Het meten van regelnaleving

Een voorstudie in opdracht van
het Expertisecentrum Rechtshandhaving
van het Ministerie van Justitie

rapport

prof. dr. P.G.M. van der Heijden, prof. dr. J.J. Hox, Dr. H. Elffers

Onderzoekschool IOPS, afdeling Utrecht
Faculteit Sociale Wetenschappen
Universiteit Utrecht

Nederlands Studiecentrum Criminaliteit en Rechtshandhaving (NSCR) Leiden

25 augustus 2002

Wetenschappelijk directeur IOPS:
Prof. Dr. P. G. M. van der Heijden
IOPS, FSW, UU
Postbus 80.140, 3508 TC Utrecht
tel. 030-2534688, fax 030-2535797
email p.vanderheijden@fss.uu.nl

Voorwoord

Binnen de Rijksoverheid wordt steeds vaker systematisch onderzoek gedaan naar naleving van regels door het bevragen van potentiële overtreders. Het Expertisecentrum Rechtshandhaving (ERh) van het Ministerie van Justitie richt zich op het ontwikkelen van instrumentarium voor dit type onderzoek.
Dit type onderzoek maakt doorgaans gebruik van twee onderzoeksmethoden, de zogenaamde 'Tafel van elf' voor het meten van motieven, kennis, attituden die in verband staan met naleving van de te onderzoeken regel, en het meten van de mate van regelovertreding met 'randomized response technieken'. Met beide methoden zijn inmiddels positieve ervaringen opgedaan, maar er blijven nog enkele belangrijke vragen te beantwoorden, om de onderzoeksmethoden nog geschikter te maken voor toepassing. Dit rapport beoogt aan te geven welke kennis nog ontbreekt en hoe die kan worden aangevuld om dit te realiseren.

Daartoe hebben zeven experts op het terrein van 'randomized response' en de 'Tafel van elf' een voorstudie verricht. Op grond van deze voorstudie zijn ook onderzoeksvragen voor een onderzoeksprogramma geformuleerd.

De auteurs zijn Ardo van den Hout, Marianne Junger, Edith de Leeuw, Gerty Lensvelt, allen werkzaam bij de Universiteit Utrecht en ondergetekenden. Ger van Gils van BeleidsOnderzoek & Advies (BOA) heeft de redactie van de rapportage gevoerd.

Prof. Dr. P. van der Heijden Prof. Dr. J. Hox Dr. H. Elffers

Capaciteitsgroep Methodenleer en Statistiek, Faculteit Sociale Wetenschappen, Universiteit Utrecht

Nederlands Studiecentrum Criminaliteit en Rechtshandhaving (NSCR), Leiden

Inhoudsopgave:

Samenvatting

Het Expertisecentrum RechtsHandhaving (ERH) van het Ministerie van Justitie richt zich op het ontwikkelen van instrumentarium voor onderzoek naar naleving van regels. Daarbij wordt gebruik gemaakt van de 'Tafel van 11' voor het meten van motieven, kennis, attituden die in verband staan met naleving van de te onderzoeken regel en van 'randomized response technieken' voor het meten van de regelovertreding. Met beide methoden zijn in de uitgevoerde onderzoeken positieve ervaringen opgedaan, maar om de onderzoeksmethoden nog geschikter te maken voor toepassing, zijn nog enkele belangrijke vragen te beantwoorden. Dit rapport beoogt aan te geven welke kennis nog ontbreekt en hoe die kan worden aangevuld om dit streven te realiseren. Op grond van een voorstudie zijn onderzoeksvragen voor een onderzoeksprogramma geformuleerd, maar ook voorstellen voor ontwikkeling van materialen om het gebruik van het bestaande instrumentarium te optimaliseren. Wij geven ook kort aan op welke manier de gestelde vragen te beantwoorden zijn en hoe de materialen ontwikkeld kunnen worden. De verschillende onderzoeksvragen en ontwikkelopdrachten kunnen vervolgens worden uitgezet in de academische wereld.

De onderzoeksvragen hebben betrekking op drie belangrijke onderdelen van onderzoek naar naleving en overtreding van overheidsregels. Het betreft:
1. Dataverzameling met behulp van randomized response en eventueel met andere technieken om de vertrouwelijkheid van gevoelige informatie te verhogen;
2. de analyse van data verzameld met behulp van randomized response;
3. de theorie en meting van achtergronden van naleving en overtreding, vervat in de T11.

Het doel van deze samenvatting is de geïnteresseerde lezer zo kort mogelijk een overzicht te geven van de belangrijkste voorstellen tot optimalisatie van het instrumentarium om regelovertreding te onderzoeken. De hoofdstukken over de verschillende onderdelen bevatten echter ook nog enige ideeën en voorstellen voor onderzoek en optimalisatie van bestaande instrumenten die te gedetailleerd zijn om hier aan de orde te komen.

Onderdeel 1: dataverzameling voor gevoelige onderwerpen

Inleiding

Hieronder bespreken we drie gebieden waarop nader methodologisch onderzoek naar het verzamelen van sensitieve informatie wenselijk is.
1) Allereerst zijn dat de mogelijke alternatieven voor randomized response. Dit is zinvol omdat voor randomized response veel grotere steekproeven nodig zijn om dezelfde nauwkeurigheid van schattingen te bereiken.
2) Vervolgens worden de mogelijkheden bekeken voor alternatieven binnen de randomized response. De nadruk ligt daarbij enerzijds op:
   a) het verhogen van de efficiëntie (kleinere steekproeven) en
   b) het vergroten van de effectiviteit (eerlijker antwoorden) van randomized response, en anderzijds op
   c) het inschakelen van de computer bij de dataverzameling.
3) Tot slot formuleren we voorstellen voor het onderzoeken van de voor- en nadelen van verschillende RR-methoden om kwantitatieve gegevens te achterhalen.

Alternatieven voor randomized response technieken

Het is niet voldoende duidelijk in hoeverre randomized response beter is dan andere methoden die de dreiging van gevoelige vragen eveneens verminderen. Daarom is een vergelijkend onderzoek gewenst, waarin serieuze concurrenten voor de RR techniek zorgvuldig worden geïmplementeerd en de resultaten worden vergeleken. Computer Assisted Self Interviewing (CASI) of zelfs Audio-Casi techniek in een face-to-face interview zijn serieuze concurrenten van RR die statistisch efficiënter zijn.

Voorafgaand aan een (A-)CASI onderzoek zouden in een zorgvuldige ontwerpstudie randvoorwaarden opgesteld moeten worden voor het succesvol implementeren van deze technieken bij gevoelige vragen en

speciale groepen, om vervolgens met cognitieve labs het protocol te optimaliseren. Hetzelfde geldt uiteraard voor Computer Assisted Randomized Response (CARR). Vervolgens dient het veldonderzoek bij voorkeur betrekking te hebben op een vraagstelling binnen criminologisch regelovertredend gedrag.

Alternatieven binnen randomized response technieken

In het navolgende zullen we verschillende varianten van de RR-techniek en verder nog verfijningen binnen de belangrijkste varianten bespreken waarvan sommige vooral voordelen bieden op het vlak van efficiëntie en mogelijkheden om te corrigeren voor cheaten, andere meer op het vlak van de effectiviteit, oftewel het vergroten van de validiteit van de resultaten. Er lijkt in een aantal opzichten een balans te zijn tussen efficiëntie en effectiviteit. Een grotere efficiëntie kan ten koste gaan van de effectiviteit en omgekeerd kan een effectievere techniek ook grotere steekproeven vragen. Een scherp onderscheid tussen technieken om efficiëntie en effectiviteit te vergroten is daarom niet altijd te maken. In het navolgende bespreken we eerst de mogelijkheden die vooral de efficiëntie vergroten en vervolgens die in eerste instantie gericht zijn om het verbeteren van de effectiviteit.

a. Efficiëntie van verschillende randomized response methode

Er zijn drie statistische verfijningen waarmee de efficiëntie van randomized response onderzoek kan worden vergroot. Een eerste methode is het opsplitsen van de steekproef in twee subgroepen met verschillende kansen om de gevoelige vraag te moeten beantwoorden. Een andere manier is het gebruik van een twee-traps procedure. In de eerste fase wordt d.m.v. een randomiser bepaald of de respondent een gevoelige vraag moet beantwoorden (kans T: Ik heb mezelf wel eens schuldig gemaakt aan….) of dat hij of zij een tweede randomisatie moet doen (kans 1-T). In de tweede fase wordt voor de laatste groep wederom een randomisatie gedaan met kans p op het gevoelige statement ‘Ik heb mezelf wel eens schuldig gemaakt aan…….’, en kans 1-p op het statement ‘ik heb me nog nooit schuldig gemaakt aan….’. Een bijkomend voordeel van beide methoden is dat het in principe mogelijk wordt om uitspraken over ‘cheating’ te doen.

De derde statistische verfijning heeft uitsluitend betrekking op het unrelated question design (URQD). Bij een URQD moet een deel van de respondenten antwoord geven op een neutrale vraag, waarvan het antwoord voor de populatie reeds bekend is. Een mogelijk voordeel van URQD is, dat het waarschijnlijk minder snel aanleiding geeft tot cheaten, omdat de respondent nooit ‘gedwongen’ ‘ja’ hoeft te antwoorden op een gevoelige vraag, zoals wel bij Forced response voorkomt. Een belangrijk nadeel is dat het aantal mogelijke vragen beperkt is. Er is een grens aan het aantal vragen waarvoor op voorhand het antwoord voor de populatie bekend is. Wij stellen voor een alternatieve toepassing van het originele URQD te onderzoeken die mogelijk een oplossing biedt. Respondenten in het tweede sub-sample krijgen dan nooit de gevoelige vraag, maar alleen de neutrale vraag. Hierdoor hoeven de populatie parameters voor de neutrale vraag niet te worden geschat. Volgens Moors is de methode efficiënter dan andere varianten van het unrelated question design.

Onderzoeksvoorstellen

Bovenstaande overwegingen geven aanleiding tot de volgende voorstellen voor onderzoek:

1. Het is aan te raden om alle als uitvoerbaar beschouwde statistische modificaties uit te werken en uit te zoeken welke van deze verfijningen efficiënter is.
2. Het is tevens van belang om uit te zoeken welke bijkomende psychologische voor- (en na)delen de methoden met zich mee brengen en welke mogelijkheden ze bieden om te corrigeren voor cheating. De huidige methoden die rekening houden met cheaters maken het niet mogelijk om een verband te leggen tussen kunnen cheating en achtergrondvariabelen. Dit is een groot nadeel. In het kader van de statistische- en softwareontwikkeling (zie onderdeel 2 van deze samenvatting) zou onderzocht moeten worden of hier iets aan te doen is.

De efficiëntie van het opsplitsen van de steekproef en de tweetrapsprocedure, alsmede van de extra voordelen die het detecteren van- en controleren voor cheaten, kan worden onderzocht met behulp van statistische analyses van bestaand onderzoek. De efficiëntie van de door Moors voorgestelde variant van het URQD kan worden onderzocht met simulaties. De psychologische voordelen van de procedures kunnen worden onderzocht in het cognitieve laboratorium.

b. Effectiviteit van RR

In deze paragraaf bespreken we verschillende onderzoeksrichtingen die (vooral) de validiteit van onderzoeksresultaten kunnen vergroten en daardoor tot zuiverder schatters leiden.

Een methode is effectiever wanneer zij resulteert in betere schatters van ware populatie gemiddelden en dus minder vertekening (bias) oplevert. Het is onbekend welke vorm van randomized response onderzoek het meest effectief is. Om deze vraag te beantwoorden zouden nieuwe analyses moeten worden gedaan bij bestaande data-sets, die (nog) niet in de eerder door onderzoekschool IOPS-Utrecht uitgevoerde meta-analyse zijn betrokken. Aanvullend zouden hierbij door middel van een meta-analyse de effecten van verschillende RRT's moeten worden vergeleken met elkaar. Ook is daarbij te onderzoeken in hoeverre de effectiviteit van de RRT samenhangt met de intrensieke of extrensieke bedreiging van de bevraagde onderwerpen. Intrinsiek bedreigend zijn vragen die een respondent confronteren met zichzelf en zijn gevoelens van een positieve identiteit kunnen aantasten. Extrinsiek bedreigend zijn vragen waarbij de respondent sancties van anderen vreest (bijvoorbeeld afkeuring van anderen of wettelijke straffen).

De belangrijkste assumptie van randomized response onderzoek is dat de respondent een gevoel van veiligheid heeft, doordat zijn privacy gewaarborgd wordt. Dat gevoel is echter nog niet gelijk aan de objectieve statistische bescherming. Het is bekend dat mensen in het algemeen weinig correcte intuïties hebben ten aanzien van kansen en kansberekening. Daardoor kan de waargenomen subjectieve bescherming groter zijn dan de statistisch objectieve bescherming. Onderzoek naar de relatie tussen objectieve en subjectieve veiligheid kan bijdragen aan het vergroten van effectiviteit (en efficiëntie) van randomized response onderzoek.

Dit onderzoek kan bestaan uit beperkte literatuurstudie en kleine experimenten in een cognitief laboratorium die varianten uitproberen en mensen vragen hoe zij hun veiligheid inschatten.

Tot slot stellen wij voor een instrument te ontwikkelen waarmee onderzoekers eenvoudig een keuze kunnen maken voor een optimale vorm van RR, met name voor een optimale waarde voor de p-vraag. De keuze voor een methode hangt dan af van de verwachtingen van de onderzoeker ten aanzien van de prevalentie van een fenomeen in de populatie en de daarmee samenhangende keuze van een p-vraag, die enerzijds de veiligheid van de respondent voldoende garandeert, en anderzijds een voldoende efficiëntie belooft.

Wij stellen voor om bij eerder uitgevoerd monitoronderzoek te inventariseren wat realistisch te verwachten prevalenties zijn. Vervolgens kan een voorkeurstabel worden opgesteld met aanbevolen waarden voor de p-vraag bij verschillende te verwachten prevalenties van het onderzochte gedrag.

c. Randomized response technieken bij computergestuurd interviewen

Het is wenselijk computergestuurde varianten van RRT te onderzoeken. Computergestuurd onderzoek biedt veel voordelen. Het is efficiënt in tijd en geld, de data-invoer verloopt automatisch, de non-respons is lager en er is waarschijnlijk een grotere perceptie van privacy. De belangrijkste vraag in dit verband is de vraag naar de meest effectieve manier om randomized response technieken in een computergestuurde vragenlijst te implementeren. Dit betreft een vergelijking van het effect op het begrip en vertrouwen van de respondenten van een externe (bijvoorbeeld het werkelijk gooien van dobbelstenen) versus een interne randomizer (bijvoorbeeld het virtueel 'gooien' van een dobbelsteen op het computerscherm). Deze vraag kan onderzocht worden door gebruik te maken van een cognitief laboratorium.

Wanneer de beste vorm van CARR is vastgesteld, dan blijft de vraag of het gebruik van een combinatie van methoden, computer gestuurd onderzoek gecombineerd met het randomized response format, tot betere resultaten leidt wanneer onderzoek wordt gedaan naar zeer gevoelige onderwerpen of bij speciale populaties, dan wanneer enkel CASI (of eventueel audio-CASI) onderzoek was gedaan. Deze vraag vormt een onderdeel van de vergelijking van RR en alternatieven (zie voorstellen tot vergelijking onder 1, boven).

Randomized response voor kwantitatieve gegevens

RR-methoden om kwantitatieve gegevens te achterhalen zijn bruikbaar voor onderzoek naar regelhandhaving.

Er zijn verschillende methoden om de bevraagde getallen te randomiseren. Wij stellen voor om het onderzoek te beperken tot een vergelijking van directe - met additieve – en multiplicatieve vraagmethoden. Bij additieve methoden vraagt men de respondent om het gevraagde getal op te tellen bij een ander, willekeurig gekozen getal uit een bekende verzameling, bij multiplicatieve vraagt men om het er mee te vermenigvuldigen.

