INTERBEOORDELAAR-BETROUWBAARHEID

Het coderen van kenmerken van survey items voor de constructie van vragenlijstprofielen

Tegenwoordig worden dezelfde surveys vaak afgenomen via meerdere modes, zoals in persoon, via de telefoon en het web. Afhankelijk van de wijze van afname kunnen de antwoorden van respondenten echter verschillen. Om inzicht te krijgen in voor wat voor soort vragen deze antwoorden verschillen, hebben we de items van enkele grote surveys, waaronder de Enquête Beroepsbevolking van het Centraal Bureau voor de Statistiek, door meerdere beoordelaars op hun kenmerken laten coderen. Uit het onderzoek blijkt dat de interbeoordelaar-betrouwbaarheid relatief laag is voor een groot aantal van de kenmerken. Dit suggereert dat onderzoekers voorzichtig moeten zijn in het beoordelen en het gebruik van kenmerken om inzicht te krijgen in mode-specifieke meetfouten.

Frank Bais, Barry Schouten & Vera Toepoel

Tegenwoordig worden steeds meer surveys afgenomen via meerdere survey modes. Een gebruikelijke vorm hiervan is in eerste instantie afname via het web en aansluitend via telefoon en/of in persoon. Deze vorm is relatief gunstig, omdat de aanvankelijke afname via het web kosteneffectief is en de follow up afname via telefoon en/of in persoon zorgt voor een verhoging van het responspercentage. Afname via het web is immers relatief goedkoop en door in meerdere modes de survey af te nemen wordt een groter gedeelte van de groep beoogde respondenten aangesproken voor potentiële deelname. Een gevolg van afname via meerdere modes is dat antwoorden van respondenten op dezelfde vragen in verschillende modes niet per definitie vergelijkbaar zijn (Klausch, Hox & Schouten, 2013). De verschillende modes kunnen tijdens het antwoordproces leiden tot verschillende meetfouten, welke mode-specifieke meetfouten worden genoemd. Deze mode-specifieke meetfouten maken het vergelijken van antwoorden in verschillende modes moeilijk en onbetrouwbaar. Rekening houden met mode-specifieke meetfouten wordt dikwijls onvoldoende gedaan, omdat dit moeilijk is, veel tijd kost, of omdat men zich hier niet van bewust is. In ons onderzoek proberen we inzicht te krijgen in het optreden en de omvang van deze mode-specifieke meetfouten. We gebruiken hier onder andere de Enquête Beroepsbevolking (EBB) van het Centraal Bureau voor de Statistiek (CBS) voor, wat één van de surveys is die dient voor de productie van de officiële CBS-statistieken. Het doel is uiteindelijk om te proberen om specifieke typen surveys en specifieke survey modes op elkaar toe te spitsen om zo mode-specifieke meetfouten te minimaliseren.

Het optreden en de omvang van mode-specifieke meetfouten worden onder andere bepaald door de kenmerken van de vragenlijst (Beukenhorst, Engelen, Van der Laan, Meertens en Schouten, 2013). We willen deze kenmerken samenvatten in wat we vragenlijstprofielen zullen noemen en onderzoeken in hoeverre deze vragenlijstprofielen variatie in antwoordgedrag kunnen verklaren. Indien die variatie substantieel blijkt, zou het mogelijk moeten zijn om op basis van deze profielen uiteindelijk respondent, vragenlijst en survey mode op elkaar toe te spitsen. Als een eerste stap willen we vragenlijstprofielen construeren op basis van kenmerken van een groot aantal items van enkele omvangrijke nationale surveys. Hiervoor zal elk item in de vragenlijst op de aanwezigheid van elk kenmerk afzonderlijk beoordeeld moeten worden. Voordat dergelijke vragenlijstprofielen geconstrueerd kunnen worden, zal onderzocht moeten worden in hoeverre betrouwbaar kan worden vastgesteld dat er sprake is van een bepaald kenmerk voor een bepaald item. In het huidige onderzoek zijn daarom alle items op elk geselecteerd kenmerk gecodeerd door meerdere beoordelaars om te onderzoeken in hoeverre items betrouwbaar gecodeerd kunnen worden. Oftewel, voor welke proportie items zijn meerdere beoordelaars het erover eens dat een bepaald kenmerk zich op bepaalde wijze al dan niet manifesteert?

