geval door items met een even nummer in de ene subschaal te plaatsen en items met een oneven nummer in de andere subschaal. De Spearman-Brown split half betrouwbaarheidscoëfficiënt van de schaal, die wij ontwikkeld hebben om de tevredenheid van studenten over hun studie te meten, is bij een dergelijke „systematische" splitsing gelijk aan 0.9243.

Ook is het mogelijk om de Spearman-Brown split half betrouwbaarheidscoëfficiënt te berekenen voor een willekeurige splitsing van de items in twee subschalen. Als we een dergelijke „random" splitsing uitvoeren is de betrouwbaarheidscoëfficiënt van de ontwikkelde schaal gelijk aan 0.9093. Omdat de betrouwbaarheidscoëfficiënt afhangt van hoe de schaal gesplitst wordt, is het informatiever om te kijken naar het gemiddelde van bijvoorbeeld 1000 betrouwbaarheidscoëfficiënten verkregen na „random" splitsing. Dit gemiddelde is 0.8818.

[4] De maximale waarde die geen uitschieter is, is het grootste datapunt dat binnen 1.5 maal de interkwartielafstand vanaf het derde kwartiel verwijderd ligt. Analoog geldt dat de minimale waarde die geen uitschieter is, het kleinste datapunt is dat binnen 1.5 maal de interkwartielafstand vanaf het eerste kwartiel ligt.

Op basis van de berekende Spearman-Brown split half betrouwbaarheidscoëfficiënten bij zowel een „systematische" als een „random" splitsing van de schaal, kan verondersteld worden dat de ontwikkelde schaal een goede betrouwbaarheid heeft. Uiteraard is in R ook een functie geprogrammeerd waarmee de *Cronbach's* α betrouwbaarheidscoëfficiënt van een vragenlijst berekend kan worden. Uit deze functie blijkt dat de vragenlijst een Cronbach's α heeft gelijk aan 0.8796. Een 95%-betrouwbaarheidsinterval voor deze schatter is gelijk aan [0.8512, 0.9081]. Uit de waarde van Cronbach's α en het betrouwbaarheidsinterval wordt de eerdere conclusie over de betrouwbaarheid van de schaal bevestigd. De Cronbach's α betrouwbaarheidscoëfficiënt is namelijk een stuk groter dan het veelgebruikte niveau van 0.70. Ook is de ondergrens van het betrouwbaarheidsinterval groter dan dit niveau. Het is echter de moeite waard om te onderzoeken of de kwaliteit van deze vragenlijst verder verbeterd kan worden.

Item-totaal correlaties

Om een eerste inzicht te krijgen in items, die ongunstige indicatoren zijn van het construct, zal voor ieder item de correlatie tussen de item- en totaalscores berekend worden. In tabel 4.2 zijn deze correlaties voor de ontwikkelde schaal te vinden. Ook is voor ieder item de overschrijdingskans berekend als we de item-totaal correlatiecoëfficiënten toetsen met behulp van de Pearson's correlatietoets in R. Met behulp van deze correlatietoets kunnen wij onderzoeken of de item- en totaalscores positief samenhangen. Zodoende hebben we de volgende toets:

H_0: de correlatie tussen de itemscore en de totaalscore is kleiner of gelijk aan 0.
H_1: de correlatie tussen de itemscore en de totaalscore is groter dan 0.

Indien aan de hand van de correlatietoets een overschrijdingkans (p-value) gevonden wordt, die kleiner is dan de gekozen onbetrouwbaarheidsdrempel (α_0), dan zal de nulhypothese verworpen worden. In dat geval is het aannemelijk dat de item- en totaalscore positief samenhangen. Items met een positieve samenhang met de totaalscore zou je het liefst in de vragenlijst willen houden. De correlatietoets is in dit geval een eenzijdige toets, waarbij α_0 gelijk aan 0.05 genomen wordt.

Item nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Item-totaal corr.	0.6157	0.6061	0.4852	0.5907	0.5716	0.4548	0.5512
P-value	0.0000	0.0000	4.4e-16	0.0000	0.0000	4.2e-14	0.0000

8	**9**	**10**	**11**	**12**	**13**	**14**
0.5965	0.5773	0.6169	0.6082	0.5047	0.4427	*0.1631*
0.0000	0.0000	0.0000	0.0000	0.0000	2.2e-13	0.0054

15	**16**	**17**	**18**	**19**	**20**	**21**
0.4440	*0.3160*	0.4848	0.5408	0.4661	0.5628	0.4299
1.8e-13	2.4e-07	4.4e-16	0.0000	8.3e-15	0.0000	1.2e-12

22	**23**	**24**	**25**	**26**	**27**	**28**
0.3862	0.4978	*0.3484*	0.3996	0.7246	0.7210	0.4146
2.3e-10	1.1e-16	1.2e-08	5.0e-11	0.0000	0.0000	8.2e-12

Tabel 4.2: Item-totaal correlaties met overschrijdingskansen correlatietoets

Uit bovenstaande tabel valt op te merken dat drie items een relatief lage item-totaal correlatie hebben. Hierbij gaat het om de items 14, 16 en 24.

Uit de overschrijdingskansen van de correlatietoets volgt dat voor de items 14, 16 en 24, zoals voor alle items, de nulhypothese wordt verworpen. De overschrijdingskansen zijn namelijk kleiner dan de onbetrouwbaarheidsdrempel van 0.05. We kunnen dus niet stellen dat er geen of een negatieve samenhang is tussen deze items en de totaalscores. Op basis van de correlatietoets kunnen we deze items dus niet zomaar uit de vragenlijst weglaten. Aan de hand van andere analyses hopen we meer te kunnen zeggen over de bruikbaarheid van deze items in de vragenlijst.

Item-rest correlaties

Zoals in paragraaf 3.3 aan bod is gekomen is het ook interessant te kijken naar de item-rest correlaties. Dit criterium geeft de samenhang weer tussen een afzonderlijk item en de gehele schaal, gebaseerd op de overige items. In onderstaande tabel kunnen de item-rest correlaties van de items in de vragenlijst teruggevonden worden. Tevens zijn ook de overschrijdingskansen van de correlatietoets op de item-rest correlaties in deze tabel opgenomen.

Item nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Item-rest corr.	0.5785	0.5621	0.4354	0.5489	0.5330	0.3741	0.5148
P-value	0.0000	0.0000	5.8e-13	0.0000	0.0000	8.6e-10	0.0000

8	**9**	**10**	**11**	**12**	**13**	**14**
0.5489	0.5365	0.5674	0.5682	0.4589	0.3793	*0.0831*
0.0000	0.0000	0.0000	0.0000	2.3e-14	4.9e-10	0.0985

15	**16**	**17**	**18**	**19**	**20**	**21**
0.3765	*0.2350*	0.4248	0.4916	0.3873	0.5120	0.3708
6.7e-10	0.0001	2.3e-12	2.2e-16	2.0e-10	0.0000	1.2e-09

22	**23**	**24**	**25**	**26**	**27**	**28**
0.3319	0.4319	*0.2574*	*0.3039*	0.6878	0.6797	0.3645
5.8e-08	9.2e-13	2.4e-05	6.9e-07	0.0000	0.0000	2.4e-09

Tabel 4.3: Item-rest correlaties met overschrijdingskansen correlatietoets

Uit bovenstaande tabel met item-rest correlaties valt ons op dat de items 14, 16 en 24 wederom een relatief lage correlatie vertonen. Als we rekening houden met de veelgebruikte ondergrens van 0.35, die in paragraaf 3.3 is besproken, komen ook de items 22 en 25 in aanmerking voor nader onderzoek. De items met item-rest correlaties iets groter dan 0.35 dienen uiteraard ook aandachtig bestudeerd te worden. Uit de overschrijdingskansen van de correlatietoets, die allemaal kleiner zijn dan α_0, valt op te maken dat niet verondersteld mag worden dat de samenhang tussen de items en de som van de restscores negatief is dan wel afwezig is. Wederom geeft de correlatietoets ons geen ondersteuning om ongeschikt lijkende items uit de vragenlijst weg te laten.

Inter-itemcorrelaties

Naast de item-totaal correlaties en de item-rest correlaties is het bekijken van de inter-itemcorrelaties een manier om na te gaan in hoeverre de items in een schaal hetzelfde construct meten. De meest opvallende inter-itemcorrelaties zijn in tabel 4.4 opgenomen. Om iets meer te kunnen zeggen over de samenhang van de geselecteerde items is ook de correlatietoets op de inter-itemcorrelatiecoëfficiënten toegepast. De overschrijdingskansen van deze toets zijn ook in tabel 4.4 te lezen.

Item nr.	*Item nr.*	*Correlatie-coëfficiënt*	*Overschrijdings-kans*
12	**24**	0.0060	0.4631
13	**19**	-0.0143	0.5875
14	**1**	-0.0249	0.6502
14	**2**	-0.0781	0.8875
14	**5**	-0.0044	0.5274
14	**6**	-0.1030	0.9453
14	**8**	-0.1014	0.9426
14	**9**	-0.0187	0.6142
14	**10**	-0.0442	0.7535
14	**16**	-0.0358	0.7103
14	**24**	-0.0164	0.6001
16	**17**	0.0066	0.4591
16	**21**	-0.0453	0.7592
16	**22**	0.0056	0.4656
16	**23**	-0.0022	0.5136
16	**24**	-0.0289	0.6729
16	**28**	-0.0131	0.5808

Tabel 4.4: Inter-item correlaties met overschrijdingskansen correlatietoets

We zien dat de correlatiecoëfficiënten in bovenstaande tabel allemaal dichtbij 0 liggen. De overschrijdingskansen van de correlatietoets geven aan dat de nulhypothese in geen enkel geval verworpen mag worden. Het is dus aannemelijk dat de bovenstaande items onderling geen samenhang vertonen of misschien zelfs negatieve samenhang. Wat ons opvalt, is dat item 14 en item 16 regelmatig terugkomen in bovenstaande tabel. Ook in eerdere analyses is gebleken dat deze twee items ongunstig zijn voor de schaal. Het is dus zeer waarschijnlijk dat de kwaliteit van de schaal toeneemt als we deze twee items uit de vragenlijst verwijderen. In vervolganalyses hopen we dit te kunnen aantonen.