Het onderzoek naar RR voor kwantitatieve gegevens kan in vier stappen worden uitgevoerd:

1. Statistische bureau studies (desk-research) waarbij de efficiëntie van verschillende vormen van kwantitatieve randomized response methoden wordt vergeleken.
2. Statistisch bureau onderzoek naar methoden om de relatie tussen kwantitatieve randomized response variabelen en verklarende variabelen en tussen kwantitatieve randomized response variabelen onderling te leggen.
3. Experimenteel onderzoek naar de manier waarop respondenten omgaan met additieve en multiplicatie methoden. Onderzocht moet worden hoe eventueel rekenwerk voor respondent zo makkelijk mogelijk kan worden gemaakt, zodat de kans op fouten en dus vertekening van de resultaten zo klein mogelijk wordt.
4. Onderzoek is nodig naar de mogelijkheden van het toepassen van kwantitatieve RR-methoden binnen computer gestuurd onderzoek.

Wij stellen voor om de onderzoeksvoorstellen met betrekking tot deze methoden te integreren in het onderzoek naar de efficiëntie en effectiviteit van de randomized response in het algemeen (zie ook het schema in paragraaf 4 op de volgende pagina). Als bijvoorbeeld gekozen wordt voor een vorm van forced response dan kunnen waarschijnlijk alle bij 'efficiëntie en effectiviteit' beschreven aanpassingen ook voor kwantitatieve toepassingen van de randomized response methode worden onderzocht.

Onderzoeksvoorstel vergelijkend onderzoek
In de laatste fase worden op basis van de uitkomsten van de voor dit onderdeel dataverzameling voor gevoelige onderwerpen voorgestelde onderzoeken veelbelovende varianten van RRT gespecificeerd. De volgende stap is om deze varianten met behulp van cognitieve lab methoden te optimaliseren en ze vervolgens te vergelijken in een empirisch veldonderzoek. Hierbij moet ook de mogelijkheid van computergestuurd RR onderzoek (CARR) worden meegenomen.

4. Schematisch overzicht

Een overzicht van de verschillende onderzoeksvoorstellen en de wijze waarop ze kunnen worden uitgevoerd is te vinden in het volgende schema

Onderzoek	**Doel**	**Dichotome randomized response**	**Kwantitatieve Randomized response**
Statistisch (bureau)onderzoek	Efficiëntie	Splitsen steekproef	Splitsen steekproef
		Tweetrapsprocedure	Tweetrapsprocedure
		Modificaties UQT	Modificaties UQT
			Additieve en muliplicatiemodellen vergelijken met FR /UQT
Keuzemoment: Welke RRT(s) optimalisatie(s) heeft (hebben) voorkeur? Dan vervolgonderzoek in veld:			
Veldonderzoek	Inhoudelijk	Mogelijkheden implementatie binnen computergestuurd onderzoek	Mogelijkheden implementatie binnen computergestuurd onderzoek
	Effectiviteit	Cognitieve laboratorium studies naar mogelijkheden om effectiviteit van de gekozen RRT(s) te vergroten	Vergelijken van UQT /FR met additieve en multiplicatie technieken, bij respondenten
	Effectiviteit en efficiëntie	Experiment waarbij (meerdere) RRT(s) wordt(en) vergeleken met (A)-CASI	Gekozen kwantitatieve RRT vergelijken met directe vragen via (A)-CASI

Onderdeel 2: analyse van randomized response data

Software voor de analyse van randomized response data

Voor de analyse van het verband tussen onafhankelijke variabelen en één afhankelijke RR-variabele is op dit moment een bèta-versie van een computerprogramma beschikbaar. Dit is geschikt voor het analyseren van de relatie tussen enerzijds een of meer kwantitatieve - of categorische onafhankelijke variabelen en anderzijds één RR variabele. Het is wenselijk dat dit programma in het publieke domein komt, zodat elke marktpartij die in opdracht van het Ministerie van Justitie RR onderzoek doet, ook logistische regressies kan uitvoeren. Een van de zaken die hiervoor dienen te worden geregeld is een adequate handleiding voor het programma. Verder dient het bestaande programma professioneler te worden gemaakt, en uitgebreid te worden met een aantal zinvolle opties. Het gaat om de volgende zaken:

1) Professionalisering:
   a) Meer gebruikersgemak bij de bediening, bijvoorbeeld m.b.t. de behandeling van missing data.
   b) Het programma moet zelfstandiger kunnen werken. Nu moeten bij elke nieuwe analyse enkele zaken worden veranderd om het programma goed te laten lopen. Dit kost tijd, maar kan ook tot fouten leiden.
   c) Bij analyse van een groot aantal verklarende variabelen loopt het programma soms vast. Het programma zou hiervan de oorzaak dienen te rapporteren. Waarschijnlijk is het vastlopen geen gevolg van een programmafout maar van kenmerken van het gebruikte algoritme.
   d) Het maken van een gebruikersvriendelijke handleiding.
2) Uitbreidingen
   a) Berekening van effectmaten voor van tevoren door de gebruiker op te geven waarden van de verklarende variabelen is een belangrijke uitbreiding van de mogelijkheden van het analyseprogramma. Tot nu toe geeft de presentatie van de resultaten van logistische regressie alleen in woorden de richting van het verband weer en niet een maat voor de sterkte van het verband tussen onafhankelijke - en RR-variabelen. Zogenaamde ‘effectsizes’ geven informatie over zowel de richting als de sterkte van het verband. Bovendien zijn deze maten zeer goed te interpreteren. Het huidige analyseprogramma berekent de effectsizes echter niet standaard: de gebruiker dient de benodigde gegevens zelf in te vullen in formules.
   b) De mogelijkheid om gewichten in te voeren. In het analyseprogramma voor logistische regressieanalyse kan nog niet worden gewogen. Hierdoor levert dit programma geen zuivere schattingen op in de tabel met effectsizes.
   c) Zinvolle uitbreidingen die ook beschikbaar zijn in gewone programma’s voor logistische regressie, bijvoorbeeld classificatietabellen die nu handmatig moeten worden uitgevoerd met alle risico’s van fouten.

De puntschattingen voor RR-variabelen, een eenvoudiger vorm van analyse van RR-data, worden op dit moment met de hand berekend. Om fouten te beperken is het wenselijk om te onderzoeken of deze analyses ook kunnen worden uitgevoerd met software. De software die op de markt is voor Latente klasse analyse, is daarbij een goede kandidaat. Verder dient te worden nagegaan of men met deze software ook de relatie tussen categorische verklarende variabelen en één RR variabele, als ook de relatie tussen twee RR variabelen kan analyseren en zonder extra stappen maten voor de sterkte van relaties tussen de variabelen kan verkrijgen. De thans gebruikte software biedt deze mogelijkheid niet, is daarom niet gebruikersvriendelijk en leidt gemakkelijker tot fouten. Indien de gewenste mogelijkheden bestaan, zou een handboek gemaakt dienen te worden voor dit onderwerp. Het Ministerie kan zo uitvoerders van onderzoeken snel en met minder kans op fouten analyses laten verrichten.

Telvariabelen

Er is in het kader van onderzoek met RR voor het Ministerie van SoZaWe een statistische procedure ontwikkeld voor het tellen van het aantal overtreden regels. Om een aantal redenen dient onderzoek te worden gedaan naar het probleem van de telvariabele:

- Telvariabelen zijn beleidsmatig van groot belang voor het beantwoorden van de vraag hoeveel regels per persoon zijn overtreden, en, hierdoor ook, hoeveel mensen minimaal één regel overtreden.
- Telvariabelen hebben over het algemeen een hogere psychometrische betrouwbaarheid en bieden meer mogelijkheden om relaties met andere variabelen vast te stellen. Afzonderlijke vragen hebben over het algemeen een lage betrouwbaarheid. Hierdoor is het soms lastig een relatie tussen zo’n losse vraag te ontdekken, terwijl die relatie er in de populatie wel is. Indien vragen hetzelfde meten, neemt de betrouwbaarheid van de somscore van de telvariabele toe.

- De problematiek rondom telvariabelen biedt de mogelijkheid om de onderschatting tengevolge van het overtreden van de spelregels van de RR-procedure te kwantificeren en/of te elimineren. Door dit onderzoek kan RR-onderzoek mogelijk meer valide uitkomsten opleveren.

De huidige methode van het construeren van een telvariabele leidt echter tot ongerijmde resultaten. In de 'kaartmethode' gehanteerd bij het Ministerie van SoZaWe bleek het aantal personen dat bij bijvoorbeeld drie items 'zwart-zwart-zwart' antwoordt groter te zijn dan volgens de onderzoeksopzet mag worden verwacht. Een mogelijke oorzaak van dit probleem is dat een groep personen de randomized respons procedure niet volgt ('*cheaters')*. Het gevolg is dat bij het gebruik van de ene schattingsmethode (zogenaamde 'momentschatters') negatieve schattingen ontstaan, en bij het gebruik van een andere methode ('maximum likelihood schatters') kansschattingen voor meerdere variabelen resulteren die niet te rijmen zijn met de kansschattingen voor de afzonderlijke variabelen. Een voorbeeld van dat laatste is een schatting waarbij 8% van de respondenten 3 regels zou overtreden, terwijl een of meer van de drie regels afzonderlijk door minder dan 8% van respondenten zou worden overtreden. Je zou verwachten dat als voor alle drie de regels 8 % wordt gevonden, elk van de drie afzonderlijke percentages minimaal 8 % is.

Te onderzoeken vragen en deelproblemen met betrekking tot de problematiek van de telvariabele zijn:

1. Speelt het probleem van negatieve kansen bij telvariabelen evenzeer bij andere randomised response methoden? Dit kan een criterium zijn om een bepaalde methode te prefereren boven een andere. Hiertoe dienen data uit verschillende onderzoeken die van verschillende RR-methoden (m.n. 'forced response methoden') gebruik maakten, opnieuw te worden geanalyseerd.
2. Wanneer zijn de waargenomen frequenties van ('zwart'-'zwart'-'zwart') zo groot dat deze de designkansen overstijgen en er geen sprake meer kan zijn van toeval, maar van 'cheaten'? Stel dat er geen enkele regelovertreder is. Bij gebruik van de speelkaartmethode verwacht je dan een proportie van .2 antwoorden 'rood' gemiddeld, omdat er in de 'nee-stapel' 20% rode kaarten zijn opgenomen. Door het toeval kan deze proportie voor een specifieke vraag lager zijn, want er kan toevallig vaker een rode kaart zijn getrokken in de linkerstapel. We stellen voor een norm te ontwikkelen waarmee we kunnen bepalen of de afwijking zo groot is, dat de waargenomen frequenties afwijken van toevalsfluctuaties en er wel sprake van 'cheaters' moet zijn.
3. Ontwikkeling van een techniek om te onderzoeken in hoeverre de variabelen die de telvariabele vormen, hetzelfde meten en een maat voor de betrouwbaarheid van die telvariabele en het gemeten 'construct' te verkrijgen. Telvariabelen hebben een verhoogde betrouwbaarheid (en daardoor een verhoogde power) indien de variabelen die de grondslag voor de telvariabele vormen, hetzelfde meten. In deze context zou een mogelijke invulling van dit "hetzelfde meten" iets kunnen zijn als "neiging tot regelovertreding".
4. Een volgende logische vraag is hoe verklarende variabelen gerelateerd zijn aan de telvariabele. We stellen voor regressiemodellen te ontwikkelen waarbij de telvariabele de afhankelijke variabele is. Het zgn. ordinale regressiemodel lijkt hiertoe het meest geëigend. Dit dient bekeken te worden voor verschillende randomized response methoden, bijvoorbeeld dobbelsteen-methode ('forced response'), en kaartmethode ('Kuk').
5. Ontwikkeling van een andere statistische methode voor het creëren van telvariabelen teneinde het probleem van negatieve kansen te vermijden. Momenteel worden telvariabelen gemaakt via het aantal gegeven 'zwart'-antwoorden. Dus, bij 3 regels zijn er 8 antwoorden en moeten er 8 regelnalevingscombinaties worden geschat (bijvoorbeeld "wel-niet-niet' is een van de acht). Omdat de interesse naar een telling uitgaat, lijkt het echter beter op basis van het aantal 'rood' antwoorden het aantal "wellen" te schatten. Bij drie regels zijn er vier kansen (0 keer "wel" tot drie keer "wel"), en dit vermindert het probleem dat kansen negatief zouden kunnen worden.
6. Ook dit dient bekeken te worden voor verschillende randomized response methoden, bijvoorbeeld dobbelsteen-methode ('forced response'), en kaartmethode ('Kuk'). Het schatten van de onder- en bovengrenzen van het percentage 'cheaters'. Deze grenzen geven aan hoe goed respondenten de RR-procedure hebben gevolgd, en geven hiermee een indicatie van de kwaliteit van de verzamelde data. Daarnaast krijgen we betere schattingen van het onderzochte verschijnsel, namelijk schattingen die ontdaan zijn van 'cheaters'.

Onderdeel 3: theorie en meting van achtergronden naleving en overtreding

Inleiding

Op basis van de kritische bespreking van de methodologie van de T11 is in het laatste hoofdstuk een aantal problemen geïdentificeerd en zijn voorstellen gedaan voor onderzoek en ontwikkeling. De T11 is een hulpmiddel waarmee men in het algemeen tracht na te gaan welke factoren van de perceptie van de regelgeving onder de beregelden samenhangen met (de neiging tot) overtreden. Tafel-van-Elf is eigenlijk meer de aanduiding van een gedachteschema om over handhaving en naleving te denken, dan dat het een concreet meetinstrument is[1]. Het is geïnspireerd vanuit het rationele mensbeeld zoals dat is neergezet door Becker in de economie en door Fishbein-Ajzen in de sociale psychologie. De T11 richt de aandacht op elf dimensies zoals de bekendheid en helderheid van de regels, de kosten en baten van naleven en overtreden, de acceptatie van regels, sociale normen, kans op verschillende manieren van ontdekking van overtreding en elementen van sanctie. De elf dimensies zijn eigenlijk niet meer dan labels voor op zich alweer ruime gebieden, en in veel T11 -toepassingen worden een of meer concepten weer op verscheidene manieren geoperationaliseerd.

De voorstellen voor onderzoek en ontwikkeling van materialen hebben betrekking op de volgende drie stappen bij het gebruik van de T11 :

1. Standaardisatie en codificatie op grond van de bestaande empirische T11-studies om zo tot een grotere eenvormigheid en eenduidigheid van terminologie en gebruik van de T11 te komen
2. De opzet van een T11–studie en de dataverzameling, met name de ontwikkeling van richtlijnen en hulpmiddelen voor operationalisatie van de verschillende dimensies.
3. De analyse en interpretatie van T11 – data.

Het grootste probleem bij het gebruik van de T11 is de onduidelijkheid die bestaat over hoe men een T11–profiel dient te interpreteren. In die zin zijn aanbevelingen 12 tot en met 14 hierbeneden de meest wezenlijke.

Standaardisatie en codificatie

Het is niet gemakkelijk toegang te krijgen tot de T11 literatuur, omdat het deels om niet openbaar gepubliceerde rapporten gaat. Dit leidt tot de *eerste aanbeveling* om ergens, bijvoorbeeld bij het Expertisecentrum Rechtshandhaving (ERH), een literatuurbank te onderhouden, waar alle betrokken rapporten, liefst ook elektronisch, voorhanden zijn en een up-to-date literatuurlijst wordt bijgehouden. Daarbij is ook een beschikbaar komen van de databestanden van de betreffende studies, in ieder geval van de doelgroepstudies, alleszins aan te raden.

De *aanbevelingen 2 tot en met 4* hebben betrekking op het ontwikkelen van een standaardbeschrijving van het meetinstrument en de dataverzameling, en van een handleiding voor analyse en interpretatie, met daarbij een beschrijving van specificatie, uitvouwing en condensatie van de dimensies en sub-dimensies van de T11. De aanleiding voor deze actie is de onhelderheid over wat een T11 –studie is en hoe de resultaten geanalyseerd, geïnterpreteerd en gepresenteerd kunnen worden. Het voorstel betreft onder meer ordening van bestaand werk en het doen van voorstellen voor een standaardterminologie. Voor het opstellen van een handleiding voor analyse en interpretatie dient te worden gewacht tot de aanbeveling 12 tot 14 zijn uitgevoerd.

