

Lezingendag Nederlandstalig Platform voor Survey Onderzoek (NPSO) Innovatie in survey onderzoek – jonge onderzoekers aan het woord

Datum: 19 mei 2015

Tijd: 10:15 – 16:30

Locatie: Centraal Bureau voor de Statistiek, Henri Faasdreef 312, Den Haag

Organisatie vanuit NPSO: Fannie Cobben (fcbn@cbs.nl), Barry Schouten (bstn@cbs.nl)

en Vera Toepoel (v.toepoel@uu.nl)

Samenvatting: In Nederland en Vlaanderen vindt onderzoek plaats naar allerlei thema's binnen de survey methodologie; van vragenlijstontwerp tot prijsindexcijfers. Dit onderzoek leidt op termijn tot innovatie in de waarneming, verwerking en analyse van surveys. We willen met de lezingendag jonge onderzoekers de mogelijkheid bieden om hun werk aan een breed publiek te presenteren en hun resultaten en plannen te confronteren met de praktijk. In tegenstelling tot andere NPSO bijeenkomsten richt de lezingendag zich op het gehele terrein aan survey methodologie. Daarnaast willen we met de dag een discussie voeren over sterke en zwakke plekken binnen het onderzoek: waar doen we voldoende en waar laten we mogelijkheden onbenut? We vragen alle deelnemers en presentatoren om met voorstellen te komen voor innovatie in survey methodologisch onderzoek. Deelnemers kunnen op deze ideeën en voorstellen stemmen en in het programma is tijd gereserveerd voor een plenaire discussie.

Programma

09:45 – 10:15	Registratie met koffie en thee	
10:15 – 10:30	Algemene opening	
10:30 – 11:00	Visualisatie van statistische output – Jan van der Laan en Edwin de Jonge (CBS)	
11:00 – 11:10	Korte pauze (wisseling van zalen)	
11:00 – 12:20	Parallel sessie 1	Parallel sessie 2
	Arnaud Wijnant (UvT)	Laura Boeschoten (UvT)
	Job van den Berg (TNS-Nipo)	Lianne Ippel (UvT)
	Celine Wuyts (KU Leuven)	Sander Scholtus (CBS-VU)
	Babs Broekema (UL)	Susanna Gerritse (UU)
12:20 – 13:10	Lunch	
13:10 – 14:30	Parallel sessie 3	Parallel sessie 4
	Marieke Haan (RUG)	Roline Kamphuis (UvT)
	Margreet Kerkmeer (Pearson)	Jurian Meijering (LUW)
	Thomas Klausch (UU)	Anja Boeve (RUG)
	Oksana Balabay (KUL)	Jannine van de Maat (UL)
14:30 – 14:50	Pauze (wisseling zalen)	
14:50 – 15:50	Plenaire discussie Innovatie in survey onderzoek	
15:50 – 16:20	Innovatie in survey onderzoek - Robert van Ossenbruggen (Customer Central)	
16:20 – 16:30	Sluiting	
16:30 – 17:00	Borrel	

Samenvattingen presentaties

Het meten van studiegedrag in het hoger onderwijs : Problemen met bestaande maten en een agenda voor toekomstig onderzoek - Anja Boeve (RUG)

Er wordt ontzettend veel belang gehecht aan innovatie en onderzoek om de kwaliteit van het hoger onderwijs te verbeteren. Dit onderzoek is in zeer grote mate afhankelijk van data verkregen uit vragenlijsten. De kwaliteit van de data laat echter vaak te wensen over, en ten grondslag aan dit probleem is het gebrek aan zorgvuldige vragenlijst ontwerp en onderzoek gericht op de evaluatie van vragenlijsten.