α indien item verwijderd

Een manier om te onderzoeken of de kwaliteit van een schaal verbetert wanneer een bepaald item uit de schaal verwijderd wordt, is *α indien item verwijderd* (zie ook paragraaf 3.3). De resultaten van deze techniek op de ontwikkelde vragenlijst zijn te vinden in onderstaande tabel.

Item nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
α item verwijderd	0.8731	0.8728	0.8756	0.8734	0.8742	0.8778	0.8748

8	**9**	**10**	**11**	**12**	**13**	**14**
0.8729	0.8738	0.8722	0.8731	0.8753	0.8769	***0.8846***

15	**16**	**17**	**18**	**19**	**20**	**21**
0.8770	*0.8812*	0.8757	0.8744	0.8773	0.8737	0.8770

22	**23**	**24**	**25**	**26**	**27**	**28**
0.8777	0.8756	*0.8817*	*0.8811*	0.8695	0.8691	0.8771

Tabel 4.5: *α indien item verwijderd* originele vragenlijst

Uit bovenstaande tabel kunnen we afleiden dat voor vier items geldt dat *α indien item verwijderd* groter is dan de Cronbach's α van de originele vragenlijst (0.8796). Het gaat hierbij om de items 14, 16, 24 en 25. Ook uit eerdere analyses is gebleken dat onder andere deze items de kwaliteit van de schaal negatief beïnvloeden. We zien dat de hoogste Cronbach's α wordt bereikt wanneer item 14 uit de schaal wordt verwijderd. De stelling, die bij dit item hoort, "*Next to my study I keep enough time to do other things"*, hoeft inderdaad niet de tevredenheid van studenten over hun studie te meten. Andere zaken kunnen namelijk ook bepalen of iemand veel of weinig tijd over heeft naast de studie.

De Cronbach's α van de vragenlijst zonder item 14 is dus gelijk aan 0.8846 en het 95%-betrouwbaarheidsinterval is gelijk aan [0.8567, 0.9124]. Deze verkregen schaal gaan we opnieuw analyseren met als doel om eventueel andere items te vinden die uit de schaal

verwijderd kunnen worden. Wederom zal *α indien item verwijderd* voor de items in deze schaal berekend worden. De resultaten zijn te vinden in onderstaande tabel.

Item nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
α item verwijderd	0.8783	0.8779	0.8811	0.8788	0.8793	0.8828	0.8802

8	**9**	**10**	**11**	**12**	**13**	**14**
0.8779	0.8790	0.8773	0.8785	0.8807	0.8824	

15	**16**	**17**	**18**	**19**	**20**	**21**
0.8826	*0.8864*	0.8810	0.8796	0.8828	0.8791	0.8822

22	**23**	**24**	**25**	**26**	**27**	**28**
0.8830	0.8809	***0.8870***	*0.8866*	0.8748	0.8743	0.8824

Tabel 4.6: *α indien item verwijderd* schaal zonder item 14

Voor drie items geldt nu dat er en hogere waarde van Cronbach‟s α wordt bereikt wanneer het item uit de schaal wordt weggelaten. Dit zijn de items 16, 24 en 25. Het verwijderen van item 24 leidt tot de hoogste waarde van de betrouwbaarheidscoëfficiënt. Ook is dit item in vorige analyses naar boven gekomen als een kwalitatief ongunstig item voor de schaal. De stelling die bij dit item hoort, namelijk *"The library of the faculty of sciences is easily accessible"*, meet dus waarschijnlijk niet de tevredenheid van studenten over hun studie.

De Cronbach‟s α betrouwbaarheidscoëfficiënt van de schaal, die overblijft wanneer item 14 en 24 worden weggelaten, is dus gelijk aan 0.8870. Het 95%-betrouwbaarheidsinterval van deze coëfficiënt is [0.8593, 0.9146].

De procedure van *α indien item verwijderd* is vervolgens een aantal keer toegepast totdat de schaal niet verder verbeterd kon worden. De uitgevoerde R-code is terug te vinden in Appendix D. Deze procedure heeft achtereenvolgens tot de volgende resultaten geleid:

- Na item 14 en 24 blijkt dat het verwijderen van item 25 ("*The faculty provides sufficient places for self-study*") te leiden tot een hogere betrouwbaarheid van de schaal. De resulterende schaal heeft een betrouwbaarheid van 0.8912 met een 95%-betrouwbaarheidsinterval gelijk aan [0.8639, 0.9184].
- Vervolgens leidt het verwijderen van item 16 ("*I attend the lectures regularly*") tot een hogere betrouwbaarheid. We houden nu een schaal over met een betrouwbaarheid van 0.8935 ([0.8663, 0.9207]) .
- Daarna blijkt het verwijderen van item 19 ("*In the faculty there are sufficient computers available*") te resulteren in een hogere betrouwbaarheid, namelijk 0.8941 ([0.8667, 0.9215])
- Ten slotte wordt item 6 (*"Some courses are a waste of time"*) uit de vragenlijst verwijderd. Na het weglaten van dit item houden we een schaal over met een betrouwbaarheid van 0.8949 ([0.8673, 0.9226]). Het berekenen van *α indien item verwijderd* voor de overgebleven items in de schaal geeft de resultaten, die te vinden zijn in tabel 4.7.

Tabel 4.7 laat zien dat de betrouwbaarheid van de schaal niet verder toeneemt wanneer één van de overgebleven items uit de vragenlijst wordt weggelaten. Dat betekent dat we de hoogste waarde voor Cronbach‟s α hebben bereikt. De vragenlijst in Appendix B kan op basis van deze resultaten worden aangepast. Deze aangepaste vragenlijst kan als meetinstrument voor de tevredenheid van studenten worden toegepast op een grotere doelgroep.

Item nr.	1	2	3	4	5	6	7
α item verwijderd	0.8878	0.8881	0.8916	0.8887	0.8891		0.8901

8	9	10	11	12	13	14
0.8887	0.8892	0.8879	0.8888	0.8907	0.8935	

15	16	17	18	19	20	21
0.8948		0.8937	0.8909		0.8919	0.8934

22	23	24	25	26	27	28
0.8944	0.8937			0.8847	0.8846	0.8931

Tabel 4.7: *α indien item verwijderd* schaal zonder item 14, 25, 24, 16, 19 en 6

Het is ook interessant om te kijken of de Spearman-Brown split half betrouwbaarheids-coëfficiënt van de overgebleven schaal groter is dan die van de originele schaal. Deze coëfficiënt is berekend in het geval van een „systematische" splitsing van de schaal. Ook is het gemiddelde berekend van 1000 Spearman-Brown split half betrouwbaarheidscoëfficiënten verkregen na „random" splitsing van deze schaal. In tabel 4.8 zijn de belangrijkste indicatoren van betrouwbaarheid voor de originele en overgebleven schaal samengevat.

	Originele schaal	**Overgebleven schaal**
Cronbach's α	0.8796	0.8949
95%- Betrouwbaarheidsinterval α	[0.8512, 0.9081]	[0.8673, 0.9226]
Split half na 'systematische' splitsing van de schaal	0.9243	0.9274
Gemiddelde van 1000 'random' split half coëfficiënten	0.8818	0.8962

Tabel 4.8: *Indicatoren van betrouwbaarheid voor de originele en overgebleven schaal*

Uit bovenstaande tabel is op te maken dat naast Cronbach"s α ook andere maatstaven van betrouwbaarheid een iets beter resultaat laten zien voor de overgebleven schaal. We hebben dus nu via verschillende betrouwbaarheidscoëfficiënten aangetoond dat de kwaliteit van de schaal is toegenomen na het verwijderen van de zes items.

4.3.3 Conclusie

Na de analyses in deze paragraaf zijn een aantal logische items uit vragenlijst verwijderd. Zo zien we onder andere dat item 20 (*"The computing facilities are good"*) heel veel lijkt op het verwijderde item 19. Ook voor de andere stellingen is er iets voor te zeggen dat ze niet of nauwelijks de tevredenheid van studenten over hun studie bepalen. Toch kan een onderzoeker ervoor kiezen om sommige verwijderde vragen toch in de vragenlijst te houden. De oorspronkelijke vragenlijst laat namelijk een relatief goede betrouwbaarheid zien. Daarnaast merken we op dat we bij een andere doelgroep misschien tot andere resultaten waren gekomen. Desalniettemin zijn we ervan overtuigd dat we een voldoende grote steekproef hebben genomen en we dus met enige zekerheid kwalitatief ongunstige items uit de vragenlijst hebben verwijderd.
In de bovenstaande analyses is niet expliciet onderzocht of de items een eendimensionaal construct meten. Om dit te onderzoeken kunnen bijvoorbeeld de inter-itemcorrelaties bestudeerd worden. Aan de hand van deze correlaties kunnen we niet met zekerheid stellen dat de items een sterke samenhang vertonen. De meeste correlaties tussen de itemscores zijn namelijk niet erg groot. Echter is de maat voor de interne consistentie van onze multiple item schaal, oftewel Cronbach"s α, relatief hoog. Op basis van Cronbach"s α is het aannemelijk dat

de items een eendimensionaal construct meten in plaats van een meerdimensionaal construct. Om hier meer zekerheid over te krijgen zou eventueel een factor analyse kunnen worden uitgevoerd. In dit werkstuk besteden we verder geen aandacht aan deze techniek.