De *tweede aanbeveling* is om een degelijke standaardbeschrijving te vervaardigen van wat exact verstaan moet worden onder “Tafel-van-Elf-onderzoek” , hetgeen in twee delen uiteenvalt: een beschrijving van het meetinstrument enerzijds, en een beschrijving van de analysemethoden anderzijds. De *derde aanbeveling* is dan ook om een heldere handleiding te ontwikkelen die analyse en interpretatie van de uitslagen van Tafel-van-Elf-onderzoek weergeeft.

Een T11 studie bestaat allereerst uit het vaststellen welke (sub)-dimensies, al of niet ontvouwd, worden meegenomen, onmiddellijk gevolgd door een operationalisatie van de gekozen sub-dimensies. Het maken van een keus uit de T11 –(sub)–dimensies is de fase van de specificatie en, omdat sommige dimensies van de T11 als het ware worden uitgepakt, ook wel uitvouwing van de T11. De ‘specifikatie-uitvouwing’–gedachte leidt tot de *vierde aanbeveling*: in een handleiding voor de methode van de T11 dient men expliciet aandacht te besteden aan de noodzaak dat elke toepassing aanvangt met reflectie op de voor die specifieke toepassing op te nemen dimensies, en op de mate waarin binnen dimensies nader onderscheid wordt gemaakt. Bij het opstellen van de handleiding dient ook expliciet aandacht te zijn voor de noodzaak of mogelijkheid tot condensatie van meerdere dimensies of sub-dimensies die sterk gecorreleerd zijn. Condensatie brengt mogelijk per te evalueren wet een andere groepering van dimensies met zich mee.

Dataverzameling: operationalisatie

De aanbevelingen 5 tot en met 7 betreffen onderzoek met betrekking tot meervoudige operationalisaties van T11-vragen om een hogere betrouwbaarheid te bewerkstelligen.

Aanbeveling 5 is om bij nieuwe studies systematisch meervoudige operationalisaties mee te nemen om vervolgens te analyseren in welke mate daarmee betrouwbaarheidswinst geboekt wordt. *Aanbeveling 6* is om een inventarisatie en analyse van bestaande meervoudige operationalisaties te verrichten om te bezien of deze aanpak een aanwijzing geeft over de mate van betrouwbaarheid.
De *zevende aanbeveling* heeft specifiek betrekking op de operationalisatie van kansvragen. Over de betrouwbaarheid daarvan is twijfel gerezen. Verschillende vraagformaten, bijvoorbeeld: 'hoe groot is die kans in procenten?' respectievelijk 'is die kans zo groot dat u daar rekening mee houdt?' blijken geenszins identieke uitkomsten op te leveren. De 'rekening houden met' – vragen blijken voor de respondenten erg onduidelijk. Ook categorische kansmeting (vijfpunts-schaal, heel klein – heel groot) en percentuele kansmeting (hoeveel procent) sporen nauwelijks. Vooral bij de wat grotere kansen blijken respondenten die aangaven een kans 'heel groot' te achten, daar –in procenten uitgedrukt– toch heel verschillende waarden mee te bedoelen.

Een kwalitatief onderzoek naar de formulering van de vragen in een cognitief laboratorium kan uitsluitsel bieden welk format de voorkeur verdient. Verder lijkt er wel aanleiding tot wat nader spitten in deze materie om te bezien of een secundaire analyse van de monitorstudie-kansvragen meer inzicht in het effect van verschillende vraagformaten oplevert. Zo kan worden beoordeeld of het advies bij de meting van T11 -kansdimensies tot één vraagformaat te beperken gerechtvaardigd is.

De achtste aanbeveling luidt om een onderzoek te doen naar de operationalisatie van conditionele gebeurtenissen. Nogal wat dimensies van de T11 adresseren conditionele gebeurtenissen. Bijna altijd gaan de vragen over wat er zou gebeuren indien iemand, of de respondent zelf, de regels wel of niet zou hebben overtreden. Soms is die conditionaliteit nog ingewikkelder. Vanuit het 'objectieve' perspectief van de controle-experts is een ontleding van de controle in de trits [controlekans / pakkans bij controle / sanctiekans bij betrapping] zinvol. Er is echter twijfel aan de begrijpelijkheid voor respondenten van meervoudig samengestelde vragen met hypothetische gebeurtenissen. De vraag hierbij is of en in welke mate respondenten in staat zijn de verscheidene stappen in het controleproces te onderscheiden, te interpreteren zoals bedoeld, en daar een duidelijke mening over te geven. In een cognitief laboratorium kan dit vraagstuk empirisch worden ontleed.

Bij problemen zijn twee opties het overwegen waard. Een is om de kwaliteit van de conditionele vragen belangrijk te verbeteren. Dat zal uitdraaien op omvangrijker vragenlijsten en meer gecompliceerde vragen. De andere optie is afzien van de pretentie om het controleproces zo sterk in stappen uiteen te gooien. Ten aanzien van de doelgroep is het vaak genoeg te weten of de controle bedreigend is, en daarbij is een 'pakket-schatting' voldoende. Dat sluit ook aan bij de elders gemaakte observatie dat condensatie veelal tot clustering van de hele formele controle - en sanctiebatterij leidt.

Als *negende aanbeveling* stel ik voor over deze optie in een studieconferentie van T11 – experts serieus van gedachten te wisselen. Uiteraard gaat het dan om een majeure herstructurering van de T11 –benadering voor doelgroepensurveys.

Analyse en interpretatie

Aanbeveling 10: Het neoklassieke rationele keuzemodel stelt dat afschrikking gemeten kan worden als verwachte strafzwaarte. Bijvoorbeeld: de kans op een sanctie wordt vermenigvuldigd met de ernst van de sanctie (volgens de respondent) om tot een nieuwe variabele 'gewicht van de sanctie' te komen.
Op voorhand lijkt twijfel op zijn plaats, zowel vanwege de kwaliteit van de metingen als vanwege het conceptueel model. Ook conceptueel lijkt het neoklassieke model te simpel te zijn. Volgens nieuwere economisch-psychologische inzichten blijken mensen in hun omgang met kansen zich niet conform dit model te gedragen. De aanbeveling is dan ook om zich de moeite te besparen.

Aanbeveling 11 heeft betrekking op de mogelijkheden tot wetsonafhankelijke condensatie van dimensies. In de praktijk van de condensatie-analyse blijkt er vaak een min of meer overeenkomende groepering van dimensies plaats te vinden die geldt voor meerdere wetten. Een secundaire analyse van verschillende beschikbare datasets is te overwegen. Deze moet meer inzicht geven in de vraag of er een kern van gecondenseerde T11 –dimensies bestaat die voor een grote groep regelingen en wetten nagenoeg identiek is.

De *twaalfde en dertiende aanbeveling* hebben betrekking op de constructie, presentatie en interpretatie van een T11-profiel op. Het gaat daarbij om een op zichzelf staand proffiel, afgezien van eventuele samenhangen met regelovertreding. Aanleiding voor de aanbevelingen is het feit dat voor deze taken geen uitgekristalliseerde methodologie bestaat. Vaak wordt een lijngrafiek van gemiddelden gepresenteerd, al of niet met standaarddeviaties geprojecteerd in de grafiek. In andere gevallen wordt een gestapeld staafdiagram weergegeven met percentages van de respondenten die bepaalde antwoordopties hebben gekozen. Naast een veelheid van presentatiestijlen komen we ook verschillende standaarden tegen van welke scores men nu als wel of niet kwetsbaar moet beschouwen.

Er dient met conceptuele verheldering en secundaire analyse verder gewerkt te worden aan analysemethoden die in de literatuur zijn voorgesteld om de interpretatie van T11-profielen te vergemakkelijken. *Aanbeveling 12* is te proberen de conceptuele verheldering te verkrijgen door aan eventuele voorstellen tot interpretatie van de T11 een studieconferentie met T11 –experts te wijden. De *dertiende aanbeveling* is een secundaire analyse van enkele eerdere studies te ondernemen met behulp van de uitbijtende groepenanalyse en de comparatieve profielenanalyse, om te bezien of die aanpak er inderdaad toe leidt dat interpretatie wordt vergemakkelijkt. Met de analyse van uitbijtende groepen is het mogelijk deelgroepen in de doelgroep te identificeren die op en dimensie een duidelijk afwijkend gemiddelde score hebben. Enigszins in de zelfde lijn ligt de mogelijkheid om comparatieve profielen te maken, waarbij deelgroepen qua T11-profiel worden vergeleken. Als voorbeeld kan men dan denken aan: mensen die ooit gecontroleerd zijn versus mensen die dat niet zijn, of bestraften tegen niet-gestraften, nieuwkomers vs. oude rotten onder de regeling.

Het lijkt voor de hand te liggen aan eventuele voorstellen tot interpretatie van de T11 een studieconferentie met T11 –experts te wijden. Uiteraard dient dit dan te worden gecombineerd met de bij aanbeveling 9 voorgestelde bijeenkomst, die de inkrimping van de tafel tot onderwerp heeft.

Aanbeveling 14 betreft de ontwikkeling van een methode voor de analyse van het verband tussen T11 variabelen en naleving. De grond voor deze aanbeveling is gelegen in problemen met de interpretatie van het verband tussen de T11 en naleving. Bijvoorbeeld, als een T11 –dimensie nagenoeg geen variantie heeft, kan hij geen correlatie vertonen met naleving. Ook als de spreiding op een dimensie zich vrijwel geheel aan een zijde van de kwetsbaarheidgrens bevindt, zal er geen correlatie optreden.

Nodig hiervoor zijn theoretisch werk en de secundaire analyse van bestaand datamateriaal, met name bij randomised response methoden. Daarbij wordt gekeken of loglineaire analyse van het verband tussen nalevingsprofiel en nalevingsrapportage haalbaar en inzichtgevend is, en of bestaande methoden van automatische interactiedetectie daarbij zinvol zijn.

1. Inleiding

Binnen de Rijksoverheid wordt op verschillende plaatsen systematisch onderzoek gedaan naar naleving van regels door het bevragen van (potentiële) overtreders. Het Expertisecentrum RechtsHandhaving (ERH) van het Ministerie van Justitie richt zich op het ontwikkelen van instrumentarium voor dit type onderzoek. Door de overheid, o.a. in het regeringsstandpunt 'Handhaven op niveau', is de behoefte geformuleerd aan een toegepaste nalevingsmonitor voor een groot aantal te vergelijken wetten.

De reeds uitgevoerde nalevingsonderzoeken hebben zich op twee beloftevolle onderzoeksmethoden geconcentreerd, te weten de zogenaamde 'Tafel van 11' voor het meten van motieven, kennis en attituden die in verband staan met naleving van de te onderzoeken regel, en het meten van de regelovertreding met 'randomized response technieken'. Met beide methoden zijn in de uitgevoerde onderzoeken positieve ervaringen opgedaan, maar er blijven nog enkele belangrijke vragen te beantwoorden, om de onderzoeksmethoden nog geschikter te maken voor toepassing. Dit rapport beoogt aan te geven welke kennis nog ontbreekt en hoe die kan worden aangevuld om dit streven te realiseren.

De 'Tafel van 11' is een instrument dat gebruikt wordt om naleving en overtreding van overheidsregels te verklaren. Het is geïnspireerd vanuit het rationele mensbeeld zoals dat is neergezet door Becker in de economie en door Fishbein-Ajzen in de sociale psychologie. De T^{11} bevat elf dimensies, T_1 tot en met T_{11}, onderverdeeld in twee groepen. De eerste groep, bestaande uit de factoren T_1 tot en met T_5, bestaat uit factoren die niet direct door de handhavinginstantie zijn te beïnvloeden. Zij worden *spontane nalevingsfactoren genoemd*. De tweede groep factoren. T_6 tot en met T_{11} hebben betrekking op de activiteiten van de handhavinginstantie en heten *afgeleide nalevingsfactoren*.

Randomized response is een manier om tijdens een interview gevoelige vragen te stellen. Om de privacy van de geïnterviewde te verzekeren heeft Warner (1965) een methode ontwikkeld die deze privacy volledig beschermt, om zo de weerstand van geïnterviewden te verminderen om informatie te verstrekken over sensitieve onderwerpen. Sindsdien zijn meerdere vormen van deze onderzoekstechniek ontwikkeld. Allemaal hebben ze gemeen dat aan de sensitieve vraag een kanselement wordt toegevoegd, waardoor de interviewer niet weet welke vraag wordt beantwoord en de privacy van de respondent volledig wordt beschermd.
Tot dusver zijn in Nederland drie vormen van de randomized response techniek in onderzoek gebruikt:

- De 'forced response methode': de geïnterviewde gooit twee dobbelstenen, zegt, ongeacht het ware antwoord 'ja' als de som gelijk is aan 2, 3 of 4, 'nee' als de som gelijk is aan 11 of 12, en geeft eerlijk antwoord op de sensitieve vraag als de som ligt tussen de 5 en 10.
- Het 'unrelated question design': de geïnterviewde gooit een dobbelsteen. Als de worp 1 of 2 is, beantwoordt hij of zij een neutrale vraag, bij een uitkomst tussen de 3 en 6 beantwoordt de respondent de gevoelige vraag.
- De 'kaartenmethode' De geïnterviewde heeft twee stapels kaarten voor zich, met in de linkerstapel 80% rode kaarten en 20% zwarte en in de rechterstapel 20% rode kaarten en 80% zwarte. Voordat de vraag wordt gesteld dient de respondent, buiten het zicht van de enquêteur, uit beide stapels een kaart te trekken. Vervolgens wordt de vraag gesteld en noemt de respondent de kleur van de linkerkaart als het antwoord op de vraag 'ja' is en de kleur van de rechterkaart als het antwoord op de gevoelige vraag 'nee' is.

Er bestaan nog meer varianten van de randomized response methode, maar uit de literatuur blijkt dat voornoemde methoden 92% uitmaken van het totale randomized response onderzoek tussen 1975 en 2000 (Lensvelt-Mulders and Hox, 2000).

Wij hebben een studie verricht om te bepalen welke kennis over randomized response en de T^{11} momenteel beschikbaar is en welke kennis ontbreekt. Op grond van deze studie formuleren wij onderzoeksvragen voor een onderzoeksprogramma om de ontbrekende kennis aan te vullen. Verder geven wij ook kort aan op met welke vormen van onderzoek de gestelde vragen te beantwoorden zijn. De verschillende onderzoeksvragen kunnen vervolgens worden uitgezet in de academische wereld.

Dit rapport heeft betrekking op drie belangrijke onderdelen van onderzoek naar naleving en overtreding van overheidsregels. Deze onderdelen worden in drie opeenvolgende hoofdstukken van het rapport door verschillende auteurs behandeld. Het betreft:

1. dataverzameling met behulp van randomized response en eventueel met andere technieken om de vertrouwelijkheid van gevoelige informatie te verhogen wordt beschreven in hoofdstuk 2 van Joop Hox. Gerty Lensvelt, Marriane Junger en Edith de Leeuw;
2. de analyse van data verzameld met behulp van randomized response in hoofdstuk 3 door Peter van der Heijden en Ardo van den Hout;
3. de theorie en meting van achtergronden van naleving en overtreding, vervat in de T^{11}, komt aan bod in hoofdstuk 4 van de hand van Henk Elffers.

Wij zullen in het navolgende de kennis en vragen met betrekking tot deze onderdelen bespreken. Alle besproken onderzoeksvoorstellen gaan uit van de wens de nalevingsmonitor en meer in het algemeen onderzoek met randomized response en de T^{11} betrouwbaarder, meer valide, beter te analyseren en toch goedkoper en eenvoudiger in het gebruik te maken.