Een belangrijk voorbeeld is het meten van studiegedrag. Met name in het onderwijs van de sociale wetenschappen wordt een grote mate van zelfstandigheid van studenten verwacht. Innovatie in het cursus ontwerp van hoger onderwijs worden ingevoerd om het studiegedrag van studenten te sturen zodat de prestaties verbeteren. Bij de evaluatie van de implementatie is het studiegedrag van studenten daarom een cruciaal element om te onderzoeken.

Vaak wordt studiegedrag echter bevraagd door bijvoorbeeld tijdsbesteding, waarbij studenten moeten schatten hoeveel uur per week ze gemiddeld genomen aan een bepaald vak hebben besteed. Uit onderzoek is regelmatig geen relatie tussen studietijd en prestaties gevonden, en gegeven deze manier van meten is dat niet verassend, en levert ook geen bijdrage aan de evaluatie van onderwijsinnovaties.

Anderzijds, is het perspectief dat het juist gaat om de kwaliteit van hoe studenten leren, en om dit te meten zijn er vragenlijsten die ingaan op het typeren van studie gedrag als bijvoorbeeld 'diep leren' of 'oppervlakkig leren'. Hierbij krijgen studenten stellingen over gedrag, waarbij ze kunnen aangeven in hoeverre ze dit nooit tot altijd doen. Ook met deze bestaande, en zogenaamd gevalideerde vragenlijsten, is het lastig om innovatie in onderwijs te evalueren.

Innovatie op het gebied van vragenlijst onderzoek om studiegedrag te meten is dus noodzakelijk. Ik illustreer hoe ik in een van mijn studies naar onderwijs innovatie het studiegedrag van studenten heb gemeten op een manier die probeert met deze problemen van zowel tijdsbesteding als de kwaliteit van hoe studenten leren rekening te houden. Tot slot zal ik ingaan op het ontwerp van onderzoek wat staat ingepland om deze manier van het meten van studiegedrag te evalueren en te verbeteren.

De gevolgen van automatische optimalisatie voor mobiele gebruikers in een web panel - Arnaud Wijnant (Centerdata, UvT)

Websurveys worden steeds meer bezocht door respondenten met een smartphone of tablet. Omdat deze apparaten andere schermgroottes en meestal geen fysiek toetsenbord en muis hebben, is het verstandig om hiermee rekening te houden in de manier waarop deze vragenlijsten worden weergegeven. Een manier om dit te doen is het toepassen van een liquid layout in de vragenlijsten. In eerder onderzoek binnen ons panel zagen we echter dat er een groep panelleden is die nog steeds een PC gebruikt en deze liquid layout minder waardeert. Om deze panelleden een optimale gebruikerservaring te geven, zouden alle vragenlijsten daarom zowel in een PC versie als een mobiele versie beschikbaar moeten zijn. Het dubbel programmeren van vragenlijsten verhoogt echter de kosten van vragenlijsten aanzienlijk. Om deze reden ontwikkelden wij de zogenaamde amfibische vragenlijsten. Dit zijn vragenlijsten die zowel een PC layout als een mobiele layout automatisch kunnen genereren. Er wordt iedere keer gekeken welk apparaat de vragenlijst bezoekt en aan de hand daarvan de geschikte layout gekozen. Dit heeft als voordeel dat er maar één data model is, waardoor mensen gemakkelijk kunnen wisselen tussen verschillende apparaten binnen een vragenlijst. Daarnaast hoeft er maar 1 vragenlijst geprogrammeerd te worden.

In deze presentatie laat ik de resultaten zien van een onderzoek over de gevolgen van het toepassen van een amfibische vragenlijst in een panel. In een split-ballot experiment hebben we gekeken naar 4 aspecten die beïnvloed kunnen worden door het toepassen van deze

vragenlijsten. Deze zijn: measurement error, response error, usability scores en tevredenheid van de respondent. Met de resultaten van dit onderzoek zullen we de basis leggen voor een web panel dat rekening houdt met zowel de gebruikers van PCs als mobiele apparaten.