4.4 Analyse van de resultaten

In deze paragraaf zullen kort een aantal resultaten van het onderzoek worden geanalyseerd. We zijn hierbij voornamelijk geïnteresseerd of verschillende groepen studenten van elkaar verschillen qua tevredenheid. De uitgevoerde analyses hebben betrekking op de antwoorden van 243 respondenten, die zijn overgebleven na het verwijderen van één uitbijter uit de data. Bij het analyseren van de resultaten is ook alleen gekeken naar de antwoorden op 22 items. Dit zijn de overgebleven items van paragraaf 4.3, die leiden tot de hoogste betrouwbaarheid van de schaal. Een aantal numerieke samenvattingen met betrekking tot de somscores van de respondenten op de items zijn terug te vinden in tabel 4.9.

Gemiddelde somscore	97.86
Variantie somscore	126.78
Minimale somscore	44
Maximale somscore	129

Tabel 4.9: Numerieke samenvattingen m.b.t. somscores

Uit deze tabel valt ons meteen de hoge variantie van de somscores op. Een verklaring hiervoor is niet eenvoudig te geven. Verder valt op te merken dat de gemiddelde somscore van studenten erop duidt dat FEW studenten over het algemeen tevreden zijn over hun studie. Een gemiddelde somscore van 97.86 betekent namelijk dat de gemiddelde itemscore 4.45 (97.86/22) is. Numerieke samenvattingen met betrekking tot de itemscores zijn opgenomen in onderstaande tabel.

Item nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Gemiddelde itemscore	4.5967	4.5062	4.4568	4.6049	4.6420		4.6626
Variantie itemscore	0.6053	0.8047	0.7285	0.6945	0.5614		0.4724

8	**9**	**10**	**11**	**12**	**13**	**14**
4.3004	4.4444	3.9136	4.5226	4.8107	3.9465	
0.9052	0.6364	1.0462	0.6803	0.6500	1.0591	

15	**16**	**17**	**18**	**19**	**20**	**21**
3.9918		4.0864	4.1770		4.6214	4.3416
1.1983		1.0380	0.8157		0.9222	0.9035

22	**23**	**24**	**25**	**26**	**27**	**28**
4.5103	4.5967			4.5885	4.5144	5.0288
0.7137	1.2747			0.9622	1.1682	0.6397

Tabel 4.10: Numerieke samenvattingen m.b.t. itemscores

Ook de gemiddelden van de itemscores duiden aan dat de studenten over het algemeen tevreden zijn over hun studie. Bijna alle gemiddelden zijn groter dan 4 (*tend to agree*). De varianties van de itemscores zijn nu niet erg groot, wat natuurlijk ook te maken heeft met het feit dat respondenten slechts uit zes antwoordcategorieën kunnen kiezen.

Tijdens het analyseren van de resultaten van de vragenlijst leek het ons ook interessant om de volgende onderzoeksvragen te beantwoorden:

1 *Is er een significant verschil tussen de tevredenheid van bachelor en master studenten?*
2 *Is er een significant verschil tussen de tevredenheid van mannelijke en vrouwelijke studenten?*
3 *Zijn studenten, die relatief lang staan ingeschreven, minder tevreden over hun studie dan studenten die relatief kort staan ingeschreven?*

Deze drie onderzoeksvragen worden in de volgende subparagrafen beantwoord.

4.4.1 Tevredenheid van bachelor en master studenten

Om antwoord te vinden op de eerste onderzoeksvraag is de dataset met de antwoorden van alle respondenten verdeeld in twee aparte datasets. De eerste dataset bevat alleen de antwoorden van bachelor studenten en de tweede dataset alleen die van master studenten. Respondenten, die geen antwoord hebben gegeven op deze vraag zijn niet opgenomen in één van de twee datasets. Nu kunnen we de somscores van deze twee typen respondenten met elkaar vergelijken.

In totaal hebben 130 bachelor en 94 master studenten de vragenlijst beantwoord. De overige respondenten zijn of bezig met een pre-master programma of hebben simpelweg deze vraag onbeantwoord gelaten. Enkele numerieke samenvattingen over de twee datasets zijn te vinden in onderstaande tabel.

	Bachelor studenten	**Master studenten**
Gemiddelde somscore	98.85	96.80
Variantie somscore	91.17	181.80
Minimale somscore	73	44
Maximale somscore	129	124

Tabel 4.11: Numerieke samenvattingen m.b.t. somscores van bachelor en master studenten

Het is uit tabel 4.11 niet in één oogopslag duidelijk of bachelor studenten meer tevreden zijn over hun studie dan master studenten of andersom. Het gemiddelde van de somscores liggen namelijk dicht bij elkaar. Wel valt op te merken dat de variantie van de somscores van master studenten ongeveer twee keer zo groot is als die van bachelor studenten. We gaan middels een statistische toets onderzoeken of er een significant verschil is tussen de tevredenheid van beide type studenten.

Eerst gaan we onderzoeken of de somscores van de respondenten een normale verdeling hebben. Met behulp van de Shapiro-Wilk toets kan onderzocht worden of waarnemingen normaal verdeeld zijn. Deze toets heeft in ons geval als nulhypothese dat de somscores een normale verdeling bezitten. Aangezien deze toets een tweezijdige toets is, wordt de onbetrouwbaarheidsdrempel gelijk aan 0.025 ($\alpha_0/2$) genomen. De overschrijdingskans van de toets is voor bachelor studenten 0.1133 en voor master studenten is deze 0.00028. De overschrijdingkans is in één geval groter dan de onbetrouwbaarheidsdrempel van 0.025 . Dat betekent dat de somscores van bachelor studenten niet van een steekproef uit een normale verdeling zijn te onderscheiden. Daarentegen wordt de hypothese dat de somscores van master studenten normaal verdeeld zijn, verworpen.

Aangezien we geen zekerheid hebben over de verdeling van de somscores van master studenten kan het gebruik van de t-toets om de twee steekproeven met elkaar te vergelijken misleidend zijn. Om die reden zullen we gebruik maken van een verdelingsvrije toets. De twee univariate steekproeven zijn ongepaard en we nemen aan dat de twee steekproeven onafhankelijk zijn van elkaar. Een geschikte toets om de locaties van de twee steekproeven

met elkaar te vergelijken is de Mann Whitney-toets (Wilcoxon-twee-steekproeventoets). Met deze toets kunnen we onderzoeken of er een significant verschil is tussen de tevredenheid van bachelor en master studenten.

De *nulhypothese* is: De somscores van bachelor en master studenten zijn gelijk.
De *alternatieve hypothese* is: De somscores van bachelor en master studenten verschillen.
De onbetrouwbaarheidsdrempel wordt gelijk aan 0.025 genomen omdat we te maken hebben met een tweezijdige toets.

De overschrijdingskans, die uit de Mann Whitney-toets volgt, is 0.536. Deze waarde is groter dan de onbetrouwbaarheidsdrempel, waardoor de nulhypothese niet verworpen mag worden. Het is dus aannemelijk dat de gemiddelde somscore van bachelor studenten niet significant verschilt van die van master studenten. Dat betekent dat we kunnen veronderstellen dat bachelor en master studenten over het algemeen even tevreden zijn over hun studie.

4.4.2 Tevredenheid van mannelijke en vrouwelijke studenten

Om de tweede onderzoeksvraag te beantwoorden is dezelfde procedure doorlopen als bij de onderzoeksvraag van paragraaf 4.4.1. In totaal hebben tenminste 173 mannelijke en 68 vrouwelijke studenten de vragenlijst beantwoord (twee respondenten hebben hun geslacht niet opgegeven). Enkele numerieke samenvattingen over de twee datasets zijn te vinden in onderstaande tabel.

	Mannen	**Vrouwen**
Gemiddelde somscore	97.53	98.54
Variantie somscore	129.96	121.98
Minimale somscore	63	44
Maximale somscore	129	121

Tabel 4.12: Numerieke samenvattingen m.b.t. somscores van mannelijke en vrouwelijke studenten

Aan de hand van tabel 4.12 kunnen we niet afleiden of mannelijke studenten meer tevreden zijn over hun studie dan vrouwelijke studenten of andersom. Het gemiddelde van de somscores liggen namelijk dicht bij elkaar. Ook de varianties van de somscores verschillen niet veel van elkaar. Op het eerste gezicht lijkt er dus weinig verschil te zitten tussen de tevredenheid van beide geslachten.

Met behulp van de Shapiro-Wilk toets is eerst onderzocht of de somscores van mannelijke en vrouwelijke studenten een normale verdeling hebben. De overschrijdingskans van deze toets is voor mannelijke studenten 0.0039 en voor vrouwelijke studenten is deze 7.013e-06. De overschrijdingkans is in beide gevallen kleiner dan de onbetrouwbaarheidsdrempel van 0.025, waardoor de nulhypothese in beide gevallen verworpen dient te worden. Het is dus aannemelijk dat zowel de somscores van mannelijke studenten als de somscores van vrouwelijke studenten niet normaal verdeeld zijn.

Wederom gaan we middels de Mann Whitney-toets onderzoeken of er een significant verschil is tussen de tevredenheid van beide type studenten. Uit deze toets volgt een overschrijdingskans van 0.4708. Aangezien deze overschrijdingskans veel groter is dan de onbetrouwbaarheidsdrempel van 0.025 is het dus zeer waarschijnlijk dat er geen significant verschil te onderkennen is tussen de tevredenheid van mannelijke en vrouwelijke studenten.