Opzet coderingsonderzoek

In het huidige onderzoek zijn alle items van de Enquête Beroepsbevolking van het CBS en de tien kernstudies Bezit, Wonen, Inkomen, Gezondheid, Persoonlijkheid, Religie & Etniciteit, Politiek & Waarden, Familie & Huishouden, Werk & Scholing en Sociale Integratie & Vrije Tijd van het LISS Panel (geadministreerd door CentERdata) door twee of drie beoordelaars gecodeerd op 16 item-kenmerken. Aan de hand van typologieën van Saris en Gallhofer (2007) en Gallhofer, Scherpenzeel en Saris (2007), een aantal focusgroepen en een proefcodering is een lijst met deze uiteindelijke 16 item-kenmerken geconstrueerd. De beoordelaars van zowel deze focusgroepen en proefcodering als de daadwerkelijke codering zijn werknemers van Universiteit Utrecht (UU), het CBS en CentERdata en hebben relatief veel kennis van en ervaring met survey onderzoek. Tijdens de coderingen werden de beoordelaars geacht zich strikt te houden aan de omschrijving en categorieën van de kenmerken. Zeven kenmerken hebben volgens de literatuur relatief veel invloed in het opwekken van mode-specifieke meetfouten en zullen we derhalve de lastige kenmerken noemen: Inhoud van de vraag, lastig taalgebruik van de vraag, emotionele lading, potentieel vermoeden van een filtervraag, gevoelige informatie, centraliteit en complexiteit van de antwoordopties. Deze kenmerken zijn gecodeerd door drie beoordelaars. De overige negen kenmerken tijd, condities, herinnering, hypothetische situatie, berekeningen, dubbelloops, mismatch, formulering en toelichting noemen we de makkelijke kenmerken. Deze kenmerken zijn gecodeerd door twee beoordelaars. Zie tabel 1.

Uitkomsten coderingsonderzoek

Voor elk kenmerk is de interbeoordelaar-betrouwbaarheid over alle items van alle surveys berekend. Zie tabel 1. Deze bestaat uit de totale proportie items waarvoor door beide of alle drie de beoordelaars dezelfde categorie gecodeerd is. Als vuistregel is aangehouden dat een interbeoordelaar-betrouwbaarheid van 0,80 of hoger acceptabel is. Voor wat betreft de makkelijke kenmerken bleken de kenmerken formulering en toelichting niet betrouwbaar gecodeerd te kunnen worden. Voor wat betreft de lastige kenmerken bleken inhoud van de vraag, lastig taalgebruik van de vraag, emotionele lading, potentieel vermoeden van een filtervraag, gevoelige informatie en centraliteit niet betrouwbaar gecodeerd te kunnen worden. Slechts het aangepaste kenmerk inhoud van de vraag met één objectieve en één subjectieve categorie en het kenmerk complexiteit van de antwoordopties van de lastige kenmerken bleken betrouwbaar gecodeerd te kunnen worden.

Verklaringen voor lage interbeoordelaar-betrouwbaarheid

Een eerste mogelijke verklaring voor de gevonden lage interbeoordelaar-betrouwbaarheid is het idee dat het relatief moeilijk is om de kenmerken sluitend te definiëren. Ook al zijn de kenmerken gebaseerd op de literatuur en extensieve discussies tussen de beoordelaars, het is voor bepaalde kenmerken vermoedelijk relatief moeilijk om te bepalen voor de beoordelaars welke categorie van toepassing is, doordat de definitie niet doorslaggevend blijkt voor relatief veel items. Veel items blijken niet eenduidig te coderen op basis van de afgesproken definitie. Daarom verwachten wij dat er voor de meeste surveys relatief veel items zullen zijn die niet eenvoudig beoor-