En de winnaar is... - Babs Broekema & Joop van Holsteyn (Universiteit Leiden)

Al ruim tien jaar organiseert het actualiteitenprogramma EenVandaag de verkiezing van Politicus van het Jaar. Half december wordt in een speciale uitzending de winnaar op televisie bekend gemaakt. Wie winnaar is, wordt bepaald door de deelnemers van het EenVandaag Opiniepanel. Echter, hoe deze verkiezing precies is opgezet en wordt uitgevoerd, kan gevolgen hebben voor de uitkomsten ervan. En kan bepalen wie met de eer van Politicus van het Jaar gaat strijken, een niet onbelangrijke titel in een tijd waarin het persoonlijke en individuele in de politiek van voornaam belang wordt geacht. In onze bijdrage gaan we in op onderzoekstechnische aspecten van de verkiezing van Politicus van het Jaar in een tijd van voortschrijdende personalisering van de politiek.

De interviewer in de schoenen van de respondent: Wat kunnen we leren uit hoe interviewers zelf surveyvragen beantwoorden? - Celine Wuyts (KU Leuven)

In surveys met face-to-face interviews nemen de interviewers een sleutelpositie in. Hij legt contact met de respondent, overtuigt om deel te nemen, stelt de vragen en vult de antwoorden in. De ene interviewer slaagt er al beter in dan de andere om de verschillende taken te vervullen zoals dat van hem verwacht wordt. Dit leidt tot de vraag wat de 'betere interviewers' kenmerkt. Meer ervaring? Lagere werkbelasting? Een specifieke persoonlijkheid? Heel wat interviewerkenmerken kunnen overwogen worden. Op basis van administratieve bronnen zijn geslacht, leeftijd en ervaring gewoonlijk vrij direct beschikbaar. Specifieke interviewervragenlijsten kunnen ontwikkeld worden om andere kenmerken te verzamelen, maar dit vereist een extra inspanning van onderzoekers en interviewers. In deze presentatie licht ik een derde, ongebruikelijke, bron van interviewer informatie toe: de respondentenvragenlijst, ingevuld door de interviewers. Voor ESS België worden de interviewers gevraagd de respondentenvragenlijst in te vullen voorafgaand aan het veldwerk om kennis te maken met de vragen. Indirect creëert deze activiteit een potentiële schat aan interviewer informatie. Enerzijds kunnen formele indicatoren worden afgeleid: kenmerken zoals snelheid, volledigheid, straightlining geven een indicatie van de zorg en moeite waarmee de vragenlijst wordt ingevuld. We veronderstellen dat 'interviewer satisficing' zich uit zowel in deze taak van de interviewer-als-respondent, als in de eigenlijke taken van de interviewer. Anderzijds is het mogelijk dat de vragenlijst ook inhoudelijk bruikbaar is, via de bevroegde thema's zoals vertrouwen. De analyse is gebaseerd op data van de zesde ronde van ESS in België en interviewer data van dezelfde vragenlijst.

Vraageffecten in Websurveys - Jannine van de Maat (Universiteit Leiden)

Dat opiniepeilingen in aantal toenemen en dat bij tijd en wijle politici worden opgeroepen om rekening te houden met deze vorm van publieke opinie kan geen verrassing vormen, maar de vraag blijft hoe bruikbaar en waardevol de informatie voortkomend uit opiniepeilingen is. Het is immers bekend dat tal van elementen van (het ontwerp van) een vragenlijst invloed hebben op de uitkomsten (zie onder meer Bradburn, Sudman & Wansink 2004; Schuman & Presser 1981). Hierbij is één belangrijk vraagstuk of 'de publieke opinie' werkelijk is gebaseerd op de verzameling van individuele *meningen*. Hebben mensen eigenlijk meningen over de vele, uiteenlopende onderwerpen die aan bod komen, zeker in een tijd van eenvoudige en goedkope internetpeilingen, en, specifiek, wat zijn de mogelijkheden en de effecten van alternatieve manieren om respondenten de gelegenheid te bieden aan te geven in voorkomende gevallen geen opinie te hebben? Deze vraag, die niet origineel is maar waarvan de bevredigende

beantwoording (zeker in de Nederlandse context) nog op zich laat wachten, staat centraal in het project (promotietraject) waarover in deze presentatie zal worden bericht.