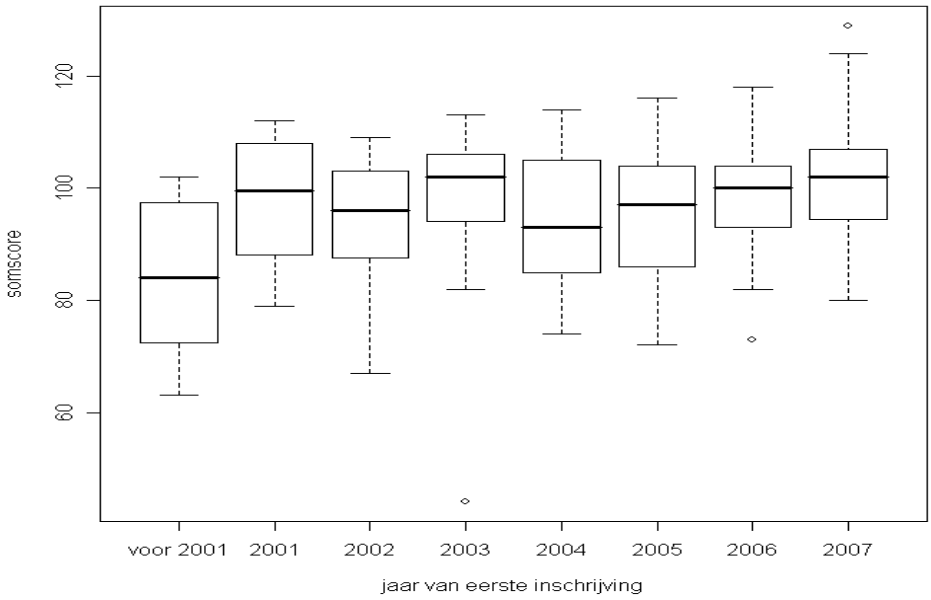
4.4.3 Jaar van eerste inschrijving en tevredenheid

In deze paragraaf zal onderzocht worden of de tevredenheid van FEW studenten gerelateerd is aan het jaar waarop zij zich voor het eerst ingeschreven hebben voor de studie. Hierbij zijn we benieuwd of studenten die al een aantal jaar staan ingeschreven over het algemeen minder tevreden zijn over hun studie dan studenten die relatief kort staan ingeschreven. In tabel 4.13 zijn voor verschillende jaren van eerste inschrijving het aantal respondenten en de gemiddelde somscore van deze studenten gerapporteerd.

Jaar van eerste inschrijving	**<2001**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
Aantal studenten	7	8	20	33	17	30	45	79
Gemiddelde somscore	84.14	97.75	93.9	98.33	95.12	95	98.69	101.27

Tabel 4.13: Per jaar van eerste inschrijving het aantal respondenten en hun gemiddelde somscore

Uit bovenstaande tabel is het niet eenvoudig om conclusies te trekken over de verhouding tussen jaar van eerste inschrijving en tevredenheid. De gemiddelde somscore blijkt voor studenten van 2006 en 2007 weliswaar groter te zijn dan het gemiddelde van studenten die zich voor 2006 hebben ingeschreven. Voor studenten van 2002, 2004 en 2005 geldt daarentegen dat de gemiddelde somscores kleiner zijn dan die van studenten van het jaar daarvoor. Om meer inzicht te krijgen in de verhouding tussen jaar van eerste inschrijving en tevredenheid zijn ook boxplots van de somscores van studenten gemaakt. In onderstaande figuur kan men voor de verschillende jaren van eerste inschrijving een boxplot van de somscores vinden.



Figuur 4.3: Boxplots van de somscores van studenten voor verschillende jaren van eerste inschrijving

Uit de boxplots valt ons op dat een stijgende trend is waar te nemen in de tevredenheid van studenten tussen “voor 2001” en 2003. Hetzelfde geldt voor de tevredenheid van studenten tussen 2004 en 2007. Tussen 2003 en 2004 zien we een daling in de tevredenheid van studenten. Omdat het aantal respondenten per jaar van eerste inschrijving niet bijzonder groot is, leek het ons verstandig om verschillende respondenten samen te voegen. Tabel 4.13 en de boxplots in figuur 4.3 bestuderend heeft ons doen besluiten om de studenten in vier groepen te verdelen. Deze groepen zijn als volgt:

Groep 1: Studenten die zich voor 2001 voor het eerst hebben ingeschreven.
Groep 2: Studenten die zich in 2001 of 2002 voor het eerst hebben ingeschreven.
Groep 3: Studenten die zich tussen 2003 en 2005 voor het eerst hebben ingeschreven.
Groep 4: Studenten die zich in 2006 of 2007 voor het eerst hebben ingeschreven.

Voor alle vier de groepen geldt dat de gemiddelde somscores binnen een groep niet veel van elkaar verschillen. Dit blijkt tevens uit de resultaten van de Mann Whitney-toets op de somscores van studenten binnen een groep. Deze somscores vertonen namelijk geen aanzienlijke discrepantie.

Nu zal onderzocht worden of studenten in een hogere groep significant meer tevreden zijn over hun studie dan studenten in een lagere groep. Dit zal nagegaan worden door de eenzijdige Mann Whitney-toets op de somscores van de verschillende groepen uit te voeren. De nulhypothese en alternatieve hypothese van deze toets worden als volgt geformuleerd:

H_0: De somscores van groep i zijn significant kleiner dan of gelijk aan de somscores van groep j.
H_1: De somscores van groep i zijn significant groter dan de somscores van groep j.
De onbetrouwbaarheidsdrempel wordt gelijk aan 0.05 genomen omdat we te maken hebben met een eenzijdige toets.

Voor alle combinaties tussen de groepen is de eenzijdige Mann Whitney-toets uitgevoerd. De overschrijdingskansen van deze toets zijn in onderstaande tabel terug te vinden.

i	*j*	**Overschrijdingskans**
2	***1***	0.0378
3	***1***	0.0175
3	***2***	0.2338
4	***1***	0.0035
4	***2***	0.0284
4	***3***	0.0464

Tabel 4.14: Overschrijdingskansen van de eenzijdige Mann Whitney-toets op de somscores van de groepen

Uit tabel 4.14 is op te maken dat in één geval de nulhypothese niet verworpen mag worden (i=3 en j=2). Dat betekent dat we niet mogen veronderstellen dat studenten die zich tussen 2003 en 2005 voor het eerst hebben ingeschreven significant meer tevreden zijn over hun studie dan studenten die zich tussen 2001 en 2002 voor het eerst hebben ingeschreven. Voor alle andere gevallen kunnen we aannemen dat studenten die zich in een hogere groep bevinden, en zich dus later hebben ingeschreven, over het algemeen meer tevreden zijn over hun studie dan studenten in een lagere groep. Door het onderscheiden van deze vier groepen is dus enigszins aangetoond dat studenten die relatief lang aan de faculteit staan ingeschreven minder tevreden zijn over hun studie dan studenten die korter staan ingeschreven.

4.4.4 Conclusie

Uit de analyses van deze paragraaf blijkt dat er geen wezenlijk verschil te ontdekken is tussen de algemene tevredenheid van bachelor en master studenten. Hetzelfde geldt voor de algemene tevredenheid van mannelijke en vrouwelijke studenten. Ook uit de gemiddelde somscores blijkt dat verschillende typen studenten over het algemeen nauwelijks van elkaar verschillen wat betreft hun tevredenheid over de studie. Als laatst is onderzocht of studenten, die relatief lang staan ingeschreven, minder tevreden zijn over hun studie dan studenten die

relatief kort staan ingeschreven. Door het onderscheiden van vier groepen studenten en het vergelijken van de somscores van deze groepen is aangetoond dat hier min of meer sprake van is. Om meer zekerheid te krijgen over de relatie tussen jaar van eerste inschrijving en tevredenheid zouden eigenlijk de somscores van studenten van verschillende jaren van eerste inschrijving rechtstreeks met elkaar vergeleken moeten worden. Omdat in ons geval per jaar van eerste inschrijving te weinig data beschikbaar is, hebben we de respondenten in verschillende groepen ingedeeld.
Een andere methode om te onderzoeken of de tevredenheid van FEW studenten afneemt naarmate zij met hun studie vorderen is door de studenten over een langere periode en bijvoorbeeld ieder jaar te onderwerpen aan dit onderzoek. Zo kan onderzocht worden of de tevredenheid van dezelfde studenten over het algemeen afneemt of misschien zelfs toeneemt. Aangezien voor ons onderzoek beperkte tijd beschikbaar was, hebben we de relatie tussen jaar van eerste inschrijving en tevredenheid onderzocht door de tevredenheid van studenten van verschillende periodes van eerste inschrijving met elkaar te vergelijken.

Het aantal respondenten dat voor de andere analyses in deze paragraaf zijn genomen, is mogelijk ook te klein om betrouwbare uitspraken te kunnen doen. In deze analyses zijn namelijk hetzelfde aantal respondenten genomen als bij de item-analyse om de kwaliteit van de vragenlijst te verbeteren (paragraaf 4.3). Normaalgesproken wordt na de kwaliteitsverbetering van een vragenlijst deze opnieuw uitgezet onder een grotere steekproef. De resultaten van de vragenlijst worden vervolgens geanalyseerd aan de hand van de antwoorden van deze steekproef. Door beperkte tijd hebben we ervoor gekozen om dezelfde groep respondenten te gebruiken voor de item-analyse en analyse van de resultaten.

Appendix A Afleiding Cronbach's alpha formule

In deze Appendix wordt stilgestaan bij de totstandkoming van Cronbach's α formule. Deze formule is als volgt:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} s_{Y_i}^2}{s_X^2}\right). \qquad (A.1)$$

Alvorens een begin wordt gemaakt met de afleiding van formule (A.1) zullen ter verduidelijking eerst de variabelen, die we in het vervolg gebruiken, gedefinieerd worden:

Notatie

s_x^2	(Steekproef)variantie van de stochastische variabele X. $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$.
s_x	(Steekproef)standaarddeviatie van de stochastische variabele X.
s_{xy}	(Steekproef)covariantie tussen de stochastische variabelen X en Y. $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$.
r_{xy}	Correlatiecoëfficiënt tussen de stochastische variabelen X en Y. $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$.
K	Aantal items in de vragenlijst.
n	Aantal respondenten, die de vragenlijst hebben ingevuld.
X	Waarnemingsscore op een vragenlijst (Somscore).
Y_i	Waarnemingsscore op item i, met $1 \leq i \leq K$.
T_i	Ware score (true score) op item i, met $1 \leq i \leq K$.
ε_i	Meetfout (error score) op item i, met $1 \leq i \leq K$.
α	Cronbach's alpha.