deeld kunnen worden op bepaalde kenmerken, ongeacht hoe strak deze kernmerken conceptueel ook gedefinieerd mogen zijn. De interpretatie van items kan niet alleen van beoordelaar tot beoordelaar verschillen, maar is ook afhankelijk van zaken zoals persoonlijkheid, voorgeschiedenis en stemming van de dag, welke allemaal hun invloed kunnen hebben op de wijze waarop een kenmerk geïnterpreteerd wordt en vervolgens hoe bepaalde items worden gecodeerd voor het desbetreffende kenmerk. Vanuit dit perspectief bezien zal de interbeoordelaar-betrouwbaarheid gedeeltelijk afhangen van welke beoordelaars de items coderen. De verwachting is dat zowel an-

KENMERK	OMSCHRIJVING	CODENUMMER EN CODECATEGORIE	α
Tijd	Naar welke tijd refereert het item?	1 verleden 2 heden 3 toekomst	0,85
Condities	Bevat het item condities of voorwaarden?	0 nee 1 ja	0,89
Herinnering	Is er sprake van een bepaalde herinnering?	0 geen herinnering 1 herinnering non-specifiek 2 herinnering < 1 maand 3 herinnering > 1 maand	0,85
Hypothetische situatie	Is er sprake van een concrete, specifieke hypothetische situatie?	0 nee 1 ja	0,98
Berekeningen	Dient respondent een bepaalde berekening uit te voeren?	0 nee 1 ja	0,94
Dubbelloops	Bevat het item meerdere vragen of is het potentieel verwarrend?	0 nee 1 ja	0,96
Mismatch	Vraag en antwoordcategorieën sluiten niet op elkaar aan.	0 nee 1 ja	0,98
Formulering	Is de vraag geformuleerd als een stelling?	0 nee 1 ja	0,57
Toelichting	Is er toelichting aanwezig bij een item?	0 nee 1 ja	0,71
Inhoud van de vraag 1	Naar wat voor inhoudelijk aspect wordt gevraagd?	1 feitelijk gedrag 2 overig feitelijk 3 mening 4 tevredenheid 5 overig subjectief	0,56
Inhoud van de vraag 2	Naar wat voor inhoudelijk aspect wordt gevraagd?	1 objectief 2 subjectief	0,90
Lastig taalgebruik vraag	Bevat vraag moeilijke woorden of zinsconstructies?	0 nee 1 ja	0,61
Emotionele lading	Bevat het item emotionele woorden of lading?	0 nee 1 ja	0,75
Vermoeden van filtervraag	Kan respondent vermoeden dat vraag filtervraag zou kunnen zijn?	0 nee 1 ja	0,62
Gevoelige informatie	Bevat het item gevoelige informatie van maatschappelijke of huishoudelijke aard?	0 nee 1 ja	0,53
Centraliteit	Valt de vraag buiten de kennis- of belevingswereld van respondent?	0 nee 1 ja	0,59
Complexiteit antwoordopties	Bevatten antwoordopties moeilijke woorden of zinsconstructies?	0 nee 1 ja	0,91

Tabel 1. Omschrijving, categorieën en interbeoordelaar-betrouwbaarheden α van de kenmerken

dere beoordelaars als dezelfde beoordelaars die dezelfde items nog een keer opnieuw coderen, zullen resulteren in verschillende interbeoordelaar-betrouwbaarheden.

Omgaan met lage interbeoordelaar-betrouwbaarheid

Hoe nu om te gaan met lage interbeoordelaar-betrouwbaarheid? Een eerste optie is het simpelweg negeren van alle items waarvoor geen consensus gevonden is in het coderen en om de vragenlijstprofielen puur te baseren op de items waarvoor alle beoordelaars het met elkaar eens waren voor een bepaald kenmerk. Dit heeft niet de voorkeur, omdat er op deze wijze informatie ongebruikt blijft en vragenlijstenprofielen vooral voor wat betreft de lastige kenmerken gebaseerd zullen zijn op een beperkt aantal items. Een tweede optie is om de kenmerken te verfijnen en herdefiniëren op basis van alle items waarvoor geen consensus bereikt is. De nieuwe definitie wordt dan aangepast en aangevuld op basis van de inhoud van deze items met behulp van de literatuur. Op deze wijze bestrijken de definities alle beoordeelde items en kunnen vragenlijstprofielen voor elk kenmerk geconstrueerd worden op basis van alle items.