In de presentatie zullen de resultaten van een *split-ballot* surveyexperiment worden gepresenteerd. Hierbij is onderzocht wat de effecten zijn van variaties in vraagstelling op de item nonrespons en de distributie van opinies. Verder zal een onderscheid gemaakt worden tussen onderwerpen en in hoeverre daarbij de vraageffecten van sterkte verschillen.

Hoe WhatsApp de respons onder jongeren kan verhogen bij online surveys - Job van den Berg (TNS NIPO)

Eén van de grootste uitdagingen binnen online survey-onderzoek is het doen van kwalitatief hoogwaardig onderzoek onder jongeren. De respons onder deze groep is vaak laag en er is vaak een hoge uitval bij longitudinale (panel) studies. Een oplossing om de respons en retentie te verhogen, kan zijn om jongeren via WhatsApp te benaderen. Deze app is populair onder jongeren en sluit wellicht beter aan bij hun belevingswereld dan een e-mailuitnodiging. In deze presentatie ga ik in op de resultaten van een grootschalig experiment dat TNS NIPO onlangs heeft uitgevoerd waarin deze hypothese wordt getoetst. In het experiment is gekeken naar de mate waarin WhatsApp in vergelijking tot email en sms een geschikte methode is om jongeren te benaderen voor deelname aan online surveys. Hoe ervaren jongeren de 'WhatsApp-methode' en wat is het effect op de respons? Kunnen we specifieke typen jongeren via WhatsApp bereiken? Is WhatsApp geschikt als volwaardige methode om jongeren voor onderzoek te benaderen? Deze en andere vragen zullen in de presentatie beantwoord worden.

Ranglijsten van Europese groene steden: de methodologische kenmerken in kaart gebracht – Jurian Meijering (LUW)

Diverse partijen hebben ranglijsten van groene steden ontwikkeld. De data die nodig zijn voor de ontwikkeling van deze ranglijsten worden vaak door middel van een gestandaardiseerde vragenlijst verzameld. De partijen claimen vaak dat hun ranglijst in één oogopslag duidelijk maakt welke steden goed en slecht presteren. Het probleem is echter dat er maar zelden kritisch gekeken wordt naar de methodologie van ranglijsten. Het doel van het door ons uitgevoerde onderzoek was dan ook om de methodologische kenmerken van zes prominente ranglijsten van Europese groene steden te identificeren en te evalueren.

Het onderzoek begon met een literatuurstudie die tot doel had om belangrijke methodologische kwesties te identificeren die een rol spelen bij de ontwikkeling van ranglijsten. Na de literatuurstudie zijn de methodologische kenmerken van zes ranglijsten van Europese groene steden gedetailleerd in kaart gebracht door het bestuderen van de officiële websites, eindrapporten en alle beschikbare (methodologische) achtergrond documenten. Ook zijn de ontwikkelaars van de ranglijsten geïnterviewd om zoveel mogelijk ontbrekende informatie boven tafel te krijgen.

De resultaten laten zien dat de methodologische kenmerken van de zes ranglijsten sterk variëren. Daarbij vertonen alle ranglijsten bepaalde methodologische zwaktes. Opzienbarend is met name dat vijf van de zes ranglijsten geen duidelijke benaming en definitie hadden van het overkoepelende concept. Zodoende blijft het bij deze ranglijsten onduidelijk waarop de steden nu daadwerkelijk gerangschikt zijn. Bovendien moet de vraag worden gesteld hoe zonder een duidelijk gedefinieerd

concept een degelijke selectie van indicatoren en bijbehorende vragenlijst kon worden gemaakt. Ook bij de analyse van de eenmaal verzamelde kwantitatieve en kwalitatieve data kunnen een aantal belangrijke methodologische kanttekeningen worden gemaakt.