2. Dataverzameling

Gerty Lensvelt, Edith de Leeuw, Joop Hox,
Marianne Junger

Universiteit Utrecht

2.1 Inleiding

Wat is nodig voor valide survey-onderzoek?

Valide survey-onderzoek dient aan een aantal methodologische eisen te voelden. Centraal staat het minimaliseren van vier potentiële fouten: dekkingsfouten, steekproeffouten, non-responsfouten, en meetfouten.

Dekkingsfouten ontstaan wanneer een bepaalde groep uitgesloten wordt van het onderzoek, dus geen kans heeft om in de steekproef te worden opgenomen. Een voorbeeld is geheime nummers bij een steekproef uit het telefoonboek. Wanneer mensen met een geheim nummer verschillen van mensen met een 'gewoon' telefoonnummer dan kan vertekening optreden. Bij het meten van regelnaleving, is het belangrijk om na te gaan dat er geen dekkingsfouten optreden met betrekking tot de 'overtreders'.

Steekproeffouten ontstaan doordat niet iedereen in de populatie wordt ondervraagd, maar een kleinere steekproef. Wanneer de populatie goed omschreven is en wanneer de steekproef aselect uit deze populatie getrokken wordt, dan kan de steekproeffout via statistiek berekend worden. Bij de toepassing van randomized response wordt de steekproeffout groter doordat er ruis aan de data wordt toegevoegd. Vanwege de efficiency van de dataverzameling is het belangrijk deze extra fout zo klein mogelijk te houden.

Non-respons ontstaat wanneer een bepaalde groep potentiële respondenten de vragen niet beantwoord, omdat ze het niet willen en weigeren, of niet kunnen omdat ze bijvoorbeeld de taal niet goed spreken. Als de non-respondenten op de belangrijkste doelvariabelen sterk afwijken van de respondenten kan serieuze vertekening optreden. Wanneer toepassing van randomized response tot minder weigeringen leidt, wordt de non-responsfout verminderd.

Tenslotte zijn er de *meetfouten*: het antwoord dat gegeven wordt wijkt af van het 'ware' antwoord. Dit kan samenhangen met het stellen van complexe vragen.

Bovengenoemde fouten vormen een bedreiging van de validiteit van ieder onderzoek, maar onderzoek naar gevoelige vragen is extra gevoelig voor potentiële bedreiging van de validiteit. Hieronder wordt daarop nader ingegaan.

Wat is speciaal aan gevoelige vragen?

Onderzoeksvragen kunnen op twee manieren bedreigend voor respondenten zijn. Ze kunnen een *intrinsieke* of een *extrinsieke* dreiging vormen (Lee, 1993).

Er is sprake van een *intrinsieke dreiging* wanneer de vragen betrekking hebben op onderwerpen die zeer persoonlijk zijn, pijnlijk zijn, of stressvol. De dreiging komt dan vanuit de respondent zelf. Voorbeelden hiervan zijn vragen naar echtscheiding, dood van een ouder, etc. Ook vragen die respondenten dwingen om hun zelfbeeld negatief bij te stellen, worden als bedreigend ervaren. Soms kan het voor respondenten pijnlijk of stressvol zijn om 'tegenover zichzelf' toe te geven dat zij zich niet gedragen zoals ze eigenlijk volgens zichzelf

zouden moeten. Er is dan sprake van internalisatie van sociale normen. Zo zal regelovertredend gedrag in het algemeen ondergerapporteerd worden.

Er is sprake van *extrinsieke dreiging* als de antwoorden sancties kunnen oproepen. Dit kunnen sociale sancties zijn. Wanneer er sprake is van strafbaar gedrag, dan kunnen de vragen bedreigend zijn omdat eerlijk antwoorden strafrechtelijke consequenties kan hebben. De kern van extrinsieke dreiging ligt in de angst van de respondent dat de antwoorden bekend worden of zelfs doorgegeven worden aan officiële instanties.

Bij onderzoek naar het naleven van regels onder potentiële overtreders (meten van regelnaleving) heeft de onderzoeker zowel te maken met intrinsieke dreiging als extrinsieke dreiging. Bij het beschouwen van alternatieven voor of optimalisaties van randomized response methoden dienen dan ook beide vormen van bedreiging in beschouwing te worden genomen.

Validiteit van zelfrapportage van delinquentie

Zelfrapportage van delinquentie is inmiddels een courante onderzoeksmethode. Het ligt echter niet zonder meer voor de hand dat respondenten bereid zijn informatie over gevoelige onderwerpen, zoals delicten, te rapporteren tijdens een interview. Wanneer men naar globale schattingen kijkt dan is ongeveer 2/3 van de respondenten waarvan bekend is dat zij contacten hadden met politie/justitie bereid informatie te geven over delinquent gedrag. De conclusie in de literatuur is daarom dat zelfrapportage van delinquentie in het algemeen als redelijk valide kan worden beschouwd. Een belangrijke vraag is of er echter of deze globale conclusie enige bijstelling behoeft als de validiteit met meer precisie wordt onderzocht. Een potentieel probleem hierbij is de *Differentiële validiteit*.

Bij *differentiële validiteit* gaat het om de vraag in hoeverre de validiteit van zelfrapportage gelijk is voor afzonderlijke categorieën respondenten. Bijvoorbeeld, zijn mannen en vrouwen even open tijdens een interview over door hen gepleegde delicten? Indien mannen gemakkelijker delicten toegeven tijdens een interview dan vrouwen, dan kan niet zonder meer worden geconcludeerd - op basis van zelfrapportage - dat vrouwen minder delicten plegen. Differentiële validiteit is daarmee een belangrijke bedreiging voor het gebruik van zelfrapportage, zeker wanneer men wil *beschrijven* hoe het met de betrokkenheid bij criminaliteit is gesteld van diverse demografische groepen. Dit geldt temeer wanneer de validiteit van zelfrapportage samenhangt met het onderwerp van onderzoek, namelijk criminaliteit. De vraag daarbij is of respondenten met een sterke betrokkenheid bij criminaliteit meer of minder geneigd zijn correct antwoord te geven dan respondenten die nagenoeg geen delicten plegen?

Op het gebied van de differentiële validiteit is helaas niet zeer veel onderzoek voorradig. Er is echter wel enig onderzoek verricht.

Leeftijd en sekse lijken geen zeer belangrijke variabelen bij differentiële validiteit. De belangrijkste factor bij differentiële validiteit is vermoedelijk etniciteit. In de enige Nederlandse studie naar de validiteit van zelfrapportage onder verschillende etnische groepen zijn relatief grote verschillen gevonden (Junger, 1989). In deze studie is, bij jongens met officieel geregistreerde politiecontacten, onderzocht wat hun antwoorden waren op vragen naar delinquent gedrag en politiecontacten. Onder Nederlandse en Surinaamse jongens bleek 13% van de jongens geen enkel delict te vermelden terwijl zij een geregistreerde politiecontact hadden, en onder Marokkaanse en Turkse jongens was dit respectievelijk 37% en 44%. Omdat Turkse en Marokkaanse jongens, d.w.z. jongens met een Noord-Afrikaanse en/of Islamitische achtergrond minder delicten rapporteerden dan groepen van Centraal-Afrikaanse of Aziatische afkomst, lijkt het erop dat, in Nederland, met betrekking tot zelfrapportage, de culturele factor belangrijker is dan de etnische factor.

Ook Amerikaanse studies vonden etnische verschillen in de validiteit van zelfrapportage van misdrijven. Tenminste drie grote studies vermeldden dat respondenten van Afrikaanse afkomst minder geneigd waren delicten toe te geven dan respondenten van andere etnische origine (zie Hindelang et al., 1981; Maxfield et al., 2000; Thornberry & Krohn, 2000). Er zijn echter ook studies die geen verschillen vonden tussen etnische groepen (Farrington et al., 1996; Junger-Tas & Haen-Marshall, 1999).

Wanneer men de onderzoeksresultaten in samenhang beschouwt lijken er twee antwoordtendensen te zijn.

- Over sommige zaken schaamt men zich en die verzwijgt men liever. Zeer afkeurenswaardige misdrijven en recente feiten rapporteren respondenten minder vaak dan minder laakbare feiten en oudere feiten. Intrinsieke dreiging speelt hier kennelijk een rol.
- De meest deviante respondenten, dit wil zeggen diegenen die de meeste delicten plegen (volgens de officiële registraties) zijn het meest open tijdens interviews en zij lijken dus het minst last te hebben van schaamte.

De studie van Junger (1989) ondersteunt deze laatste stelling. Hieruit kwam naar voren dat de meest problematische respondenten, namelijk diegenen die relatief tolerant stonden ten opzichte van het plegen van delicten en die zich weinig gebonden voelden aan traditionele waarden en tradities, de meeste delicten toegaven tijdens een interview. Meer conventionele jongeren waren geneigd vaker delicten te verzwijgen (ook al hadden zij een geregistreerd politiecontact).

Petersilia (1978) gaf hiervoor de volgende verklaring. Verwijzend naar Teilmann (1976): *'The answer may be that most people are not willing to lie about unambiguous facts. However, behaviours which are ambiguous as to their definition of offences, and which are engaged in frequently so that their number is difficult to remember, may well be subject to self-enhancing definitions in threatening situations. The occasions in which a person takes money from someone by force are probably remembered fairly clearly. To fail to report these offences would require a knowing outright lie on the part of the respondents' (Teilmann, 1976).* Petersilia concludeert daarom (1978) dat '*... the importance of an event to an individual and its integration with other life events affect his reporting of them.*'

De implicatie hiervan is dat - in termen van schaamte - respondenten het enerzijds moeilijk vinden om zeer laakbare zaken te vermelden, maar het anderzijds ook moeilijk vinden om antwoorden te geven die ver bezijden de waarheid liggen. Respondenten kunnen het voor zichzelf verantwoorden om een deel van hun verhaal te verzwijgen maar vinden het klaarblijkelijk moeilijk om grote delen ervan te achter te houden.

Samenvattend en vooruitblik

De conclusie uit het voorafgaande is dat zelfrapportage van delinquentie in principe valide gegevens kan opleveren. Wanneer respondenten bereid zijn om aan een interview mee te doen, zijn zij in principe ook bereid sensitieve vragen te beantwoorden. Om de kwaliteit van de gegevens te optimaliseren zijn de standaard kwaliteitseisen van survey-onderzoek van kracht, met name die maatregelen waarvan doorgaans wordt aangenomen dat zij het beantwoorden van sensitieve vragen bevorderen. Daarnaast zijn specifieke maatregelen belangrijk die de intrinsieke en extrinsieke dreiging verminderen. Randomized response is één techniek hiervoor, er zijn echter alternatieven. Hieronder bespreken we drie gebieden waarop nader methodologisch onderzoek naar het verzamelen van sensitieve informatie wenselijk is.

1. Allereerst zijn dat de mogelijke alternatieven voor randomized response. Dit is zinvol omdat randomized response wat betreft de steekproeffout inefficiënt is, zodat veel grotere steekproeven nodig zijn om dezelfde nauwkeurigheid te bereiken.
2. Vervolgens worden de mogelijkheden bekeken voor alternatieven binnen de randomized response. De nadruk ligt daarbij enerzijds op het verhogen van de efficiëntie, en anderzijds op het inschakelen van de computer bij de dataverzameling. Wat betreft de dataverzameling zou de kennis die in deze twee methodologische onderzoekslijnen wordt verzameld gebundeld dienen te worden in ‘current best methods’ handleiding.
3. Tenslotte is er ook gekeken naar ‘slimme onderzoekopzetten’, een aantal mogelijke verbeteringen die niet onder de twee eerdere gebieden vallen. Deze deelstudie heeft geen concrete onderzoeksvoorstellen opgeleverd.

Onderzoeksvoorstellen

Bij onderzoek naar gevoelige vragen in het algemeen dient men een situatie te creëren waarin de respondent zich zo goed mogelijk voelt. Algemene aanbevelingen uit de literatuur hierover omvatten o.a. (zie De Leeuw, 1999): uitleg aan de respondent waarom deze vragen gesteld worden (nut, doel), specifieke eisen aan de vraagverwoording en antwoordcategorieën, privacygarantie, en een sterk respondentgericht enquêteontwerp.

Het is zinvol om op basis van een literatuuronderzoek binnen de sociaal-wetenschappelijke en criminologische onderzoeksliteratuur een ‘current best methods’ handleiding op te stellen voor het uitvoeren van enquêteonderzoek naar sensitieve vragen, in het bijzonder vragen naar regeloverschrijdend gedrag. Een current best methods handleiding dient te voldoen aan de volgende voorwaarden: 1) gebaseerd op wetenschappelijke kennis, 2) concreet genoeg dat het als checklist gebruikt kan worden om de kwaliteit van het onderzoek te waarborgen. Hiervoor hoeft geen nieuw empirisch onderzoek te worden uitgevoerd, maar wel een grondig literatuuronderzoek te worden uitgevoerd. Daarbij moet met name de grijze literatuur worden meegenomen, omdat de methodologische verantwoording van veel evaluatie- en beleidsonderzoek d.m.v. interne rapporten wordt gerapporteerd. Voorbeelden zijn rapporten van onderzoeken naar sensitieve informatie uitgevoerd door de Rand Corp. (prostitutie) en Research Triangle Institute (RTI, drugs, alcohol, HIV-risico).

Een ‘current best methods’ handleiding heeft als doel het verzekeren van de kwaliteit van het proces van gegevensverzameling. Daarnaast kan deze tevens functioneren als een pakket van eisen dat bij het uitzetten van onderzoek wordt gehanteerd.

2.2 Alternatieven voor randomized response technieken

Stand van zaken

Uit vergelijkend onderzoek naar dataverzamelingsmethoden is gebleken dat men bij schriftelijke vragenlijsten opener en eerlijker antwoord geeft op gevoelige vragen. Voor een overzicht zie De Leeuw (1992). Wanneer het niet mogelijk is een schriftelijke vragenlijst of postenquête te gebruiken, dient het interview zodanig te zijn ingericht dat dit voor de geïnterviewde zo min mogelijk stress geeft. Het spreekt vanzelf dat men bij een interview de aanwezigheid van derden dient te vermijden. Maar ook de interviewer zelf is een toehoorder, die de privacy kan aantasten. Daarom is het belangrijk om bij gevoelige vragen andere methoden binnen het interview toe te passen die de privacy van de respondent zo veel mogelijk garanderen. Men noemt dit wel een 'mixed-mode' benadering. Een veel gebruikte en succesvolle methode is om tijdens het interview de respondent zelf een korte vragenlijst te laten invullen en deze daarna in een gesloten en dichtgeplakte envelop aan de interviewer mee te geven. Wanneer er gebruik wordt gemaakt van computergestuurd interviewen, kan de interviewer aan de respondent vragen om zelf achter het toetsenbord plaats te nemen. Uit vergelijkend onderzoek blijkt dat een volledig computergestuurd zelfinterview nog iets meer openheid in de antwoorden oplevert dan een door de respondent zelf ingevulde papieren vragenlijst (De Leeuw, Hox, & Snijkers, 1995; Weisband & Kiesler, 1996; De Leeuw, 2001). Hierbij is het van belang dat *in de optiek van de respondent* de antwoorden volstrekt vertrouwelijk zijn. Zo vond Beebe (1998) dat in een onderzoek naar drugsgebruik via CASI (Computer Assisted Self Interview) de afstand tussen computers in een computerruimte van cruciaal belang was voor de openheid in antwoord. Wanneer computers te dicht bij elkaar staan, of wanneer de mogelijkheid bestaat dat anderen meekijken, wordt het gevoel van privacy van de respondent aangetast.