Om het subjectieve karakter van het coderen te voorkomen, is een derde optie het computeriseren van de definities van de kenmerken en het eigenlijke codeerproces. Door de definities en de beslissingen voor een categorie van alle kenmerken te formaliseren, worden subjectieve interpretaties van de beoordelaars voor items die binnen het eerder genoemde grijze gebied vallen, vermeden. Het valt echter te bezien in hoeverre dit computeriseren realistisch blijkt, wat met name voor de lastige kenmerken vermoedelijk moeilijk zal zijn. Bovendien zijn ook hier beoordelaars en programmeurs nodig en ontkomt men niet aan het subjectieve karakter van het opstellen en programmeren van regels en beslissingen. Om zowel het interpretatieve karakter van het herdefinieren als het computeriseren van de kenmerken te vermijden, is een vierde optie het construeren van schalen met meerdere categorieën in plaats van enkel de categorieën van toepassing en niet van toepassing. Dit betekent dat bijvoorbeeld geen, één, twee of drie beoordelaars voor een bepaald item hebben aangegeven dat een bepaald kenmerk van toepassing is. Hier kan een vragenlijstprofiel op gebaseerd worden dat de aanwezigheid van een bepaald kenmerk gradueel weergeeft, afhankelijk van hoeveel beoordeelaars hebben aangegeven dat sprake van het kenmerk zou zijn.

Conclusie

Survey items kunnen niet betrouwbaar gecodeerd worden. Dit geldt voor de meeste lastige kenmerken, maar ook voor de makkelijke kenmerken formulering en toelichting. De lage interbeoordelaar-betrouwbaarheid kan mogelijk verklaard worden doordat kenmerken moeilijk sluitend te definiëren zijn en dat enige subjectieve interpretatie tijdens het beoordelen onvermijdelijk lijkt en bovendien afhankelijk is van de beoordelaar. Dit is van belang voor elke onderzoeker die met item-kenmerken werkt. Dat wat de één bestempelt als een vraag over bijvoorbeeld kennis, gedrag of attitude, kan door iemand anders op andere wijze geclassificeerd worden. Dit kan een belangrijke oorzaak zijn van tegenstrijdige conclusies over item-kenmerken in de literatuur. Doordat survey items niet betrouwbaar gecodeerd kunnen worden op hun kenmerken, is het moeilijk om deze te gebruiken voor onderzoek naar mode-specifieke meetfouten. Toekomstig onderzoek zal zich moeten richten op het omgaan met lage interbeoordelaar-betrouwbaarheid.

Literatuurlijst

Beukenhorst, D., Buelens, B., Engelen, F., Van der Laan, J., Meertens, V., & Schouten, B. (2013). *The impact of survey item characteristics on mode-specific measurement bias in the Crime Victimisation Survey*. Discussion paper 201416. Den Haag: Statistics Netherlands,

Gallhofer, I., Scherpenzeel, A., & Saris, W. E. (2007). *The codebook for the SQP program*. Gevonden op <www.sqp.nl>.

Klausch, L. T., Hox, J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods and Research*, *42*, 227-263.

Saris, W. E., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1*, 29-43.

Frank Bais heeft een onderzoeksmaster psychologie gevolgd aan de Universiteit van Amsterdam met als major Methoden & Statistiek. Sinds januari 2014 is hij promovendus op de afdeling Methoden & Statistiek van de Universiteit Utrecht. Zijn positie betreft een samenwerkingsproject tussen de UU en het CBS waarin met behulp van kenmerken van respondenten en vragenlijsten geprobeerd wordt om inzicht te verkrijgen in mode-specifieke meetfouten.
E-mail: <f.bais@uu.nl>

Barry Schouten cv

Vera Toepoel cv