Op basis van de resultaten worden ontwikkelaars van ranglijsten geadviseerd om kritisch te reflecteren op de methodologische kenmerken van hun ranglijst en waar mogelijk verbeteringen aan te brengen. Ontwikkelaars dienen op z'n minst transparant te zijn over de methodologie van hun ranglijst en gemaakte keuzes zo goed mogelijk te onderbouwen. Gebruikers van ranglijsten

dienen zich eerst te verdiepen in de methodologische kenmerken van een ranglijst, en zodoende zich een beeld te vormen van de kwaliteit van de ranglijst, alvorens te handelen op de resultaten.

Van pen-en-papier naar iPads: psychologische tests voor kinderen en volwassenen – Margreet Kerkmeer (Pearson Assessment & Information)

Psychologische tests voor bijvoorbeeld intelligentie, taalontwikkeling en persoonlijkheid worden tot nu toe meestal op pen-en-papier afgenomen en handmatig gescoord. Wij zijn echter over aan het gaan om een aantal tests op tablets af te nemen. In Amerika is hier al enige ervaring mee opgedaan in een paar pilotonderzoeken. Vragen die we hebben zijn:

- 1) Zijn normen die verzameld zijn op pen en papier ook bruikbaar voor digitale afnames?
- 2) Hoe betrouwbaar en valide zijn digitale afnames?
- 3) Wat zijn verdere haken en ogen?

Verdere aandachtspunten zijn beveiliging en privacy, en training van de testleiders.

Effecten van Enquêtemethoden op Participatie en Antwoordgedrag - Marieke Haan (RUG)

Er is een trend van dalende respons in enquêteonderzoek en ook blijkt dat bepaalde groepen niet goed worden gerepresenteerd binnen enquêtes. Om te onderzoeken of het gebruik van meerdere enquêtemethoden respons zou kunnen verhogen en de representatie van groepen die moeilijk te enquêteren zijn zou kunnen verbeteren, hebben wij een experiment uitgevoerd. Echter, het gebruik van meerdere enquêtemethoden kan tevens methodespecifieke meetfouten veroorzaken. Daarom is ook onderzocht of verschillende enquêtemethoden effecten hebben op antwoordgedrag.

Er zijn vier groepen geselecteerd die bekend staan als 'moeilijke groepen' voor enquêteparticipatie; jongvolwassenen, huishoudens met meer dan één voltijdswerkende, inwoners van grote steden, en etnische minderheden (Stoop, 2005;2007). Om data te verzamelen hebben wij een experimenteel design waarin enquêtemethoden tegelijk werden aangeboden gecombineerd met een design waarin respondenten werden toegewezen aan een enquêtemethode. De huishoudens waren gerandomiseerd toegewezen aan vijf experimentele groepen. Groep 1 was persoonlijk gecontacteerd en had een keuze tussen een CAPI of Internet-enquête. Groep 2 was telefonisch gecontacteerd en had een keuze tussen een CATI of Internet-enquête. De andere steekproefeenheden zijn gerandomiseerd toegewezen aan CAPI (groep 3), CATI (groep 4), of Internet (Groep 5). In groepen 1, 2, en 5 zijn de helft van de Internet respondenten begonnen met een traditionele Internet-tekstmethode en zijn halverwege overgestapt op een Internet-videomethode, de andere helft begon met de Internet-videomethode en ging halverwege de vragenlijst over op de Internet-tekstmethode. In de Internet-videomethode werd vooraf opgenomen videomateriaal getoond aan de respondent waarin een interviewer de vragen voorlas. De respondent kon zelf het gewenste antwoord selecteren na het beluisteren van de vraag.