Functioneel analfabetisme kan de bruikbaarheid van CASI voor specifieke doelgroepen beperken. Ook bij lager opgeleiden en/ of respondenten waarbij Nederlands niet de moedertaal is kunnen leesproblemen de meetfout vergroten en zo de validiteit van de antwoorden aantasten. Speciaal voor deze groepen is A-CASI ontworpen: Audio-Computer Assisted Self Interview. In deze toepassing hoort de respondent via een koptelefoon die vraag en mogelijke antwoordcategorieën op het moment dat deze ook op het scherm verschijnt. Met name in de USA is A-CASI een standaardmethode voor gevoelige vragen geworden.
Hoewel het gebruik van (A)-CASI surveyfouten sterk reduceert zijn deze zeker niet nul, er is nog veel ruimte voor verbetering van de accuraatheid. In de USA werkt een onderzoeksgroep aan de technische verbetering van A-CASI (Thornberry, Bhaskar, Krulewitch, Wesley, Hubbard, Adamson, 2000), maar aan de methodologische optimalisatie van de effectiviteit wordt weinig aandacht besteed. Onderzoek naar het verder reduceren van de meetfouten bij efficiënte technieken zoals (A)-CASI is aan te bevelen. Een veelbelovende invalshoek is het implementeren van ‘randomized response technieken’ binnen (A)-CASI.

Desiderata

Er is weinig onderzoek waarin de verschillende methoden tegen elkaar worden afgezet. Dat randomized response beter is dan een directe vraag is duidelijk. In hoeverre randomized response beter is dan andere methoden die de intrinsieke en extrinsieke dreiging eveneens verminderen is niet duidelijk. Er is slechts één onderzoek (Van der Heijden et al.), 2000) waarin RRT vergeleken wordt met CASI. In dat onderzoek komt CASI er niet goed van af. De bevinding in dat onderzoek dat CASI slechter werkt dan een directe vraag is echter geheel in tegenspraak met de meta-analyse van Weisband en Kiessler (1996), hetgeen twijfel oproept of in dat onderzoek CASI wel optimaal is geïmplementeerd. Zoals hierboven is uiteengezet, blijkt A-CASI gevoelig voor ogenschijnlijke kleinigheden in de onderzoeksopzet, zoals de afstand tussen twee gelijktijdige respondenten. Wanneer een goed geïmplementeerde A-CASI de nauwkeurigheid van randomized response blijkt te benaderen, zou de grotere statistische efficiëntie van A-CASI een voordeel zijn voor deze methode.

Onderzoeksvoorstel

Wij stellen voor een vergelijkend onderzoek uit te voeren waarin serieuze concurrenten voor de RR techniek zorgvuldig worden geïmplementeerd en de resultaten worden vergeleken, bij voorkeur toegepast op een vraagstelling binnen criminologisch regelovertredend gedrag. Serieuze concurrenten die statistisch efficiënter zijn, zijn de CASI of zelfs Audio-Casi techniek in een face-to-face interview, en het gebruik van

een telepanel. Voorafgaand aan een (A-)CASI onderzoek zou een zorgvuldige ontwerpstudie gemaakt moeten worden. De ontwerpstudie dient op basis van elders opgedane ervaringen randvoorwaarden op te stellen voor het succesvol implementeren van (A-)CASI bij gevoelige vragen en speciale groepen, en vervolgens met cognitieve labs het protocol te optimaliseren.

2.3. Alternatieven binnen randomized response technieken

Bij zeer gevoelige onderwerpen kan gebruik gemaakt worden van meer gecompliceerde indirecte technieken zoals de randomized response. Voor het succesvol gebruiken van randomized response technieken is het van belang dat de methode helder en duidelijk aan de respondent wordt uitgelegd. De respondent moet de techniek goed kunnen toepassen, én er van overtuigd zijn dat deze ook inderdaad de vertrouwelijkheid bevordert. Sinds Warner's publicatie over de RRT zijn er meer dan 300 artikelen verschenen. Daarin worden o.a. aanpassingen in Warner's RRT beschreven die de methode statistisch efficiënter maken. Of de RRT naast statistisch efficiënt ook zuiver is, hangt af van de validiteit van de resultaten, en die hangt af van de bereidheid van de respondenten om de procedure daadwerkelijk te volgen. Sinds 1965 is er een fors aantal variaties op Warner's methode beschreven. Bij al die variaties hebben we voorbeelden van toegepast onderzoek gevonden, maar het is opvallend onduidelijk op welke gronden onderzoekers voor een bepaalde RRT kiezen. Dit wekt verbazing, omdat voor alle vormen van de RRT extra grote steekproeven nodig zijn om gelijke betrouwbaarheidsintervallen te krijgen als bij directe vragen, iets dat zich vertaald in extra kosten voor materiaal, inzet van mensen en doorlooptijd. Deze extra kosten zijn niet voor elke RRT gelijk.

Meta-analyse van randomized response onderzoek wijst uit dat randomized response technieken meer valide schatters opleveren dan meer standaard ondervragingstechnieken. Dit is speciaal het geval naarmate het onderwerp meer sociaal gevoelig of bedreigend is (Lensvelt-Mulders, Hox en Van der Heijden, submitted). Deze effecten zijn statistisch significant (niet aan toeval toe te schrijven) maar niet zeer groot. Verder is gevonden dat een groot deel van de variatie tussen RRT-studies zich niet laat verklaren door variabelen als de kwaliteit van het onderzoek, de gebruikte randomizers, verschillen in de kans op de gevoelige vraag en aanpassingen in het aantal respondenten. Dat betekent in onze ogen dat het gebruik van de randomized response techniek in toegepast onderzoek nog onvoldoende onder controle van de onderzoeker is. Er gebeuren tijdens het onderzoeksproces teveel dingen die de onderzoekers nog niet onder controle hebben.

Sinds Warner (1965) de randomized response techniek introduceerde is een groot aantal varianten ontwikkeld. Een literatuurstudie van toegepast randomized response onderzoek tussen 1975 en 2000 toont echter aan dat in 92% van het toegepaste onderzoek gebruik wordt gemaakt van slechts drie methoden:
1. De forced response methode waarbij de geïnterviewde gooit met twee dobbelstenen. Men wordt 'geforceerd' *ja* te zeggen als de som van de worp 2, 3, of 4 is, *nee* te zeggen als de som van de worp 11 of 12 is en een eerlijk antwoord te geven op de gestelde vraag als de som van de worp tussen 5 en 10 is.
2. Het unrelated question design lijkt sterk op het forced response design, met dien verstande dat nu een neutrale vraag wordt beantwoord als de som van een worp 2, 3, 4, 11, of 12 is en een eerlijk antwoord te geven als de som van de worp tussen 5 en 10 is. De unrelated question techniek kent twee vormen, één waarbij het voorkomen van de neutrale vraag in de populatie niet bekend is (werkt u als vrijwilliger?), en één waarbij dit wel tevoren bekend is (is uw moeder jarig tussen 1 januari en 1 april?). Deze laatste vorm is het meest toegepast en is technisch gezien een vorm van de Forced Response methode.
3. Warner's originele methode en de kaartenmethode van Kuk. Deze laatste methode is statistisch gelijk aan Warner's oorspronkelijke randomized response techniek, maar psychologisch sterker. De respondent heeft twee stapels kaarten voor zich liggen, een *ja* en een *nee* stapel. De ja-stapel bevat 80% rode kaarten en de nee-stapel bevat 20% rode kaarten. Voordat de vraag gesteld wordt trekt de respondent, buiten het zicht van de enquêteur, uit elke stapel een kleur en hij noemt de kleur die overeenkomt met de stapel van het juiste antwoord.

Op dit moment is het mogelijk om met behulp van randomized response technieken de kans op gevoelig gedrag te schatten, de relatie tussen randomized response variabelen en verklarende variabelen te leggen d.m.v. logistische regressie, onderzoek te doen naar relaties tussen randomized response vairabelen onderling en, op overigens nog suboptimale wijze, telvariabelen te construeren.

In de volgende paragrafen beschrijven wij de resultaten van een korte verkenning naar allereerst de efficiency en vervolgens de effectiviteit van RRT's. Daarna onderzoeken we de mogelijkheid de RRT op te nemen in een computergestuurde dataverzameling. en leiden daaruit een aantal onderzoeksrichtingen af, die kunnen bijdragen aan het verbeteren van de toepasbaarheid van randomized response onderzoek in de beleidscontext.

Efficiëntie van verschillende randomized response methode t.o.v. directe vragen.

In het onderstaande wordt het begrip ‘efficiëntie’ opgevat als de (optimale) verhouding tussen de inbreng aan tijd, geld en respondenten en de betrouwbaarheid van het gevonden resultaat. De betrouwbaarheid van het resultaat wordt bepaald door de steekproefvariantie. Hoe groter de steekproefvariantie, des te groter het aantal respondenten dat nodig is om statistisch verantwoorde betrouwbaarheidsintervallen te verkrijgen. Het aantal respondenten is uiteindelijk bepalend voor de kosten voor afname (materiaal en personeel), de doorlooptijd, en de omvang van de data-invoer. De relatieve efficiëntie van onderzoeksmethoden ten opzichte van elkaar wordt berekend als de ratio van de steekproefvarianties van beide methoden.

Bij randomized response methoden wordt de efficiëntie van elke afzonderlijke techniek vooral bepaald door de waarschijnlijkheid waarmee het te onderzoeken gedrag voorkomt in de populatie (verder prevalentie genoemd) en de kans dat de respondent de gevoelige vraag moet beantwoorden (verder p-vraag)[2]. De keuze voor een methode hangt dan af van de verwachtingen van de onderzoeker ten aanzien van de prevalentie van een fenomeen in de populatie en de daarmee samenhangende keuze van een p-vraag, die enerzijds de veiligheid van de respondent voldoende garandeert, en anderzijds een voldoende efficiëntie belooft.

In grafiek 1 wordt de relatie tussen de efficiëntie van de belangrijkste randomized response methoden in kaart gebracht. Hierbij is de kans op een gevoelige vraag variabel gemaakt tussen 50 en 100 procent (.00- 1) en het aantal respondenten is op 100 gesteld.

De ‘relatieve efficiëntie’ is een maat voor de efficiëntie van een methode, waarbij de varianties behorend bij de standaard data-collectiemethoden worden vergeleken met die van de verschillende randomized response technieken, en de efficiëntie van directe bevraging op 100% wordt gesteld.

Grafiek 1: efficiëntie bij variabele kansen op de gevoelige vraag

In bovenstaande grafiek lezen we op de y-as de relatieve efficiëntie van de gebruikte RRT ten opzichte van directe vragen af. Op de x-as zien we dat de kans dat de sensitieve vraag moet worden beantwoord (p-vraag) variabel is tussen .5 en 1.

Bij een p-vraag van .7 heeft de forced response methode een relatieve efficiëntie van 45%, dat wil zeggen dat er 222 (100/.45) in plaats van 100 respondenten nodig zijn om een zelfde betrouwbaarheidsinterval te krijgen als wanneer er directe vragen waren gesteld. Voor de methode van Warner (en Kuk, die statistisch equivalent is) zijn bij een p-vraag van .7 zelfs 833 i.p.v. 100 respondenten nodig (100/.12).

Als de onderzoeker gebruik maakt van de unrelated question techniek met onbekende populatiegemiddelden dan zijn 10 keer zoveel respondenten nodig. Bij hogere p-vraag waarden blijft de unrelated question techniek met onbekende populatie parameters ver achter bij de andere methoden.

In grafiek 2 wordt de relatie tussen de prevalentie van een fenomeen en de efficiëntie van de verschillende onderzoeksmethode in beeld gebracht. Het voorkomen van het fenomeen varieert tussen .005, een half procent en 50 procent (.5). Er is gekozen voor een p-vraag van .7 (gemiddelde p-vraag over 42 verzamelde onderzoeken). Uit de grafiek kan worden afgelezen dat wanneer gedrag bij 30% van de respondenten wordt waargenomen (bijvoorbeeld: van alle respondent met een uitkering fraudeert 30% wel eens) de efficiëntie van de forced response methode ongeveer 45% is van rechtstreekse bevraging en dat er dus iets meer dan 2 maal zoveel respondenten nodig zijn om tot gelijke betrouwbaarheidsintervallen te komen. De methoden volgens Warner en Kuk blijven achter met ongeveer 10% efficiëntie, dus ongeveer 10 keer zoveel respondenten zijn nodig voor het verkrijgen van gelijke betrouwbaarheidsintervallen.

Grafiek 2: efficiëntie bij variabele prevalentie van het gevoelige gedrag

De conclusie is dan dat randomized response onderzoek nog niet erg efficiënt is in het gebruik van respondenten en daaruit voortvloeiend tijd en geld. Van de meest populaire methoden is de forced response methode de meest efficiënte. Bovendien blijkt de efficiëntie van de forced response methode niet samen te hangen met de prevalentie van het sensitieve gedrag, wat een voordeel is.

In de literatuur worden meerdere oplossingen gegeven voor het gebrek aan efficiëntie. Voor elke oplossing geldt dat zij slechts fragmentarisch beschreven is. Voor veel oplossingen geldt zelfs dat er slechts één internationaal gepubliceerd artikel kon worden gevonden, of dat er een kleine serie artikelen bestaat van /door dezelfde onderzoeksgroep. Vergroting van de efficiëntie van randomized response onderzoek kan worden gezocht in statistische verfijningen zowel als in methodologische veranderingen. De door ons voorgestelde onderzoeksrichtingen gaan uit van het potentieel gebruik van de forced response methode, de daaraan statistisch verwante unrelated question methode met bekende populatiegemiddelde, of de kaartmethode volgens Kuk. Dit omdat daarover al veel informatie bestaat, vooral van Nederlandse toepassingen. Het is natuurlijk mogelijk om alle in de literatuur vermelde alternatieve technieken voor randomized response onderzoek op een dergelijke wijze te exploreren, maar het is ons inziens efficiënter de bestaande kennis te gebruiken en verder te vergroten.

Onderzoek naar statistische oplossingen

In deze paragraaf worden enige voorstellen besproken die in de literatuur worden gedaan om de statistische efficiëntie te vergroten. Een selectie van deze voorstellen zou nader geëxploreerd kunnen worden om te bezien of ze in de context van het Justitie-onderzoek tot een verbetering van de onderzoeksmethode zouden leiden.

Het splitsen van de steekproef

Het opsplitsen van de steekproef in twee subgroepen met verschillende kansen voor de respondenten om de gevoelige vraag te moeten beantwoorden, zou bijdragen aan het vergroten van de relatieve efficiëntie van randomized response technieken (Clark & Desharnais 1998, Kendall & Stuart 1979).

Hoewel dit contra-intuïtief is, wordt in beide artikelen statistisch aangetoond dat de relatieve efficiëntie van gesplitste groepen groter is, dan wanneer alle respondenten in een groep worden geplaatst. Alle versies van randomized response onderzoek hebben grotere steekproeven nodig dan wanneer directe bevragingsmethoden worden gebruikt, maar het splitsen van de steekproef, en het gebruik van een *p*-vraag die verschillend en symmetrisch is rond 50% levert schatters op met een kleinere variantie, en is dus efficiënter.

waarbij π staat voor de ware populatie waarde (prevalentie) en p_1 en p_2 voor de verschillende kansen per subgroep om de vraag eerlijk te moeten beantwoorden (p-vraag).

Het opsplitsen van de respondenten in twee groepen lijkt een betrekkelijk simpele ingreep om de efficiëntie te vergroten. Voordat we dit in de praktijk kunnen toepassen zouden eerst verschillende modellen met variabele verhoudingen tussen p_1 en p_2 moeten worden doorgerekend. Daarbij is het van belang om ook goede afwegingen te maken tussen de statistisch meest ideale p_1- p_2 verhouding en de veiligheid die de respondent moet worden geboden. Nadat het meest efficiënte model is gekozen kan deze methode eenvoudig worden geïmplementeerd in vervolgonderzoek, ook in de door Justitie gehanteerde monitor. In dat geval moeten ook de gevolgen in de vorm van een mogelijke trendbreuk worden onderzocht. Een bijkomend voordeel van het opsplitsen van de respondenten in twee subgroepen is dat het dan in principe mogelijk wordt om uitspraken over 'cheating' te doen. Op het 'cheating'-probleem wordt elders ingegaan.