De betrouwbaarheid van een vragenlijst kan, zoals in hoofdstuk 3 aan bod is gekomen, worden opgevat als de mate van overeenstemming tussen scores op deze vragenlijst en de scores op een parallelle vragenlijst. De parallelle vragenlijst wordt geacht hetzelfde te meten en dat betekent dus dat personen op deze vragenlijst dezelfde scores zouden moeten behalen als op de oorspronkelijke vragenlijst. De mate van overeenstemming kan worden uitgedrukt aan de hand van de correlatie tussen de resultaten op beide vragenlijsten.

De formule voor de correlatiecoëfficiënt wordt eenvoudiger als de gemiddelden op beide te correleren variabelen 0 zouden zijn. Omdat we aannemen dat een Likertschaal als intervalschaal te beschouwen is (zie paragraaf 2.3), is het toegestaan om een lineaire transformatie op een Likertschaal toe te passen. In [12] staat beschreven dat deze transformatie inhoudt dat alle geobserveerde scores verminderd mogen worden met hun gemiddelde. Het gevolg hiervan is dat we afwijkingscores verkrijgen met 0 als gemiddelde.

Om de formule voor de betrouwbaarheid af te leiden gaan we eerst uit van twee parallelle vragenlijsten, die ieder uit slechts één item bestaat. De vragenlijsten zijn door dezelfde *n* respondenten ingevuld. De afwijkingsscore van respondent *i* op vragenlijst 1 en vragenlijst 2 drukken we uit in respectievelijk X_{1i} en X_{2i}. De betrouwbaarheid als correlatie tussen beide vragenlijsten is nu als volgt uit te drukken:

$$r_{x_1x_2} = \frac{\sum_{i=1}^{n} X_{1i}X_{2i}}{(n-1)s_{x_1}s_{x_2}} = \frac{\sum_{i=1}^{n}(T_{1i}+\varepsilon_{1i})(T_{2i}+\varepsilon_{2i})}{(n-1)s_{x_1}s_{x_2}} = \frac{\sum_{i=1}^{n}(T_{1i}T_{2i}+T_{1i}\varepsilon_{2i}+T_{2i}\varepsilon_{1i}+\varepsilon_{1i}\varepsilon_{2i})}{(n-1)s_{x_1}s_{x_2}}$$

$$= \frac{\sum_{i=1}^{n}T_{1i}T_{2i}}{(n-1)s_{x_1}s_{x_2}} + \frac{\sum_{i=1}^{n}T_{1i}\varepsilon_{2i}}{(n-1)s_{x_1}s_{x_2}} + \frac{\sum_{i=1}^{n}T_{2i}\varepsilon_{1i}}{(n-1)s_{x_1}s_{x_2}} + \frac{\sum_{i=1}^{n}\varepsilon_{1i}\varepsilon_{2i}}{(n-1)s_{x_1}s_{x_2}}. \qquad (A.2)$$

In de klassieke testtheorie worden bepaalde aannamen gedaan, wat tot gevolg heeft dat bovenstaande uitdrukking voor de betrouwbaarheid beduidend eenvoudiger wordt. Deze aannamen zijn onder andere:

- $T_{1i} = T_{2i} \quad \forall i = 1,...,n$ (*de ware scores van een willekeurige respondent op parallelle vragenlijsten zijn gelijk).*
- $s_{x_1} = s_{x_2}$ (*de standaardafwijkingen van de geobserveerde scores op parallelle vragenlijsten zijn gelijk).*
- *De meetfouten en ware scores zijn onafhankelijk van elkaar. Covariantie tussen ware score en meetfout is dus gelijk aan 0.*
- *De meetfouten zijn onderling onafhankelijk verdeeld. Covariantie tussen verschillende meetfouten is ook gelijk aan 0.*

Als we de ware score van een respondent *i* op vragenlijst 1 als op vragenlijst 2 uitdrukken als T_i en de standaarddeviatie van de geobserveerde scores van de respondenten op de vragenlijsten aanduiden met s_x is bovenstaande formule voor $r_{x_1x_2}$ te herschrijven naar:

$$r_{x_1x_2} = \frac{\sum_{i=1}^{n} T_iT_i}{(n-1)s_xs_x} = \frac{s_T^2}{s_x^2}. \qquad (A.3)$$

De betrouwbaarheid is dus uit te drukken als het quotiënt van de spreiding in de ware scores (s_T^2) en de spreiding in de geobserveerde scores (s_x^2). We zien dat de eerder gedane lineaire transformatie geen invloed heeft op vergelijking (A.3). Varianties veranderen namelijk niet bij een lineaire transformatie.

Vragenlijst met 2 items

We gaan nu kijken wat er gebeurt met de betrouwbaarheid als de vragenlijst tweemaal zo lang wordt gemaakt. Een twee maal zo lange vragenlijst verkrijgt men door de parallelle vragenlijsten van hiervoor samen te voegen. Om een uitdrukking voor de betrouwbaarheid te krijgen, berekenen we zowel de spreiding in de geobserveerde scores van een tweemaal zo lange vragenlijst (s_{2x}^2) als de spreiding in ware scores van een dergelijke vragenlijst (s_{2T}^2):

$$s_{2x}^2 = s_{x1+x2}^2 = s_{x1}^2 + s_{x2}^2 + 2s_{x1x2} = s_{x1}^2 + s_{x2}^2 + 2r_{x1x2}s_{x1}s_{x2} = 2s_x^2(1+r_{x1x2}). \quad \text{(A.4)}$$

$$s_{2T}^2 = s_{T1+T2}^2 = s_{T1}^2 + s_{T2}^2 + 2r_{T1T2}s_{T1}s_{T2} = 4s_T^2. \quad \text{(A.5)}$$

Bij de berekening van s_{2x}^2 is wederom de aanname gedaan dat $s_{x_1} = s_{x_2}$ en bij de berekening van s_{2T}^2 is verondersteld dat $s_{T1} = s_{T2}$ en dat de ware scores van de parallelle vragenlijsten maximaal gecorreleerd zijn ($r_{T1T2} = 1$).

Uit bovenstaande resultaten blijkt dat indien $r_{x1x2} < 1$ een verdubbeling van het aantal items leidt tot een grotere stijging in de variantie van de ware scores dan in de variantie van de geobserveerde scores. Hieruit volgt dat de betrouwbaarheid van een tweemaal zo grote vragenlijst (r_{2x}) groter wordt. De betrouwbaarheid van een twee maal zo grote vragenlijst is als volgt uit te drukken:

$$r_{2x} = \frac{s_{2T}^2}{s_{2x}^2} = \frac{4s_T^2}{2s_x^2(1+r_{x1x2})} = 2 \cdot \frac{s_T^2}{s_x^2} \cdot \frac{1}{1+r_{x1x2}} = \frac{2r_{x1x2}}{1+r_{x1x2}}. \quad \text{(A.6)}$$

N.B. Vergelijking (A.6) komt overeen met de *Spearman-Brown split half betrouwbaarheids-coëfficiënt* van een vragenlijst met twee items.

Vragenlijst met K items

Eigenlijk zijn we geïnteresseerd in wat er gebeurt als de vragenlijst willekeurig groot wordt gemaakt. We kunnen dit analyseren door een willekeurig aantal parallelle vragenlijsten, die hetzelfde construct meten, met elkaar te combineren. Stel dat we een vragenlijst *K* maal zo groot maken, dan is de variantie in de geobserveerde scores als volgt:

$$s_{Kx}^2 = s_{x1+x2+\ldots+xK}^2 \overset{(i)}{=} \sum_{i=1}^{K} s_{xi}^2 + \sum_{i,j:i\neq j}^{K} s_{xixj} \overset{(ii)}{=} Ks_x^2 + \sum_{i,j:i\neq j}^{K} r_{xixj}s_{xi}s_{xj} \overset{(iii)}{=} Ks_x^2 + K(K-1)r_{xx}s_x^2. \quad \text{(A.7)}$$

Deze afleiding vergt enig uitleg en dat zal hieronder stapsgewijs worden gedaan:

(i) De variantie van de som van een verzameling afhankelijke stochastische variabelen is gelijk aan de som van de elementen in de covariantiematrix van deze variabelen.

(ii) $s_{x_1} = s_{x_2} = \ldots = s_{xK} = s_x$.

(iii) Hier is aangenomen dat de betrouwbaarheid voor elke subvragenlijst van gelijke lengte (vragenlijst bestaande uit één item) even groot is. Dat betekent dat de correlatie tussen de verschillende geobserveerde scores (r_{xixj}) gelijk zijn aan elkaar en daarom te schrijven als r_{xx}.

Door deze aanname en de aanname van *(ii)* zien we dat de sommatie $\sum_{i,j:i\neq j} r_{xixj}s_{xi}s_{xj}$ naar $K(K-1)r_{xx}s_x^2$ hergeschreven kan worden, mede omdat de *K* items ieder met (*K*-1) andere items verbonden kunnen worden.

De variantie in de ware scores is bij een *K* maal zo grote vragenlijst simpelweg:

$$s_{KT}^2 = K^2 s_T^2. \qquad \text{(A.8)}$$

Vergelijking (A.8) volgt uit de veronderstelling dat de ware scores op parallelle vragenlijsten gelijk zijn aan elkaar.

Gebruikmakend van (A.7) en (A.8) is de betrouwbaarheid van een vragenlijst met *K* items (r_{Kx}) in formule vorm te schrijven als:

$$r_{Kx} = \frac{s_{KT}^2}{s_{Kx}^2} = \frac{K^2 s_T^2}{K s_x^2 + K(K-1) r_{xx} s_x^2} = \frac{K s_T^2}{s_x^2 + (K-1) r_{xx} s_x^2} = \frac{K}{1+(K-1) r_{xx}} \cdot \frac{s_T^2}{s_x^2} = \frac{K r_{xx}}{1+(K-1) r_{xx}}. \qquad \text{(A.9)}$$

Hierbij is het resultaat dat $r_{xx} = \frac{s_T^2}{s_x^2}$ gebruikt.