Onze resultaten laten geen effect zien van het kiezen van een enquêtemethode op participatie. Wel blijkt dat jongvolwassenen en voltijdswerkende respondenten een voorkeur hebben voor de Internet-enquête. Aangaande het antwoordgedrag zijn er meer 'weet niet' antwoorden gevonden in CAPI en CATI in vergelijking met Internet-video. Er zijn geen methodespecifieke effecten gevonden voor non-differentiatie. Minder sociaal wenselijke antwoorden zijn gevonden in de Internet-videomethode in vergelijking met CAPI en CATI. Over het algemeen blijkt dat de Internet-videomethode meer lijkt op de Internet-tekstmethode dan op CAPI of CATI.

Consistente schattingen voor categorische data gebaseerd op een combinatie van administratieve databronnen en surveys – Laura Boeschoten (Universiteit van Tilburg)

Om kosten te besparen op de verzameling en verwerking van data, combineren statistische instituten zoals het CBS indien mogelijk administratieve databronnen met surveys. De grootste

uitdaging hierbij is om consistente (gelijke) populatieschattingen te verkrijgen. De gecombineerde dataset bevat zowel unit als item non-response, wat door zowel wegen, imputeren als macro-integratie opgelost zou kunnen worden. In dit PhD project focussen we ons voornamelijk op het imputeren van de gecombineerde dataset.

Wanneer we met een gecombineerde dataset werken, is het belangrijk dat we rekening houden met de zogenaamde "edit rules": we moeten ervoor zorgen dat de geïmputeerde data realistisch zijn, een getrouwd kind of een zwangere man is dat bijvoorbeeld niet.

Het eerste probleem dat we met dit project willen oplossen ontstaat wanneer verschillende variabelen (afkomstig uit verschillende bronnen) hetzelfde fenomeen meten. Er kunnen dan mogelijk conflicterende observaties tussen zitten. Dit kunnen we oplossen door middel van een latenteklassenmodel waarbij we de "echte waarden" van dit fenomeen schatten, en hierbij rekening houden met de edit rules.

Als het probleem van conflicterende observaties is opgelost, kan multiële imputatie in combinatie met edit rules gebruikt worden om consistente populatieschattingen te verkrijgen. Hiervoor zijn al methoden ontwikkeld, maar deze moeten zodanig verbeterd worden dat zij de relaties in de data behouden, dat zij grote tabellen en grote aantallen tabellen kunnen schatten, en dat de volgorde van de schattingen niet van invloed is. Deze verbeterde methode zal vergeleken worden met andere (bestaande) methoden en de kwaliteit van deze methode zal onderzocht worden. Vervolgens kan de methode worden uitgebreid om ook toegepast te kunnen worden op longitudinale data.

Schatten en voorspellen in data streams - Lianne Ippel (UvT)

Met de opkomst van social media en smartphones, smart watches etc. heeft data collectie een vlucht genomen. Waar data vroeger vooral verzameld werd met pen en papier, wordt data nu op vele manieren verzameld, waar een respondent zich al dan niet van bewust is. Zo wordt internet zoek gedrag bijgehouden, maar ook iemand zijn belgedrag of welke vrienden hij/zij heeft. Al deze data "stroomt" binnen en zonder goede middelen wordt deze big data al snel te veel om handelbaar te zijn.

Tijdens mijn promotietraject houd ik me bezig met big data en data die continu binnen stroomt. Ik ben hierbij vooral geïnteresseerd in data met een geneste (of anderzijds afhankelijke) structuur. Vele huidige statistische modellen die met dit soort gegroepeerde data werken, hebben op dit moment alle data nog nodig om een voorspelling te doen of om model parameters te schatten. Dit wordt problematisch op het moment dat data snel binnen komt en te veel ruimte in beslag neemt om het in het computer geheugen te houden.