Het gebruik van een tweetraps procedure

Een andere manier om randomized responseprocedures efficiënter te maken is het gebruik van een twee-traps procedure (Mangat & Singh 1990). In de eerste fase wordt d.m.v. een randomizer bepaald of de respondent een gevoelige vraag moet beantwoorden (kans T: Ik heb mezelf wel eens schuldig gemaakt aan….) of dat hij of zij een tweede randomisatie moet doen (kans 1-T). In de tweede fase, alleen voor respondenten die de 1-T kans hadden wordt wederom een randomisatie gedaan met kans p op het gevoelige statement 'Ik heb mezelf wel eens schuldig gemaakt aan…….', en kans 1-p op het statement 'ik heb me nog nooit schuldig gemaakt aan….'. Mangat beschreef deze methode speciaal voor het gebruik van Warner's randomized response techniek. Moors (Moors 1971) beschrijft in zijn commentaar de mogelijkheden van de tweetraps procedure voor toepassingen in unrelated question en forced response designs. Wanneer T zodanig gekozen wordt dat geldt:

dan is de variantie van de niet biased schatter altijd kleiner dan wanneer er een eentraps methode wordt gebruikt.

Uitgezocht zou moeten worden of deze tweetraps procedure efficiënter is dan de hierboven beschreven splitsing van de steekproef. In een dergelijk onderzoek zal niet alleen de statistische efficiëntie een rol moeten spelen maar ook het extra voordeel van de splitsingsmethoden, de mogelijkheid om te corrigeren voor cheating. Daarnaast moet er aandacht zijn voor de mogelijke psychologische voor- en nadelen van de verschillende procedures, zoals ook beschreven in de paragraaf over effectiviteit.

Het vergelijken van de statistische efficiëntie van beide methoden en de extra voordelen die het detecteren van- en controleren voor cheaten kunnen bieden, kan worden gedaan met behulp van statistische analyses. De psychologische voordelen van beide procedures kunnen worden onderzocht met behulp van cognitieve laboratorium methoden.

Modificaties specifiek voor unrelated question technieken

Forced response en unrelated question technieken met bekende populatiegemiddelden voor de neutrale vraag zijn in statistisch opzicht aan elkaar gelijk; het aantal respondenten dat nodig is om een voldoende betrouwbaar resultaat te verkrijgen is hetzelfde. Het verschil tussen beide methoden is dat de respondent in het geval van forced response gedwongen wordt 'ja ' of 'nee' te antwoorden, terwijl de gevoelige vraag

wordt gelezen of in beeld is. Wanneer het onderwerp van de gevoelige vraag erg sociaal stigmatiserend is kan de respondent die geforceerd 'ja' moet antwoorden in de verleiding komen om te cheaten d.w.z. zich niet te houden aan de randomized response regels. De unrelated question is een psychologisch sterker design, dat minder snel aanleiding zal geven tot cheaten. De respondent moet antwoord geven op een neutrale vraag, waarvan het voorkomen in de populatie bekend is. Een voorbeeld van zo'n vraag is: is/ was uw moeder jarig in januari, februari of maart. Het is gemakkelijk hierop naar waarheid 'ja' te antwoorden, het is veel lastiger om geforceerd 'ja' te antwoorden na de vraag 'heeft u wel eens gefraudeerd'.

Het nadeel van de bovenbeschreven unrelated question methode is dat het aantal mogelijke vragen beperkt is. Er is een grens aan het aantal gedragingen met een exacte prevalentie van 25% dat een onderzoeker kan verzinnen. Zeker wanneer randomized response onderzoek vaker gebruikt wordt, zoals bij monitor onderzoek, of wanneer er een fors aantal vragen in randomized response format worden gesteld is dit een nadeel.

Een oplossing is de toepassing van de unrelated question techniek met onbekende populatie parameters. Hierbij wordt de steekproef in twee subgroepen gesplitst. Beide helften krijgen een gevoelige vraag voorgelegd of een neutrale vraag (bent u geabonneerd op de Volkskrant?). De kans op de gevoelige vraag in groep1 is p_1 en op de neutrale vraag $1-p_1$. De kans op de gevoelige vraag in groep2 is p_2 en op de neutrale vraag $1-p_2$, waarbij $p_1 \neq p_2$. Door middel van statistische manipulaties is het mogelijk om op deze wijze de populatie gemiddelden voor zowel het gevoelig gedrag als voor de neutrale vraag te schatten. Helaas is deze techniek heel inefficiënt (zie figuur 1). Een oplossing voor dat probleem zou een alternatieve toepassing van dit originele URQ model kunnen zijn (Moors 1971). Gewoonlijk worden p_1 en p_2 zodanig gekozen dat zij optellen tot 1 ($p_2=1-p_1$). Moors stelt voor om de kans op p_2, op 0 te stellen, dat wil zeggen respondenten het tweede subsample krijgen nooit de gevoelige vraag te beantwoorden maar slechts de neutrale vraag. Hierdoor hoeven de populatie parameters voor de neutrale vraag niet te worden geschat. De voordelen volgens Moors zijn:

- De voor het gevoelige topic heeft een kleinere variantie dan bij de originele methode, en dus is de methode efficiënter.
- De steekproef hoeft niet te worden opgedeeld in twee gelijke subgroepen, want de neutrale eigenschap kan met veel efficiëntere methoden worden ondervraagd. Er zijn dus minder respondenten nodig.
- Het neutrale statement kan worden onderzocht op een andere wijze dan het gevoelige topic (bijvoorbeeld zonder face-to-face interview).

Onderzocht moet worden of de methode volgens Moors efficiënter is dan de twee eerdergenoemde methoden en of validiteit van de resultaten bij het gebruik van URQ inderdaad groter is dan bij forced response designs.

De efficiëntie van Moors' methode kan weer worden vergeleken met de efficiëntie van de twee eerder genoemde methoden. Een dergelijke belangrijke vergelijking is nooit eerder gemaakt. In het beschreven onderzoek wordt de efficiëntie van de aanpassingen immers alleen getoetst aan de originele randomized response techniek.

Alle drie de bovengenoemde randomized response aanpassingen hebben potentiële voordelen voor de praktijk van het randomized response onderzoek. Sommige voordelen liggen op het vlak van grotere efficiëntie en mogelijkheden om te corrigeren voor cheaten, andere voordelen liggen meer op het vlak van het vergroten van de validiteit van de resultaten. Het vergroten van de validiteit kan zo veel voordeel hebben voor de praktijk van het beleidsonderzoek dat de inefficiëntie van de methode daardoor kan worden gecompenseerd. In het navolgende zullen we verschillende onderzoeksrichtingen bespreken die de validiteit van onderzoeksresultaten kunnen vergroten en daardoor tot zuiverder schatters leiden.

Onderzoek naar de relatie 'effectiviteit en sociale gevoeligheid van een onderwerp'.

Zoals 'efficiëntie' statistisch gerelateerd is aan de steekproefvariatie van de onderzoeksresultaten, zo is 'effectiviteit' gerelateerd aan de zuiverheid van de schatters. Een methode is meer effectief wanneer zij resulteert in betere schatters van ware populatie gemiddelden en dus minder vertekening (bias) oplevert.

De relatie tussen efficiëntie en effectiviteit is nog weinig onderzocht. Directe bevraging wordt wel 100% efficiënt genoemd, maar dan wordt alleen naar de steekproefvariantie gekeken, en niet naar mogelijke systematische vertekening. Randomized response is geïndiceerd wanneer de voordelen van een meer zuivere schatter opwegen tegen de extra kosten die voortvloeien uit de kleinere efficiëntie van de methode.

In tabel 1 kan de relatie worden afgelezen tussen de sociale gevoeligheid van het te onderzoeken fenomeen en de effectiviteit van RRT's en methoden voor directe vraagstellingen. Deze data zijn ontleend aan de meta-analyse (Lensvelt-Mulders et al. submitted). Als maat voor het effectverschil tussen randomized response technieken en directe vraagmethoden is de d-probit gebruikt, een verschilscore die het toelaat de resultaten van verschillende studies met elkaar te vergelijken (Cooper & Hedges 1994). De mate van sociale gevoeligheid van een onderwerp is bepaald op de manier van Himmelfarb (Himmelfarb & Lickteig 1982). Vier beoordelaars hebben 124 verschillende onderwerpen beoordeeld met behulp van een 5-puntschaal, waarbij de 1 staat voor 'dit onderwerp ligt niet sociaal gevoelig, respondenten zullen eerlijk antwoord geven' en de 5 voor 'dit onderwerp ligt zeer sociaal gevoelig, de onderzoeker kan nauwelijks eerlijke antwoorden op rechtstreekse vragen verwachten'. Er was grote overeenstemming tussen beoordelaars onderling over de sociale gevoeligheid van de verschillende onderwerpen (r=.76).

Zoals duidelijk af te lezen is in Tabel 1 neemt de effectiviteit van RRT in het algemeen toe met een toenemende sociale gevoeligheid van het onderwerp. Het gebruik van RRT is dus effectief wanneer het te onderzoeken fenomeen sterk sociaal gevoelig is.

Tabel 1: De relatie tussen sociale gevoeligheid en de effectiviteit van Randomized response technieken vergeleken met standaard surveymethoden

Sociale gevoeligheid	*d-probit*	
1	*0,0062*	
2	*0,1980*	
3	*0,3254*	
4	*0,3063*	
5	*0,4037*	*totaal:* r = *.196***

Eerder hebben we onderscheid gemaakt tussen vragen die intrinsiek en extrinsiek bedreigend zijn. Intrinsiek bedreigend zijn vragen die een respondent confronteren met zichzelf en zijn gevoelens van een positieve identiteit kunnen aantasten, een voorbeeld hiervan zijn vragen over afwijkend seksueel gedrag. Ook vragen over regelovertredend gedrag kunnen intrinsiek bedreigend zijn. Extrinsiek bedreigend zijn alle vragen die gaan over niet legale gedragingen waarop sancties staan. Regelovertredend gedrag zal doorgaans extrinsiek bedreigend zijn. Het is onbekend hoe beide vormen van 'bedreiging' hebben bijgedragen aan de resultaten uit tabel 1. Het is mogelijk dat RRT's vooral goed werken bij extern bedreigende vragen, of juist bij intern bedreigende vragen. Ook hiernaar zou onderzoek moeten worden verricht. De kern van dit onderzoek wordt verwoord in de onderzoeksvraag 'hoe draagt de bedreiging die uitgaat van bijvoorbeeld het onderwerp 'fraude met uitkeringen' bij aan het beslissingsproces van de respondent om: (1) zich zodanig te committeren aan de randomized response techniek dat men niet vals speelt, en (2) eerlijke antwoorden te geven op de gevoelige vragen.

Welke randomized response techniek is het meest effectief?

Het is onbekend welke vorm van randomized response onderzoek het meest effectief is. Er zijn wel onderzoeken bekend die twee randomized response technieken met elkaar vergelijken. In Nederland kennen we vooral het onderzoek van de onderzoeksgroep rond van der Heijden (Landsheer et al. 1999). In dit onderzoek worden Kuk's kaartmethode en de forced response methode met elkaar vergeleken, waarbij Kuk's methode meer valide resultaten geeft, maar niet significant meer. Er zijn echter weinig van dat soort onderzoeken, en de uitgevoerde meta-analyse geeft daardoor geen uitsluitsel. Om deze vraag te beantwoorden zouden nieuwe analyses moeten worden gedaan bij bestaande datasets, die (nog) niet in de meta-analyse zijn betrokken. Hierbij zouden door middel van een meta-analyse benadering de effecten van verschillende RRT's moeten worden vergeleken met elkaar en in het licht van de sociale gevoeligheid van het onderwerp. Of dit zinvol is hangt af van het aantal daarvoor beschikbare datasets, bijvoorbeeld van onderzoek in opdracht van Justitie.

Verdieping van dit vergelijkende onderzoek, parallel aan het eerder genoemde onderzoek naar mogelijkheden om de efficiëntie van randomized response technieken te vergroten, zal zeker bijdragen aan het vergroten van de kwaliteit van randomized response onderzoek voor beide facetten van kwaliteit, betrouwbaarheid en validiteit.

Objectieve en subjectieve veiligheid en manipulatie van p-vraag

De belangrijkste assumptie van randomized response onderzoek is dat de respondent een gevoel van veiligheid heeft, doordat zijn privacy gewaarborgd wordt. Daardoor zou een respondent meer geneigd zijn om eerlijk antwoord te geven op bedreigende vragen. Deze belangrijkste assumptie van de RRT is ons inziens nog onvoldoende onderzocht.

De p-vraag is grotendeels bepalend voor het gevoel van veiligheid dat een respondent heeft, maar in figuur 1 wordt geen onderscheid gemaakt tussen de objectieve kans om een gevoelige vraag te moeten beantwoorden en de subjectieve gevoelens van veiligheid bij de respondent. De gevolgen van het complexe psychologische proces dat leidt tot vertrouwen in de RRT (Landsheer, Van der Heijden en Van Gils, 1998) zijn niet af te lezen aan grafiek 1. Hier is p-vraag immers opgevat als een absolute en objectieve maat voor veiligheid: hoe lager p-vraag, hoe minder vaak de respondent een gevoelige vraag moet beantwoorden, dus des te beter is de bescherming van zijn privacy.

Geven respondenten vaker een eerlijk antwoord waneer het *gevoel* van veiligheid groter is? Uit de meta-analyse kon de invloed van p-vraag op de validiteit van randomized response resultaten niet worden opgemaakt. We hebben slechts één onderzoek gevonden waarbij bewust p-vraag is gemanipuleerd om verschillen in antwoordtendenties te onderzoeken (Soeken & Macready 1982). We hebben geen onderzoek gevonden waarbij de psychologische manipuleerbaarheid van p-vraag is onderzocht. Het is bekend dat mensen in het algemeen weinig correcte intuïties hebben ten aanzien van kansen en kansberekening, en dus hierin enigszins gemanipuleerd kunnen worden. Daardoor kan de waargenomen subjectieve bescherming groter zijn dan de statistische objectieve bescherming. Een voorbeeld hiervan vinden we bij de toepassing van de forced response methode, waarbij de respondent twee dobbelstenen moet gooien en de resultaten moet optellen. Bij een worp van 2, 3, 4, 11 en 12 moet geforceerd antwoord worden gegeven, bij een worp van 5, 6, 7, 8, 9, en 10 moet de gevoelige vraag beantwoord worden. In de ogen van veel respondenten betekent dit dat de kans op de gevoelige vraag bijna even groot is als de kans op geforceerd antwoorden, terwijl de werkelijke kans op de gevoelige vraag 75% is. Op deze manier wordt het subjectieve gevoel van veiligheid gemanipuleerd. Dit kan echter niet zonder grenzen worden doorgevoerd, na een aantal vragen begint het zelfs de meest trage respondent te dagen. Wanneer op deze manier 100 respondenten 6 vragen moeten beantwoorden zullen immers ongeveer 18 respondenten 6 maal achtereen de gevoelige vraag moeten beantwoorden. Dat neemt niet weg dat onderzoek naar de relatie tussen objectieve en subjectieve veiligheid kan bijdragen aan het vergroten van efficiëntie en effectiviteit van randomized response onderzoek. De efficiëntie van RRT's kan worden vergroot doordat met grotere p-vraag kan worden gewerkt zonder dat de respondent zich minder veilig voelt. De effectiviteit van RRT's kan worden vergroot wanneer bij gelijkblijvende p-vraag, de gevoelens van veiligheid groter worden en dus de neiging om vals te spelen of te liegen af zal nemen.

Onderzoeksvoorstellen m.b.t. efficiëntie en effectiviteit onderzoek

Er lijkt in een aantal opzichten een balans te zijn tussen efficiëntie en effectiviteit. Om zicht te krijgen op de problematiek kunnen de volgende onderzoekingen worden uitgevoerd.