Om de betrouwbaarheid van een vragenlijst bestaande uit *K* items te schatten kan vergelijking (A.9) toegepast worden. Alleen is niet bekend hoe de correlatie tussen de geobserveerde itemscores (r_{xx}) berekend moet worden in de praktijk. Er is namelijk verondersteld dat de correlatie van een item met een parallel item gelijk is voor alle items. In de praktijk zijn de correlaties tussen de verschillende itemscores meestal niet gelijk aan elkaar en dus zal vergelijking (A.9) herschreven moeten worden om de betrouwbaarheid van een vragenlijst praktisch te kunnen uitrekenen. Aangezien r_{xx} de correlatie van een item met een parallel item voorstelt, kan deze waarde opgevat worden als de betrouwbaarheid van een item.

Cronbach's alpha

We zullen nu uitgaan van een enkele vragenlijst met *K* items. Dat betekent, volgens de notatie hierboven, dat we de itemscore van een bepaald item *i* nu met Y_i uitdrukken. De correlatie tussen de items *i* en *j* (r_{YiYj}) zijn verondersteld identiek te zijn voor alle itemparen in de vragenlijst. r_{YiYj} lossen we op uit de formule voor de variantie van de som van de *K* geobserveerde itemscores. Deze variantie is dus eigenlijk de variantie van de waarnemingscore *X* op een hele vragenlijst en is hieronder uitgewerkt:

$$s_X^2 = \sum_{i=1}^{K} s_{Yi}^2 + \sum_{i,j:i\neq j}^{K} r_{YiYj} s_{Yi} s_{Yj} = \sum_{i=1}^{K} s_{Yi}^2 + (K-1) r_{YiYj} \sum_{i=1}^{K} s_{Yi}^2. \qquad \text{(A.10)}$$

In deze afleiding hebben we wederom de aanname gedaan dat $s_{Yi} = s_{Yj}$. Tevens zijn r_{YiYj} voor $i = j = 1 \ldots K$ identiek aan elkaar en zodoende kan r_{YiYj} buiten het tweede sommatieteken gehaald worden.

Met behulp van (A.10) kunnen we een uitdrukking voor r_{YiYj} krijgen, namelijk:

$$r_{YiYj} = \frac{s_X^2 - \sum_{i=1}^{K} s_{Yi}^2}{(K-1) \sum_{i=1}^{K} s_{Yi}^2}. \qquad \text{(A.11)}$$

Omdat is aangenomen dat $r_{Y_iY_j}$ identiek zijn voor alle itemparen in de vragenlijst kunnen we deze ook schrijven als $r_{Y_iY_i}$. $r_{Y_iY_i}$ kan gezien worden als de betrouwbaarheid van één item. Nu we een praktische uitdrukking voor $r_{Y_iY_i}$ hebben, kunnen we gebruik maken van (A.9) om de betrouwbaarheid van een vragenlijst met *K* items af te leiden:

$$r_{kY_i} = \frac{Kr_{Y_iY_i}}{1+(K-1)r_{Y_iY_i}} = \frac{K\cdot\left(\frac{s_X^2-\sum_{i=1}^{K}s_{Y_i}^2}{(K-1)\sum_{i=1}^{K}s_{Y_i}^2}\right)}{1+(K-1)\left(\frac{s_X^2-\sum_{i=1}^{K}s_{Y_i}^2}{(K-1)\sum_{i=1}^{K}s_{Y_i}^2}\right)} = \frac{K\cdot\left(\frac{s_X^2-\sum_{i=1}^{K}s_{Y_i}^2}{(K-1)\sum_{i=1}^{K}s_{Y_i}^2}\right)}{\frac{\sum_{i=1}^{K}s_{Y_i}^2}{\sum_{i=1}^{K}s_{Y_i}^2}+\frac{s_X^2-\sum_{i=1}^{K}s_{Y_i}^2}{\sum_{i=1}^{K}s_{Y_i}^2}} = \frac{K\cdot\left(\frac{s_X^2-\sum_{i=1}^{K}s_{Y_i}^2}{(K-1)\sum_{i=1}^{K}s_{Y_i}^2}\right)}{\frac{s_X^2}{\sum_{i=1}^{K}s_{Y_i}^2}}$$

$$= \frac{K}{K-1}\cdot\frac{s_X^2-\sum_{i=1}^{K}s_{Y_i}^2}{\sum_{i=1}^{K}s_{Y_i}^2}\cdot\frac{\sum_{i=1}^{K}s_{Y_i}^2}{s_X^2} = \frac{K}{K-1}\cdot\left(\frac{s_X^2-\sum_{i=1}^{K}s_{Y_i}^2}{s_X^2}\right) = \frac{K}{K-1}\cdot\left(1-\frac{\sum_{i=1}^{K}s_{Y_i}^2}{s_X^2}\right) = \alpha. \qquad \text{(A.12)}$$

De formule van de betrouwbaarheid van een vragenlijst bestaande uit *K* items resulteert dus in de formule van Cronbach‟s α.

Appendix B Vragenlijst

Student satisfaction survey 2008
Answers marked with a * are required.

1. Introduction

Dear students,

The purpose of this questionnaire is to measure the satisfaction of students of the faculty of sciences (FEW) about their current study programme.
For the opinion questions please select the answer that best describes your opinion.

Thanks for your cooperation!

2. Personal data

1. What is your gender?

○ Male ○ Female

2. What is your current age?

○ 17 - 19 ○ 20 - 25 ○ 26 - 30 ○ over 30

3. What type of study programme are you following?

○ Bachelor ○ Master ○ Both

4. What programme are you studying?

5. What year of your programme are you currently in?

○ 1st ○ 2nd ○ 3rd ○ 4th ○ 5th

6. What is the year of your first registration?

3. Questionnaire

The questions in this questionnaire are about the following topics:

- Content of your study programme
- Quality of lecturers
- Study load
- Facilities of the faculty of sciences (Student Services Office/ Onderwijsbureau, Library, Computing, Printing)

Please read each question carefully and select the answer that best describes your opinion.

7. Content of your study programme. Please rate your level of agreement with the following statements. The answers you can choose from are: disagree strongly, disagree, tend to disagree, tend to agree, agree and agree strongly *

Item nr.		*disagree strongly*	*disagree*	*tend to disagree*	*tend to agree*	*agree*	*agree strongly*
1	The courses of my study programme are of good quality						
2	Generally speaking the courses of my programme are sufficiently interesting						
3	The study material (books, readers,						

	etc.) is clear						
4	The assignments/ projects in my programme are useful						
5	The study material (books, readers, etc.) is informative						
6	Some courses are a waste of time						
7	The assignments/ projects in my programme are worthwhile						

8. Quality of lecturers. Please rate your level of agreement with the following statements: *

Item nr.		*disagree strongly*	*disagree*	*tend to disagree*	*tend to agree*	*agree*	*agree strongly*
8	In general, the didactic skills of the lecturers are of good quality						
9	In general, the lecture material is explained clearly						
10	The lecturers are in general able to encourage me						
11	In general, the course material is adequately spread over the lectures						
12	The lecturers are in general easily accessible (mail, room)						

9. Study load. Please rate your level of agreement with the following statements: *

Item nr.		*disagree strongly*	*disagree*	*tend to disagree*	*tend to agree*	*agree*	*agree strongly*
13	The total study load of the courses is in proportion to the number of credits allocated.						
14	Next to my study I keep enough time to do other things						
15	In general, the courses are well spread over the year						
16	I attend the lectures regularly						

10. Facilities of the faculty of sciences (Student Services Office/ Onderwijsbureau, Library, Computing, Printing). Please rate your level of agreement with the following statements: *

Item nr.		*disagree strongly*	*disagree*	*tend to disagree*	*tend to agree*	*agree*	*agree strongly*
17	The Student Services Office is easily accessible						
18	The information provision of the Student Services Office is adequate						
19	In the faculty there are sufficient computers available						
20	The computing facilities are good						
21	The helpdesk is easily accessible						
22	The knowledge of the helpdesk staff is adequate						
23	The printing facilities are good						
24	The library of the faculty of sciences is easily accessible						
25	The faculty provides sufficient places for self-study						

11. General questions. Please rate your level of agreement with the following statements: *

Item nr.		*disagree strongly*	*disagree*	*tend to disagree*	*tend to agree*	*agree*	*agree strongly*
26	In general, I am satisfied with my study						
27	I would recommend my study to prospective students						
28	After my study there are sufficient career opportunities						

Appendix C R functies

Deze Appendix bevat de R-code van de functies, die voor de analyses in hoofdstuk 4 zijn geprogrammeerd.