Met onze nieuwe methode zijn we in staat om zowel een Multilevel model te schatten in een data stream, als ook voorspellingen te doen op individueel niveau zonder dat we alle data punten moeten opslaan.

Multilevel Modelling within the Hierarchical Bayesian Approach against Structural Unobserved Component Modelling in a Time Series Small Area Application: the Case of the Dutch Travel Survey (OVIN) – Oksana Balabay (KUL)

This work in progress intends to compare the performance of two kinds of techniques in a time series small area application. The first approach is a multilevel time series model analyzed with a hierarchical Bayesian (HB) approach. The second one is structural (unobserved component) time series model, analyzed with the Kalman filter. These time series techniques help reduce the design variances by using the sampling information accumulated over time in repeatedly conducted surveys, as well as estimate discontinuities due to a redesign of the survey process. The comparative analysis is conducted in terms of point-estimates and their standard errors.

The application chosen is the Dutch Travel Survey where estimates are produced on an annual basis and for different break-downs into domains. The variable of interest considered in this

study is the number of kilometers per person per day covered per transport modality and per motive either at the national (featuring 56 domains), or provincial level (with 672 domains). Both approaches are promising in a sense that these techniques result in more stable time series with much smaller standard errors for the different domains compared to standard design-based approaches.

Another aspect we look at is unreliable design variance estimates in the case of small domains. If these variance estimates are negatively biased, treating them as known in a multilevel setting will likely cause the model estimates to overfit the design point-estimates. This study shows how the state-space univariate analysis can help improve the design variance estimates that are used as an input in a multilevel model estimated with the HB approach.

Finally, consequences of such innovative estimation procedures for the survey design will be discussed.

Introducing Blocked Imputation for dealing with attrition in longitudinal survey data – Roline Kamphuis (UU)

Any longitudinal research project has to deal with missing data. One type of missing data is attrition, which occurs due to respondents not participating in some measurements (waves) of the study. Attrition can occur not only because participants refused to participate further, but also because they move, get ill, or change contact details without informing the researchers.

More recently attrition has been dealt with by imputation methods. However, these techniques often fail to incorporate information on why a particular case or group of cases is missing. As respondents can leave the study for different reasons, it might not be optimal to use the same imputation model for respondents who dropped out for different reasons.

This paper will introduce a new imputation model, called Blocked Imputation, that breaks the imputation procedure into separate steps, one for each type of attrition. This method is compared with multiple imputation in a simulation study.

Validiteit en vertekening van administratieve data voor het maken van statistiek – Sander Scholtus (CBS – VU Amsterdam)

Statistische bureaus gebruiken steeds vaker administratieve data uit bestaande registers als aanvulling op of zelfs vervanging van hun eigen waarneming via enquêtes. Niet elke administratieve bron is echter even geschikt voor statistische doeleinden, om verschillende redenen. Een van de aspecten die hierbij een rol speelt is de aanwezigheid van systematische of toevallige meetfouten. Een belangrijke potentiële bron van fouten in administratieve data zijn definitieverschillen tussen de variabele die gemeten wordt voor administratieve doeleinden en de statistische doelvariabele. Bijvoorbeeld: de variabele *omzet* die gehanteerd wordt door de Belastingdienst is niet per se gelijk aan de variabele *omzet* waarover men in de officiële statistiek wil publiceren; sommige economische activiteiten zijn bijvoorbeeld vrijgesteld van omzetbelasting. Het is daarom belangrijk om de meetkwaliteit van administratieve variabelen voor het maken van statistiek vast te stellen.