Methodologisch onderzoek

(1) Meta-analyse van beschikbaar materiaal. Op basis van bestaand materiaal in de meta-analyse van de onderzoeksliteratuur coderen voor intrinsieke en extrinsieke bedreiging, en onderzoeken in hoeverre de effectiviteit van de RRT hiermee samenhangt. Vervolgens bijeenbrengen van de reeds beschikbare RRT datasets voor een niet-individueel gevalideerde meta-analyse gericht op de vraag in hoeverre verschillen in de opzet van RR-onderzoek tot verschillende resultaten leiden.

(2) Inventariseren bij eerder uitgevoerd monitoronderzoek wat realistisch te verwachten prevalenties zijn. Op basis van de boven geschetste relaties tussen de gekozen RRT, prevalentie, en p-vraag bepalen op welke punten het monitordesign verbeterd kan worden door eenvoudig in te voeren wijzigingen, die de continuïteit en onderlinge vergelijkbaarheid niet aantasten. Hiervoor kan een voorkeurstabel worden opgesteld.

Onderzoeksvoorstellen statistisch onderzoek

1. Uitwerken van de als uitvoerbaar beschouwde modificaties, inclusief modificaties van grotere omvang dan die bedoeld bij punt (2) onder 'methodologisch onderzoek'.
2. Onderzoeken welke modificaties het mogelijk maken 'cheating' op te sporen. Dit heeft alleen zin wanneer het leidt tot een zuiverder schatting van de grootheid waar het werkelijk om gaat. De huidige methoden die rekening houden met cheaters kunnen niet aangevuld worden met

achtergrondvariabelen. Dit is een groot nadeel. In het kader van de statistische- en softwareontwikkeling zou onderzocht moeten worden of hier iets aan te doen is.

Onderzoeksvoorstel vergelijkend onderzoek
Als volgende stap worden op basis van de hierboven vergaarde informatie veelbelovende varianten van RRT gespecificeerd. Deze worden met cognitieve lab methoden geoptimaliseerd, en vervolgens wordt een vergelijkend empirisch onderzoek uitgevoerd. Bij voorkeur moet hierbij de mogelijkheid van computergestuurd RR onderzoek (CARR) worden meegenomen (zie volgende sectie).

Optimalisatie van randomized response technieken bij computergestuurd interviewen

Het is wenselijk computergestuurde varianten van RRT te onderzoeken. Een eerste verkenning m.b.v cognitieve laboratoria methoden door de capaciteitsgroep Methodenleer & Statistiek in samenwerking met NIPO ziet er gunstig uit: respondenten lijken de computer te accepteren en zelfs te waarderen, en het gevoel van privacy wordt door de computergestuurde techniek niet aangetast. Wel zijn enige aanpassingen in de standaard instructie nodig gebleken, en moest de interviewsoftware worden aangepast. Bij gebruik van een standaardpanel zoals CAPI@Home treden potentieel twee soorten vertekeningen op: meetfouten als gevolg van de computermethode, en dekkingsfouten door de selectieve samenstelling van het panel. Wij bespreken hieronder de mogelijke meetfouten, en gaan daarna in op mogelijke dekkingsfouten.

Computer gestuurd onderzoek biedt veel algemene voordelen. Het is efficiënt in tijd en geld, de data-invoer verloopt automatisch, de non-respons is lager, en er is waarschijnlijk een grotere perceptie van privacy. Dit laatste is zeker het geval wanneer de respondent de vragenlijst thuis in de vertrouwde omgeving en in de eigen tijd kan invullen. Een daarbij komend voordeel kan zijn dat niet alle antwoorden uit de herinnering hoeven te worden gehaald, maar dat informatie die men zich niet direct herinnert kan worden opgezocht. Bij gebruik van de computer kunnen de data bovendien via internet worden verzameld. Daarbij moet overigens worden aangetekend dat men dan meestal op panels van vaste respondenten is aangewezen. Deze brengen het risico van vertekening van de resultaten door selectie van respondenten met zich mee.

Randomized response onderzoek met behulp van een computer noemen we verder CARR, computer assisted randomized response research. Voordat we randomized response technieken kunnen implementeren in CARR zijn er drie vragen die beantwoord moeten worden:

1. Wat is de meest effectieve manier om randomized response technieken in een computergestuurde vragenlijst te implementeren
2. Wat zijn de effecten van het gebruik van de computer op het randomized response vraag en antwoord proces
3. Leidt CARR tot betere datakwaliteit CASI wanneer onderzoek wordt gedaan naar zeer gevoelige onderwerpen en /of bij speciale populaties.

Ad 1: Het ontwikkelen van *CARR*
Wat is de meest effectieve manier om randomized response technieken in een computergestuurde vragenlijst te implementeren?

De belangrijkste vraag hierbij is de te gebruiken randomizer. Moet de randomizer worden toegestuurd, of is het beter gebruik te maken van attributen die de respondent zelf in huis heeft? Of is het mogelijk om de randomizer binnen het computerprogramma te brengen? Voor methodes met externe randomizers geldt dat de respondent extra handelingen moet verrichten. Het is bekend uit methodologisch survey-onderzoek dat alle extra handelingen aanleiding geven tot fouten. Respondenten kunnen bijvoorbeeld besluiten om in plaats van het telefoonboek te raadplegen zelf nummers tussen 0 en 9 te verzinnen. Omdat onderzoek aantoont dat mensen slecht zijn in randomisatie zal dat leiden tot een vertekening van de resultaten van het onderzoek (Appelmelk et al. 2001). Het voordeel van het gebruik van externe randomizers is dat de respondent een groter gevoel van veiligheid kan hebben.

Het nadeel van het implementeren van de randomizer in het computerprogramma is dat de respondent zich sterk onveilig kan gaan voelen en daardoor niet meer zal meewerken aan de randomized response spelregels en /of zal gaan liegen bij het beantwoorden van de gevoelige vraag.

Om inzicht te krijgen in de beste manier om het randomized response format te implementeren in CASI onderzoek stellen we voor een experiment op te zetten waarbij verschillende vormen van randomizers vergeleken worden. In de eerste fase kan met behulp van kwalitatief onderzoek het effect worden onderzocht van het gebruik van interne en externe randomizers op de respondent's vertrouwen in-

en begrip van randomized response methoden. In de tweede fase worden met behulp van een validatiestudie de resultaten van verschillende vormen van CARR vergeleken.

Ad 2: Demand characteristics bij onderzoek met behulp van de computer
Wat zijn de effecten van het gebruik van de computer op het randomized response vraag en antwoord proces? Deze vraag kan onderzocht worden door gebruik te maken van de eerder genoemde cognitieve laboratorium procedures. Op basis van de literatuur zou verwacht moeten worden dat CARR leidt tot een groter gevoel van privacy, maar of dit ook geldt wanneer de randomizer ook in de computer wordt gegenereerd is nog onduidelijk.

Ad 3: Leidt CARR tot betere datakwaliteit dan CASI wanneer onderzoek wordt gedaan naar zeer gevoelige onderwerpen en /of bij speciale populaties?
Wanneer tenslotte de beste vorm van CARR is vastgesteld, dan blijft de vraag of het gebruik van een combinatie van methoden, computer gestuurd onderzoek gecombineerd met het randomized response format, leidt tot betere resultaten als wanneer enkel CASI (of eventueel audio-CASI) onderzoek was gedaan. Uit de literatuur is bekend dat beide vormen van survey onderzoek voordelen hebben ten opzichte van standaard methoden als interviews, papieren vragenlijsten, en telefonische interviews. Leidt de combinatie van de voordelen van beide methodes tot nog betere resultaten? Dit is waarschijnlijk het geval. Is deze verbetering ook significant wanneer deze wordt afgewogen tegen de grotere inefficiëntie van randomized response onderzoek? Dat is nog onhelder. Om hier inzicht in te krijgen is het noodzakelijk om een experiment op te zetten waarbij de resultaten van (audio)CASI worden vergeleken met de resultaten van CARR. Belangrijke variabelen die aanleiding kunnen geven tot verschillen tussen beide methoden zijn het sociale gevoeligheidsaspect van het te onderzoeken gedrag, kenmerken van de onderzoekspopulatie, en de verwachte prevalentie van het gedrag in de populatie.

Op grond van de nu beschikbare informatie zijn meetfouten bij CARR vermoedelijk niet het belangrijkste probleem. De mogelijk selectieve samenstelling van vaste computerpanels zoals CAPI@Home is waarschijnlijk een veel groter probleem. Het gebruik van vaste computerpanels heeft het voordeel dat het relatief goedkoop is. Uitgezocht zou moeten worden of de verschillen tussen het CAPI@Home panel en direct benaderde respondenten groot genoeg zijn om tot verschillende resultaten te leiden. Standaard wordt in survey-onderzoek selectiviteit aangepakt door *weging* van de gegevens naar bekende populatiecijfers. De vraag m.b.t. CAPI@Home is of voldoende achtergrondinformatie beschikbaar is om te wegen.

Onderzoeksvoorstellen
Onderzoeken in hoeverre en op welke wijze CARR de kwaliteit van de gegevens verhoogt dan wel de kosten van de gegevensverzameling verlaagd..

2.4 Mogelijkheden om kwantitatieve gegevens te verzamelen

Hoewel het in Nederland nog niet is toegepast, is het mogelijk om met behulp van randomized response technieken kwantitatieve gegevens, zoals frequentieschattingen, te verzamelen. In de literatuur zijn vier methoden beschreven. In het onderstaande zullen we eerst deze methoden beschrijven, daarna geven we een kort overzicht van literatuur waarin (vergelijkend) onderzoek met kwantitatieve methoden wordt beschreven en tenslotte verkennen we de hiaten die in dit onderzoek zitten en die moeten worden opgelost voordat kwantitatieve vormen van randomized response in beleidsonderzoek kunnen worden toegepast.

Greenberg's unrelated question model[3]
Het Greenberg model voor het meten van kwantitatieve gegevens is een vorm van zijn unrelated question model met onbekende populatie parameters. Hierbij worden twee vragen voorgelegd, een onschuldige vraag (kans 1-p) en een gevoelige vraag (kans p).
Het is van belang dat de antwoorden op de onschuldige vraag in dezelfde orde van grote liggen als die op de gevoelige vraag, dus vergelijkbare gemiddelden en varianties hebben. Dit om de respondent niet achterdochtig te maken. Verder moet bij dit design de steekproef in twee worden gedeeld, waarbij p (de kans op de gevoelige vraag) zo gekozen wordt dat

$p_1 + p_2 = 1$ en p1 zo ver mogelijk van .5 wordt gekozen als haalbaar is binnen de grenzen van vertrouwen en privacybescherming. Daarnaast kan men de N voor elke groep zo optimaal mogelijk kiezen. Greenberg toont aan dat de methode het meest efficiënt is wanneer

.

Greenberg geeft in zijn artikel twee voorbeelden. Een voorbeeld met vragen over het aantal abortussen dat een vrouw heeft ondergaan. De gevoelige vraag 'hoeveel abortussen hebt u in uw leven ondergaan' wordt gekoppeld aan de onschuldige vraag 'als een vrouw full time werkt hoeveel kinderen zou zij dan volgens u maximaal moeten hebben'. In het tweede voorbeeld wordt de gevoelige vraag 'hoeveel dollar heeft het hoofd van het huishouden dit jaar verdiend' gekoppeld aan de onschuldige vraag 'hoeveel dollar denkt u dat gemiddeld per maand wordt verdiend.'

Model van Lui and Chow[4]

Het bovenbeschreven model van Greenberg kan op twee manieren efficiënter worden gemaakt. Ten eerst kan worden gewerkt met een onschuldige vraag waarbij de frequentie distributie in de populatie bekend is, bijvoorbeeld 'hoeveel personen telt uw huishouden'. Deze gegevens zijn bekend, zowel de gemiddelde als de variantie is op te vragen, in Nederland bijvoorbeeld bij het CBS. Lui en Chow (1976) hebben dit model verder aangepast door het te modeleren naar het forced response model. Zij maken hiervoor gebruik van een container met twee kleuren ballen, rood en wit. Als de respondent een rode bal trekt dan moet de gevoelige vraag worden beantwoord, als de witte bal wordt getrokken moet het cijfer op de witte bal worden opgelezen. Door de cijfers op de witte bal goed te kiezen kan de methode voor alle soorten vragen worden aangepast (vergelijk abortus en inkomen), en omdat de cijfers tevoren gekozen zijn liggen gemiddelden en varianties vast. Dit maakt dat er slechts een steekproef nodig is, dus dat de methode efficiënter is dan Greenberg's model.

Additieve en multiplicatie modellen[5]

In het artikel van Pollock en Bek worden twee andere methoden besproken om door middel van randomized response kwantitatieve gegevens te krijgen. De zogenaamde additieve en multiplicatie methoden. Bij de additieve methode wordt de respondent gevraagd om bij het antwoord op de gevoelige vraag (X) een random nummer (Y) uit een bekende verdeling van nummers op te tellen. Het getal dat wordt ingevuld als uitkomst is dus Z=X+Y. Bij het multiplicatie model wordt de respondent gevraagd om het antwoord op de gevoelige vraag (X) te vermenigvuldigen met Y, ook weer met bekende distributie functie, waardoor Z=X×Y.

Het is in beide gevallen mogelijk om met behulp van Z zowel een moment schatter ( ) te berekenen als een schatting te maken van de distributie functie (*f(x)*) van het gevoelig gedrag (parametrische schatters). De mogelijkheid de distributie functie te schatten heeft als voordeel dat maximum likelihood estimation simpel is uit te voeren, wat bij de andere randomized response methoden minder het geval is.

Kort literatuuroverzicht

Naast bovengenoemde onderzoekers hebben vooral Fox en Tracy (Fox & Tracy 1980a, b, Tracy & Fox 1981) zich beziggehouden met kwantitatieve toepassingen van de randomized response methode. Zij hebben zogenaamde harde validatie studies verricht naar het aantal malen dat iemand is opgepakt voor dronken rijden (DUI) en naar het aantal malen dat een persoon met de politie in aanraking is geweest (number of arrests). In beide gevallen waren de resultaten van het randomized response onderzoek meer valide dan wanneer directe vraagmethoden werden gebruikt.

Recenter onderzoek naar kwantitatieve toepassingen van randomized response technieken is van Singh en medewerkers (Singh et al. 1994, Singh et al. 1995). In Singh et al. (1990) wordt een modificatie van het Greenberg model besproken die analoog is aan de modificatie van het algemene randomized response model van Mangat en Singh (1990). Om de randomized response methode efficiënter te maken stellen deze ook nu een twee-traps procedure voor (Mangat & Singh 1990). In de eerste fase wordt dmv een randomizer bepaald of de respondent een gevoelige vraag moet beantwoorden (kans T: hoe vaak heeft u….) of dat hij of zij een tweede randomisatie moet doen (kans 1-T). In de tweede fase, alleen voor respondenten die de 1-T kans hadden, wordt wederom een randomisatie gedaan met kans p op de gevoelige vraag 'hoe vaak heeft u ….', en kans 1-p op de neutrale vraag 'hoe vaak ….'. Singh en

medewerkers tonen aan dat hun procedure efficiënter is dan het Greenberg design. Het andere onderzoek beschrijft de toepassing van een kwantitatieve randomized response methoden in een per post toegestuurde vragenlijst.

Voorstellen voor onderzoek met betrekking op Nederlands beleidsonderzoek

Kwantitatieve methoden lijken op het eerste gezicht bruikbaar om toe te passen in onderzoek naar regelhandhaving. Wanneer we ons oriënteren op de keuze voor een bepaalde methode dan is het verstandig om een traject te volgen parallel aan dat waarbij de efficiëntie en effectiviteit van de randomized response in het algemeen wordt onderzocht. Als gekozen wordt voor een vorm van forced response of unrelated question met gekende populatie parameters dan kunnen waarschijnlijk alle bij het hoofdstuk 'efficiëntie en effectiviteit' beschreven aanpassingen ook voor kwantitatieve toepassingen van de randomized response methode worden onderzocht, zoals Singh en medewerkers (1994) dat al deden voor hun tweetrapsprocedure.