Cronbach's α

```
cronbach2 <- function(X) {
K = ncol(X)
S = cov(X)
j = matrix(c(1),nrow=K,ncol=1)
a = K/(K-1)* (1-(sum(diag(S))/(t(j)%*%S%*%j)))

a
}
```

Betrouwbaarheidsinterval Cronbach's α

```
alphaCI <- function(X,interval) {

K = ncol(X)
n = nrow(X)

S = cov(X)
j = matrix(c(1),nrow=K,ncol=1)

Q = ((K/(K-1))^2) * 2/((t(j)%*%S%*%j)^3)
*((t(j)%*%S%*%j)*(sum(diag(S%*%S))+(sum(diag(S)))^2) -
2*(sum(diag(S))*(t(j)%*%S%*%j)))

sterror = sqrt(Q/n)

a = cronbach2(X)
b = qnorm(interval+(1-interval)/2, mean=0, sd=1)

l = a - b*sterror
u = a + b*sterror

c(l,u)
}
```

Spearman-Brown split half betrouwbaarheidscoëfficiënt

Berekend a.d.h.v. twee meegegeven subschalen

```
splithalf <- function(subscale1,subscale2) {

sum1 <- rowSums(subscale1)
sum2 <- rowSums(subscale2)

rxy <- cor(sum1,sum2)
rho <- 2*rxy/(1+rxy)

rho
}
```

Berekend a.d.h.v. twee subschalen, die verkregen worden door een schaal <u>systematisch</u> (om en om) te splitsen.

```
splithalfSyst <- function(scale) {
K = ncol(scale)

sc1 <- c()
sc2 <- c()

if (K%%2 == 0) {
half = K/2

for (i in 1:half) {
sc1 <- cbind(sc1,scale[,((2*i)-1)])
}
for (i in 1:half) {
sc2 <- cbind(sc2,scale[,(2*i)])
}
}

else {
half = floor(K/2)

for (i in 1:(half+1)) {
sc1 <- cbind(sc1,scale[,((2*i)-1)])
}
for (i in 1:half) {
sc2 <- cbind(sc2,scale[,(2*i)])
}
}

splithalf(sc1,sc2)
}
```

Berekend a.d.h.v. twee subschalen, die verkregen worden door een schaal <u>random</u> te splitsen.

```
Random splitsing van een schaal in ongeveer twee even grote subschalen

splithalfRandom <- function(scale) {
K = ncol(scale)
half = round(K/2)

random <- sample(1:K)
sc1 <- c()
sc2 <- c()

for (i in random[1:half]) {
sc1 <- cbind(sc1,scale[,i])
}

for (i in random[(half+1):K]) {
sc2 <- cbind(sc2,scale[,i])
}

splithalf(sc1,sc2)
}
```

Item-totaal correlaties met overschrijdingskans eenzijdige correlatietoets

```
itemTotaalCor2 <- function(scale) {
K = ncol(scale)

totaal <- rowSums(scale)
correlaties <- matrix(,nrow=2,ncol=K)

for (i in 1:K) {
correlaties[1,i] <-cor(scale[,i],totaal)
correlaties[2,i] <- cor.test(scale[,i],totaal, alternative="g")$p.value
}

correlaties
}
```

Item-rest correlaties met overschrijdingskans eenzijdige correlatietoets

```
itemRestCor2 <- function(scale) {
K = ncol(scale)

correlaties <- matrix(,nrow=2,ncol=K)

for (i in 1:K) {
rest <- rowSums(scale)-scale[,i]
correlaties[1,i] <- cor(scale[,i],rest)
correlaties[2,i] <- cor.test(scale[,i],rest, alternative="g")$p.value
}

correlaties
}
```

Alpha indien item verwijderd

N.B. Een schaal heeft tenminste 2 items. Dus om alpha indien item verwijderd te kunnen toepassen dient een schaal minimaal 3 items te hebben

```
aIfItemDeleted <- function(scale) {
K = ncol(scale)

if (K < 3) stop("Een schaal moet tenminste uit 3 items bestaan om alpha
indien item verwijderd te kunnen toepassen")

aItemDeleted <- c()

newscale <- scale[,2:K]
aItemDeleted[1] <- cronbach2(newscale)

for (i in 2:(K-1)) {
newscale <- cbind(scale[,1:(i-1)], scale[,(i+1):K])
aItemDeleted[i] <- cronbach2(newscale)
}

newscale <- scale[,1:(K-1)]
aItemDeleted[K] <- cronbach2(newscale)

aItemDeleted
}
```

Appendix D Uitgevoerde R-code

Deze Appendix bevat de code, die in R is gegenereerd om de analyses van hoofdstuk 4 uit te voeren.

Hoofdstuk 4 Praktijkstudie: Likertschaal vragenlijst construeren, uitzetten en analyseren

4.3 Analyse van de vragenlijst

Antwoorden inlezen in R

```
> survey <- read.table("C:/Documents and Settings/Radouan/Bureaublad/BWI
werkstuk/RESULTATEN student satisfaction survey/surveyresponses.csv", sep = ";")

> scale1 <- matrix(, nrow(survey), 28)
> for (j in 1:(ncol(survey)-6)) {
+ scale1[,j] <- survey[,j+6]
+ }

scale2 <- scale1
```

Item 6 hercoderen

```
> for (i in 1:nrow(scale2)) {
+ scale2[i,6] <- 7 - scale2[i,6]
+ }

> totaal <- rowSums(scale2)
> hist(totaal,breaks=20,xlab="somscore",main = "Histogram van somscore")
> boxplot(totaal,ylab = "somscore",main="Boxplot van somscore")
```

Respondent 7 verwijderd

```
> scale2b <- rbind(scale2[1:6,],scale2[8:244,])
```

Betrouwbaarheid oorspronkelijke vragenlijst (28 items)

```
> splithalfSyst(scale2b)
> splithalfRandom(scale2b)

> splithalfrand <- c()
> for (i in 1:1000) {
+ splithalfrand[i] <- splithalfRandom(scale2b)
+}
> mean(splithalfrand)

> cronbach2(scale2b)
> alphaCI(scale2b, 0.95)
```

Item analyse oorspronkelijke vragenlijst

```
> itemTotaalCor2(scale2b)
> itemRestCor2(scale2b)
> cor(scale2b)

> cor.test(scale2b[,12],scale2b[,24],alternative="g")$p.value
> cor.test(scale2b[,13],scale2b[,19],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,1],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,2],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,5],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,6],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,8],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,9],alternative="g")$p.value
```

```
> cor.test(scale2b[,14],scale2b[,10],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,16],alternative="g")$p.value
> cor.test(scale2b[,14],scale2b[,24],alternative="g")$p.value
> cor.test(scale2b[,16],scale2b[,17],alternative="g")$p.value
> cor.test(scale2b[,16],scale2b[,21],alternative="g")$p.value
> cor.test(scale2b[,16],scale2b[,22],alternative="g")$p.value
> cor.test(scale2b[,16],scale2b[,23],alternative="g")$p.value
> cor.test(scale2b[,16],scale2b[,24],alternative="g")$p.value
> cor.test(scale2b[,16],scale2b[,28],alternative="g")$p.value
```

alpha indien item verwijderd oorspronkelijke vragenlijst (28 items)

```
> aIfItemDeleted(scale2b)
```

Item 14 verwijderd

```
> scale3b <-cbind(scale2b[,1:13],scale2b[,15:28])
> cronbach2(scale3b)
> alphaCI(scale3b,0.95)

> aIfItemDeleted(scale3b)
```

Item 24 verwijderd

```
> scale4b <- cbind(scale2b[,1:13],scale2b[,15:23],scale2b[,25:28])
> cronbach2(scale4b)
> alphaCI(scale4b,0.95)

> aIfItemDeleted(scale4b)
```

Item 25 verwijderd

```
> scale5b <- cbind(scale2b[,1:13],scale2b[,15:23],scale2b[,26:28])

> cronbach2(scale5b)
> alphaCI(scale5b,0.95)

> aIfItemDeleted(scale5b)
> a <- aIfItemDeleted(scale5b)
> max(a)
```

Item 16 verwijderd

```
> scale6b <- cbind(scale2b[,1:13],scale2b[,15],scale2b[,17:23],scale2b[,26:28])

> cronbach2(scale6b)
> alphaCI(scale6b,0.95)

> aIfItemDeleted(scale6b)
> a <- aIfItemDeleted(scale6b)
> max(a)
```

Item 19 verwijderd

```
> scale7b <- cbind(scale2b[,1:13],scale2b[,15],scale2b[,17:18],
scale2b[,20:23],scale2b[,26:28])

> cronbach2(scale7b)
> alphaCI(scale7b,0.95)

> aIfItemDeleted(scale7b)
> a <- aIfItemDeleted(scale7b)
> max(a)
```

Item 6 verwijderd

```
> scale8b <- cbind(scale2b[,1:5],scale2b[,7:13],scale2b[,15],scale2b[,17:18],
scale2b[,20:23],scale2b[,26:28])
```

```
> cronbach2(scale8b)
> alphaCI(scale8b,0.95)

> aIfItemDeleted(scale8b)
> a <- aIfItemDeleted(scale8b)
> max(a)
```

Statistieken definitieve vragenlijst (22 items)

```
> splithalfSyst(scale2b)

> splithalfrand2 <- c()
> for (i in 1:1000) {
+ splithalfrand2[i] <- splithalfRandom(scale8b)
+ }
> mean(splithalfrand2)
```

4.4 Analyse van de resultaten

Oorspronkelijke kolommen (eerste 6 vragen) aan de itemscores toevoegen:

```
gen <- cbind(survey[,1],survey[,2],survey[,3],survey[,4],survey[,5], survey[,6])
```

respondent 7 verwijderd

```
genQuestions <- rbind(gen[1:6,],gen[8:244,])

totalSurvey <- cbind(genQuestions,scale8b)
```

numerieke samenvattingen met betrekking tot de somscores van de respondenten

```
> rowSum <- rowSums(scale8b)
> mean(rowSum)
> var(rowSum)
> max(rowSum)
> min(rowSum)
```

numerieke samenvattingen met betrekking tot de itemscores van de respondenten

```
> colMeans(scale8b)

> variances <- c()
> for (i in 1:ncol(scale8b)) {
+ variances[i] <- var(scale8b[,i])
+ }

> variances
```

- Is er een significant verschil tussen de tevredenheid van bachelor en master studenten?