Uit de literatuur over vragenlijstontwerp is bekend hoe men, door middel van experimenten, de kwaliteit van gemeten enquêtevariabelen kan bepalen via zogenaamde lineaire structurele vergelijkingsmodellen. Hierbij wordt elke waargenomen variabele gemodelleerd als een niet-perfecte maat voor een onderliggende latente (niet-waargenomen) variabele die men eigenlijk had willen meten. Als de correlatie tussen de waargenomen en latente variabele hoog is, spreekt men van een valide meting. Een praktische beperking van deze methode is dat men meerdere metingen nodig heeft van de latente variabelen. Toepassing van deze modellen op administratieve data kan daarom lastig zijn. Bakker (2012) stelde voor om herhaalde metingen te verkrijgen door de administratieve data te koppelen aan enquêtedata. Bij deze aanpak mogen beide bronnen meetfouten bevatten.

In de officiële statistiek is men vaak geïnteresseerd in populatietotalen of -gemiddeldes. In dat geval is het niet alleen belangrijk om te weten of de administratieve variabele een hoge validiteit heeft, maar ook of de vertekening (*intercept bias*) klein is. In deze presentatie laat ik zien hoe men, naast de validiteit, ook de vertekening in een administratieve variabele kan vaststellen met een structureel vergelijkingsmodel. Ook bespreek ik enkele resultaten van het toepassen van deze methode op omzetdata van de Belastingdienst voor de zogenaamde Korte-termijnstatistieken van het CBS.

Gevoeligheid van de vangst-hervangst methode voor koppelingsfouten – Susanna Gerritse (UU)

De vangst-hervangst methode kan worden gebruikt om te schatten hoeveel mensen zich bevinden in Nederland. Echter de vangst-hervangstmethode heeft een aantal aannames waar niet altijd aan wordt voldaan, te weten onafhankelijkheid van de vangst en hervangst, er is sprake van een gesloten populatie tussen de momenten van vangst en hervangst, alle individuen in de vangst en hervangst behoren tot de populatie en hebben een positieve kans om tot de vangst en de hervangst te behoren, en de vangst en hervangst kunnen perfect worden gekoppeld. Met behulp van sensitiviteitsanalyses is de robuustheid van de schatter in eerder werk bepaald voor de aanname van onafhankelijkheid, in dit werk wordt de robuustheid van de schatter bepaald voor twee andere aannames: perfect koppelen en alle individuen behoren tot de populatie.

Estimating survey errors of mixed-mode designs using survey-based benchmarks – Thomas Klausch (UU)

We evaluated three types of bias – total, measurement, and selection – in three sequential mixed-mode designs of the Crime Victimization Survey (CVS): telephone, mail, and web, where nonrespondents were followed up face-to-face. In the absence of true scores, all biases were estimated against two different types of benchmarks. For the 'single-mode benchmark' (SMB), biases were evaluated against a face-to-face reference survey assuming both measurements and selection mechanism of this mode are optimal. In an alternative analysis, a 'hybrid-mode benchmark' (HMB) was used, where biases were evaluated against a mix of the measurements of a web survey and the selection bias of the face-to-face survey. The HMB is useful if the measurement of web but not the web selection mechanism is deemed optimal. We conducted a split-ballot experiment with the three mixed-mode designs (approx. n=1600 respectively) and a face-to-face benchmark survey (n=1639). In addition, we followed up the respondents in the first part of the sequential designs by face-to-face yielding a second observation on these units. We demonstrate how this data is used to multiply impute the potential answers that respondents in the mixed-mode surveys would have given under the face-to-face or web measurement benchmark modes. This information subsequently allows estimating all biases against the SMB and HMB. In the empirical part of our study, a range of 30 survey variables from the CVS was evaluated. We found that the best strategy for the CVS strongly depended on choice of the measurement but not the selection benchmark. The web and mail (but not telephone) mixed-mode strategies caused a strong total bias against face-to-face (SMB) primarily due to a measurement bias. The face-to-face follow-up could mitigate this bias to some extent but could not nullify it. Using the web/face-to-face HMB, the reverse was true: the telephone strategy showed strong measurement bias, whereas web and mail response samples did not. However, the F2F follow-up to web or mail increased the measurement bias against the HMB suggesting avoiding such a follow up to web or mail if the HMB is deemed optimal.