Verwijderen van respondenten van wie we niet weten of ze en bachelor of master studie volgen

```
> totalSurvey1 <- rbind(totalSurvey[1:3,],totalSurvey[5:98,],totalSurvey[100:243,])
```

De antwoorden van bachelor en master studenten in aparte matrices stoppen

```
bachelor <- matrix(,sum(totalSurvey1[,3]==1),ncol(totalSurvey1))
aantal <- 1
for(i in 1: nrow(totalSurvey1)) {
if(totalSurvey1[i,3]==1) {
bachelor[aantal,] <- totalSurvey1[i,]
aantal <- aantal+1
}
```

```
}

master <- matrix(,sum(totalSurvey1[,3]==2),ncol(totalSurvey1))
aantal <- 1
for(i in 1: nrow(totalSurvey1)) {
if(totalSurvey1[i,3]==2) {
master[aantal,] <- totalSurvey1[i,]
aantal <- aantal+1
}
}

> nrow(bachelor)
> nrow(master)

> somscoreBachelor <- rowSums(bachelor[,7:28])
> somscoreMaster <- rowSums(master[,7:28])

> mean(somscoreBachelor)
> mean(somscoreMaster)
> var(somscoreBachelor)
> var(somscoreMaster)
> min(somscoreBachelor)
> min(somscoreMaster)
> max(somscoreBachelor)
> max(somscoreMaster)
```

Checken of data normaal verdeeld is

```
> shapiro.test(somscoreBachelor)
> shapiro.test(somscoreMaster)
```

Uitvoeren van Mann Whitney-toets (Wilcoxon-twee-steekproeventoets)

```
> wilcox.test(somscoreBachelor, somscoreMaster)
```

- Is er een significant verschil tussen de tevredenheid van mannelijke en vrouwelijke studenten?

Analoog aan de vorige onderzoeksvraag

```
mannen <- matrix(,sum(totalSurvey1[,1]==1),ncol(totalSurvey1))
aantal <- 1
for(i in 1: nrow(totalSurvey1)) {
if(totalSurvey1[i,1]==1) {
mannen[aantal,] <- totalSurvey1[i,]
aantal <- aantal+1
}
}

vrouwen <- matrix(,sum(totalSurvey1[,1]==2),ncol(totalSurvey1))
aantal <- 1
for(i in 1: nrow(totalSurvey1)) {
if(totalSurvey1[i,1]==2) {
vrouwen[aantal,] <- totalSurvey1[i,]
aantal <- aantal+1
}
}

> nrow(mannen)
> nrow(vrouwen)

> somscoreMannen <- rowSums(mannen[,7:28])
> somscoreVrouwen <- rowSums(vrouwen[,7:28])

> mean(somscoreMannen)
> mean(somscoreVrouwen)
> var(somscoreMannen)
```

```
> var(somscoreVrouwen)
> min(somscoreMannen)
> min(somscoreVrouwen)
> max(somscoreMannen)
> max(somscoreVrouwen)

> shapiro.test(somscoreMannen)
> shapiro.test(somscoreVrouwen)

> wilcox.test(somscoreMannen, somscoreVrouwen)
```

- Zijn studenten, die relatief lang staan ingeschreven, minder tevreden over hun studie dan studenten die relatief kort staan ingeschreven?

Verwijderen van respondenten van wie we niet weten wat hun eerste jaar van inschrijving is

```
> totalSurvey2 <- rbind(totalSurvey[1:3,],totalSurvey[5:98,],totalSurvey[101:243,])

> sum(totalSurvey2[,6]<2000)
> sum(totalSurvey2[,6]==2000)
> sum(totalSurvey2[,6]==2001)
> sum(totalSurvey2[,6]==2002)
> sum(totalSurvey2[,6]==2003)
> sum(totalSurvey2[,6]==2004)
> sum(totalSurvey2[,6]==2005)
> sum(totalSurvey2[,6]==2006)
> sum(totalSurvey2[,6]==2007)
> sum(totalSurvey2[,6]>2007)

stVoor2001 <- matrix(,sum(totalSurvey2[,6]<2001),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]<2001) {
stVoor2001[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
}
somscoreStVoor2001 <- rowSums(stVoor2001[,7:28])

st2001 <- matrix(,sum(totalSurvey2[,6]==2001),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]==2001) {
st2001[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
}
somscoreSt2001 <- rowSums(st2001[,7:28])

st2002 <- matrix(,sum(totalSurvey2[,6]==2002),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]==2002) {
st2002[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
}
somscoreSt2002 <- rowSums(st2002[,7:28])

st2003 <- matrix(,sum(totalSurvey2[,6]==2003),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]==2003) {
st2003[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
```

```
}
somscoreSt2003 <- rowSums(st2003[,7:28])

st2004 <- matrix(,sum(totalSurvey2[,6]==2004),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]==2004) {
st2004[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
}
somscoreSt2004 <- rowSums(st2004[,7:28])

st2005 <- matrix(,sum(totalSurvey2[,6]==2005),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]==2005) {
st2005[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
}
somscoreSt2005 <- rowSums(st2005[,7:28])

st2006 <- matrix(,sum(totalSurvey2[,6]==2006),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]==2006) {
st2006[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
}
somscoreSt2006 <- rowSums(st2006[,7:28])

st2007 <- matrix(,sum(totalSurvey2[,6]==2007),ncol(totalSurvey2))
aantal <- 1
for(i in 1: nrow(totalSurvey2)) {
if(totalSurvey2[i,6]==2007) {
st2007[aantal,] <- totalSurvey2[i,]
aantal <- aantal+1
}
}
somscoreSt2007 <- rowSums(st2007[,7:28])

> mean(somscoreStVoor2001)
> mean(somscoreSt2001)
> mean(somscoreSt2002)
> mean(somscoreSt2003)
> mean(somscoreSt2004)
> mean(somscoreSt2005)
> mean(somscoreSt2006)
> mean(somscoreSt2007)

boxplot(somscoreStVoor2001, somscoreSt2001, somscoreSt2002, somscoreSt2003,
somscoreSt2004, somscoreSt2005, somscoreSt2006, somscoreSt2007, names=c("voor
2001", "2001", "2002", "2003", "2004", "2005", "2006", "2007"), xlab="jaar van
eerste inschrijving",ylab = "somscore")

groep1 <- somscoreStVoor2001
groep2 <- c(somscoreSt2001, somscoreSt2002)
groep3 <- c(somscoreSt2003, somscoreSt2004, somscoreSt2005)
groep4 <- c(somscoreSt2006, somscoreSt2007)

> wilcox.test(groep2, groep1, alternative="g")$p.value
> wilcox.test(groep3, groep2, alternative="g")$p.value
> wilcox.test(groep3, groep1, alternative="g")$p.value
> wilcox.test(groep4, groep1, alternative="g")$p.value
> wilcox.test(groep4, groep2, alternative="g")$p.value
> wilcox.test(groep4, groep3, alternative="g")$p.value
```

Bibliografie

[1] A. Fink. *The survey handbook*. SAGE Publications (1995).

[2] D.A. Dillman: *Mail and Internet Surveys: The Tailored Design Method (2nd ed.).* John Wiley and Sons (2007).

[3] C. Passmore, A.E. Dobbie, M. Parchman, J. Tysinger (2002), *Guidelines for Constructing a survey*. Research Series Vol. 34 No. 4.
http://www.stfm.org/fmhub/fm2002/apr02/rs1.pdf

[4] Constructing a questionnaire.
http://www.ne.jp/asahi/macgregor/classes/Resources/questionnaire.html

[5] *Eight steps in constructing questionnaire*, (Adapted from workshop materials on constructing questionnaires by Blaine R. Worthen and Vanessa D. Moss-Summers, The Evaluation Institute, July 26-27, 2000.
http://www.apssa.uiuc.edu/content/conducting_surveys/eightsteps.html

[6] W.M.K. Trochim, *Survey research*. Webcenter for social research methods.
http://www.socialresearchmethods.net/kb/survey.php

[7] Prof. Dr. P. Theuns, *Schaleren*. Aanvulling bij persoonlijke klasnota's Vrije Universiteit Brussel Faculteit voor Psychologie en Opvoedkunde.
homepages.vub.ac.be/~ptheuns/Schaleren.doc

[8] Spector, P.E.: *Summated rating scale construction: an introduction.* SAGE publications (1992).

[9] Dr. J. van Dalen, *College sheets Statistische methoden en technieken*, Erasmus Universiteit Rotterdam, Vakgroep beslissingsgerichte informatiewetenschappen.

[10] L.J. Cronbach (1951), *Coefficient alpha and the internal structure of tests*. Psychometrika 16 p. 297-334.

[11] J. Martin Bland, D. G. Altman (1997), *Statistics notes: Cronbach's alpha*. BMJ Vol. 314 p. 572.
http://www.bmj.com/cgi/reprint/314/7080/572.pdf

[12] K. M. Stokking, *Meten, meetniveaus, schaalmethoden en schaalanalyses*. Bouwstenen voor onderzoek in onderwijs en opleiding, Universiteit Utrecht.
http://studion.fss.uu.nl/Bouwstenenonline/startpagina.doc

[13] *Reliability Estimation Using a Split-half Methodology*, A guide for developing assessment programs in Illinois schools (1995).
http://www.gower.k12.il.us/Staff/ASSESS/4_ch2app.htm#Spearman-Brown

[14] N. Schmitt (1996), *Uses and abuses of Cronbach's alpha*. Psychological Assessment Vol. 8 No. 4 p.350-353.
http://socrates.berkeley.edu/~maccoun/PP279_Schmitt.pdf

[15] A. Christmann, S. Van Aelst (2006), *Robust estimation of Chronbach's alpha*. Journal of Multivariate Analysis Vol. 97 Issue 7 p. 1660-1674.

[16] J. M. van Zyl, J. Martin, N. Heinz, D. G. Nel. (2000), *On the distribution of the maximum likelihood estimator of Cronbach's alpha*. Psychometrika 65 p. 271–280.

[17] D. Iacobucci, A. Duhachek (2003), *Advancing Alpha: Measuring reliability with confidence*. Journal of consumer psychology, 13(4) p. 478-487.

[18] J. Oosterhoff, A.W. van der Vaart (2003), *Dictaat Algemene statistiek,* Vrije Universiteit Amsterdam.

[19] W. A. Brownell, *On the accuracy with which reliability may be measured by correlating test halves*. J. exper. Educ. , 1933, 1, 204-215.

[20] G. F. Kuder, and M. W. Richardson, *The theory of the estimation of test reliability*. Psychometrika, 1937, 2, 151-160.

[21] M.C.M. de Gunst, A.W. van der Vaart (2005), *Dictaat Statistische Data Analyse*, Vrije Universiteit Amsterdam.