Pollock en Lui vergelijken in hun artikel de efficiëntie van de unrelated question methode met die van de additieve en multiplicatie methoden van randomized response onderzoek. Uit hun conclusies kan worden opgemaakt dat de additieve en multiplicatie methoden potentieel waardevol zijn. Omdat de frequentie en spreiding van Y bekend is kan eenvoudig een schatter voor het populatiegemiddelde worden bepaald. Voor Greenberg's methode en voor de additieve methode is het goed om de frequentie en spreiding van Y dicht bij de verwachte frequentie en spreiding van X (gevoelig onderwerp) te kiezen. Voor de multiplicatie methode geldt dat de spreidingen gelijk moeten zijn, maar dat het gemiddelde van Y zou

overeen moeten komen met . Pollock en Lui stellen voor om toekomstig onderzoek te richten op de vergelijking tussen deze methoden en die voor directe vraagmethoden. Dit is een onderzoekslijn die ook wij voorstaan. Dit onderzoek kan in vier stappen worden uitgevoerd.

1. Statistische bureau studies (deskresearch) waarbij de efficiëntie van verschillende vormen van kwantitatieve randomized response methoden worden vergeleken. Een exercitie die parallel loopt met de al eerder voorgestelde onderzoekslijn om de efficiëntie en effectiviteit van dichotoom randomized response onderzoek te vergelijken en te verbeteren.
2. Statistisch bureau onderzoek (idem) naar methoden om bij kwantitatieve gegevens de relatie tussen randomized response variabelen en verklarende variabelen te leggen, bijvoorbeeld met behulp van (logistische) regressie en naar relaties tussen randomized response variabelen onderling (zie ook Tracy en Fox, 1981).
3. Experimenteel onderzoek naar de manier waarop respondenten omgaan met additieve en multiplicatie methoden. Het is bekend dat als een respondent meer handelingen moet verrichten de kans dat hij een fout maakt snel toeneemt. Dit vormt een bedreiging voor de validiteit van de resultaten van het onderzoek. Onderzocht dient te worden hoe eventueel rekenwerk voor respondenten zo makkelijk mogelijk kan worden gemaakt, zodat de kans op fouten en dus vertekening van de resultaten zo klein mogelijk wordt.
4. Onderzoek is nodig naar de mogelijkheden van het toepassen van kwantitatieve methoden binnen computer gestuurd onderzoek. Omdat het onderzoek naar mogelijkheden om de gewone randomized response methode binnen een CASI omgeving te gebruiken (CARR) nog maar net begonnen is, kan de ontwikkeling van kwantitatief computergestuurd randomized response onderzoek daar parallel aan lopen. Dezelfde vragen moeten tenslotte worden onderzocht. (het ontwikkelen van de meest effectieve methode, interne of externe randomizer, is kwantitatieve CARR zoveel effectiever dan (A)-CASI dat dit opweegt tegen het verlies van efficiëntie?). Deze vragen kunnen het beste worden beantwoord door het doen van experimenteel onderzoek, waarbij de antwoorden van onderzoek 1 en 2 worden gebruikt om tot een keuze van een beperkt aantal randomized response methoden die vergeleken worden met elkaar en directe bevraging (CASI).

3. Analyse van Randomized response data

Peter van der Heijden

Ardo van den Hout

Methodenleer en Statistiek, Universiteit Utrecht.

3.1 Inleiding

In onderzoeken waarin randomized response data worden verzameld, is de analyse vaak een ondergeschoven kindje. Hoewel reeds lang bekend is dat analyse van randomized response data in relatie tot andere variabelen wel degelijk mogelijk is (zie bijvoorbeeld Fox and Tracy, 1986, maar ook eerdere literatuur is te geven), wordt in het Monitorrapport van het Ministerie van Justitie ten onrechte aangegeven dat het een nadeel van RR data is dat het niet mogelijk is een relatie met andere variabelen te leggen. In onderzoek in opdracht van het Ministerie van SoZaWe wordt wel degelijk gebruik gemaakt van statistische analyses waarin randomized response variabelen zijn betrokken.

In dit hoofdstuk worden drie terreinen onderscheiden waarop nader onderzoek met betrekking tot de analyse van data verkregen met randomized response gewenst is. Als eerste behandelen we onderzoek dat moet leiden tot een grotere verspreiding van kennis over de analyse van randomized response data, m.n. door het ontwikkelen en beschikbaar stellen van gebruikersvriendelijke software voor de markt. Als tweede komt aan bod onderzoek ter verhoging van de validiteit, betrouwbaarheid en beleidsrelevantie van RR, waarbij een telvariabele wordt geperfectioneerd. De onderwerpen waarop nader onderzoek gewenst is zijn:

1. In hoeverre kunnen randomized response data met standaard software geanalyseerd worden.
2. Logistische regressie in een RR context, en
3. Telvariabelen in een RR context.

Iedere paragraaf is steeds onderverdeeld in een beschrijving van wat er bekend is (‘stand van zaken’), wat ontwikkeld zou moeten worden (‘desiderata’) en concrete onderzoeksvoorstellen.

3.2. Analyse randomized response data met standaard software

Onder standaard software verstaan we software voor statistische analyse die te koop is en die in principe bruikbaar is zonder dat men dient te programmeren. We komen dan uit bij pakketten als SPSS, SAS, STATA en Splus, maar men kan ook denken aan een special purpose programma als LEM (Vermunt, 1993)

Voor de helderheid onderscheiden we hieronder het antwoord op de RR-vraag – hiernaar refereren we met RR -, en het gevoelige gedrag waarnaar de RR vraag refereert, en die noemen we GG.

Stand van zaken

In Fox and Tracy (1986) worden al een aantal analyses genoemd die uit te voeren zijn met standaard software:

1. relatie tussen categorische verklarende variabele en één RR variabele, bijvoorbeeld geslacht met regelovertreding. Hier kan een gewone 2-weg kruistabelanalyse worden uitgevoerd op geslacht met RR, en de toets voor onafhankelijkheid geeft een juiste uitslag m.b.t. de significantie van de relatie tussen geslacht en GG in de populatie. Probleem is dat de sterkte van de relatie tussen geslacht en GG niet uit het programma komt rollen. Hier is dus weer een extra stap nodig: bijvoorbeeld, bereken voor elke categorie van geslacht afzonderlijk wat de kans op GG is.
2. relatie tussen twee RR variabelen, bijvoorbeeld twee vormen van regelovertreding die we meten met RR1 en RR2. Hier kan ook een gewone 2-weg kruistabelanalyse worden uitgevoerd op RR1 en RR2, en de toets voor onafhankelijkheid geeft een juist beeld van de significantie van de relatie tussen GG1 en GG2 in de populatie. Probleem is ook hier dat de sterkte van de relatie tussen GG1 en GG2 niet uit het programma komt rollen. Hier is dus weer een extra stap nodig die niet geheel eenvoudig uiteen te zetten is: men wil hier de schattingen van de kansen op gezamenlijk voorkomen van GG1 en GG2.

Verder is er een bètaversie van een programma voor het volgende probleem:

3. relatie tussen enerzijds een of meer kwantitatieve en/ of categorische variabelen en anderzijds één RR variabele (punt 1 is hier een speciaal geval van). Voor deze situatie is een aangepaste versie van logistische regressie geprogrammeerd in SPSS die het eigendom is van IOPS-Utrecht. Dit wordt verder uitgewerkt in sectie 2.2.

Desiderata

Opdrachtonderzoek vindt vaak plaats onder tijdsdruk. Daarom is het voor het Ministerie van Justitie gewenst dat er gebruikersvriendelijke software beschikbaar komt voor een pakket aan standaardanalyses van RR data waarmee men snel en foutloos kan werken. Het Ministerie is zo in staat offertes te doen uitgaan waarin partijen die willen offreren van deze software gebruik kunnen maken.

Het is ongewenst dat een deel van de berekeningen nog met de hand worden uitgevoerd. In de eerste plaats vergroot dit de kans op fouten. In de tweede plaats kost dit meer tijd, waardoor het onderzoek duurder wordt.

Voorstellen voor onderzoek

Hoewel dit niet algemeen bekend is, is er een nauwe relatie tussen latente klasse analyse en de analyse van RR variabelen. Zonder hier in details te treden komt het er op neer dat de conditionele kansen binnen het latente klasse model gefixeerd dienen te worden op waarden die het RR design bepalen. Bekeken dient te worden of de nadelen uit bovengenoemde punten 1 als 2 zo opgelost kunnen worden. Indien dit mogelijk is, zou een handboek gemaakt dienen te worden voor dit onderwerp. Hoewel er geen software voor latente klasse analyse binnen bestaande statistische pakketten bestaat, is er wel erg professionele software op de markt, namelijk het programma Latent Gold, dat oorspronkelijk bedoeld was om aangehaakt te worden aan SPSS. Daarnaast kan onderzocht worden wat de mogelijkheden zijn van het programma LEM. Het lijkt gewenst bij deze software aan te haken.

3.3 Logistische regressie

Stand van zaken

Er wordt in het onderzoek voor SoZaWe (Van Gils, Van der Heijden en Rosebeek, 2001; Bouts, Van Gils en Van der Heijden, 1998; Van Gils, Van der Heijden en Landsheer, 1996) standaard gebruikgemaakt van logistische regressie. Logistische regressie wordt gebruikt om een verklarende variabele (bijvoorbeeld een T11 factor) te relateren aan een randomized response variabele.

Het analyseprogramma is geprogrammeerd in SPSS Matrix. Dit heeft het voordeel dat gebruikers SPSS niet hoeven te verlaten om deze berekeningen uit te voeren. Het programma levert regressiecoëfficiënten en de significantie van elk van deze coëfficiënten. Daarnaast wordt de likelihood gegeven, waardoor verschillen tussen modellen kunnen worden getoetst.

In de analyses die worden gerapporteerd in rapporten van SoZaWe worden steeds univariate logistische regressies gebruikt, waarbij alleen wordt gekeken naar het *teken* van een significante regressiecoëfficiënt. Bijvoorbeeld, er wordt slechts gerapporteerd ‘inschatting van zwaardere sanctie leidt tot minder regelovertreding’. Effectmaten zijn te berekenen door de specifieke waarden in de regressievergelijking in te vullen, maar met de hand. Zo’n effect maat zou zijn dat, bijvoorbeeld, bij een inschatting van de kans op 80% is de kans op regelovertreding 10%, maar bij een inschatting van 50% is die kans 20 %.

We laten hier enkele voorbeelden zien. Tabel 1 (uit Van Gils, Van der Heijden en Rosebeek, 2001) is een voorbeeld waarin alleen gebruik wordt gemaakt van het teken van regressiecoëfficienten. Voorbeelden van het gebruik van het analyseprogramma zijn ook te vinden in Van der Heijden et al. (2000) en Elffers et al. (ter publicatie aangeboden).

1 | Samenhangen met overtredingen van de regels met betrekking tot het rapporteren van inkomsten naast de uitkering (WAO)[6]

Verklarende variabele	Klusjes voor betaling naast uitkering	Uitzendwerk of baan naast uitkering	Zwart werk naast uitkering
Mening inkomen			

Verklarende variabele	Klusjes voor betaling naast uitkering	Uitzendwerk of baan naast uitkering	Zwart werk naast uitkering
Goed op de hoogte van de regels			
Regels zijn duidelijk			
Kennis van de regels			
Moeite te voldoen aan regels			Meer moeite: vaker zwart werk
UVI laat mensen aan lot over	Meer mee eens: vaker klusjes		
Uvi geen oog voor persoonlijke omstandigheden			
Voordeel van fraude	Voordeliger: vaker klusjes	Voordeliger: vaker werk/ baan	Voordeliger: vaker zwart werk
Controle op baan of uitzendwerk is waterdicht	Controle beter: vaker klusjes		
Controle op zwart werk is waterdicht			
Kans op ontdekking van bijklussen			
Informele meldingskans			
Afkeuring fraude omgeving	Meer afkeuring: minder klusjes		
Wetgetrouwheid	Wetsgetrouwer: minder klusjes		Meer wetsgetrouw: minder zwart werk
Straffen zijn zwaarder geworden			
Controlekans			
Kans op regelmatige controle			
Pakkans bij overtreding			
Strafkans bij overtreding			
Door straf in moeilijkheden			
Redelijkheid van verplichtingen	Onredelijker: vaker klusjes	Onredelijker: vaker werk/ baan	

Hieronder staan twee voorbeeld tabellen van een vollediger gebruik van de mogelijkheden van het analyseprogramma. Deze voorbeelden zijn gebaseerd op Elffers, Van der Heijden en Hezemans (ter publicatie aangeboden) en hebben betrekking op de huursubsidiewet (IRS). De schalen “benefits, control, detection, sanctions’ zijn uit T11-variabelen geconstrueerd. Deze schalen worden gebruikt om de RR-variabele uit het IRS- onderzoek te voorspellen. In tabel 2 staan in het linkerdeel de resultaten van een multivariate logistische regressie met de vier schalen als verklarende variabelen, en in het rechterdeel vier uni-variate logistische regressies waarin steeds elk van de schalen afzonderlijk de verklarende variabele is. In tabel 3 worden de ‘effect sizes’ voor dezelfde (univariate) variabelen als in tabel 2 weergegeven.

Table 2: Adjusted multivariate logistic regression, IRS. Dependent variable: randomised response variable. Explanatory variables: condensed T^{11} –scales.

SCALES	IRS					
	multi-variate*			uni-variate		
	b	se(b)	signif	b	se(b)	signif
Benefits	.94	.36	x	.68	.35	x
Control	-.14	.17		-.17	.09	x

Detection	.00	.15		-.13	.06	x
Sanctions	.17	.11		-.15	.07	x

* The estimate of the constant term in the multivariate case is: 5.98 Given are parameter estimates for unstandardised regression weight b, with standard errors of b. When |estimate/standard error| > 1.96, the estimate departs significantly from zero (2-sided test). These are indicated with an 'x' in the column 'signif'. For comparison, univariate b-estimates from table 4 are given as well

Table 3: Effect size of explanatory T^{11} dimensions.
Estimated probabilities of non-compliance (EPNC) when explanatory variable is plus or minus one (sample) standard deviation from its (sample) mean. Effect size (ES) is the difference between the level estimates at + or – one standard deviation.

	IRS		
	EPNC total group: 8%		
SCALES	EPNC at – 1 st.dev	EPNC at +1 st.dev.	Effect size ES
Benefits	3%	12%	9%
Control	4%	17%	13%
Detection	5%	16%	11%
Sanctions	4%	19%	15%

We include only explanatory variables having significant regression. Grey cells indicate that the factor concerned did not reach significance for that law.

De b-waarden in tabel 2 zijn minder gemakkelijk te interpreteren dan de effectgrootte in tabel 3. Uit tabel 3 blijkt dat, mensen die 1 standaarddeviatie boven gemiddelde scoren op de verklarende variabele 'benefits', in 12 % van gevallen de regel overtreden; mensen die 1 standaarddeviatie onder gemiddelde scoren op benefits, overtreden in 3% van gevallen de regel. De effectgrootte is dan 12 – 3 = 9%. Overigens is in tabel 3 de effectsize berekend voor twee standaarddeviaties (nl. het verschil tussen het gemiddelde plus en min een standaarddeviatie). Mogelijk is het beter om in overeenstemming met hetgeen gebruikelijk is in de literatuur effectsizes te bereken voor het 25ste, 50ste en 75ste percentiel.

Tabellen met b-waarden kunnen dus gebruikt worden om tabellen met effectgroottes te berekenen. Deze tabellen met effectgroottes zijn zeer goed te interpreteren, en bieden meer informatie dan de voorbeeldtabel uit het SoZaWe-onderzoek, waarin alleen richtingen van verbanden worden aangegeven. Het probleem is echter dat het huidige analyseprogramma dit niet voor een gebruiker doet: een gebruiker dient de benodigde gegevens zelf in te vullen in formules. Dit is niet gebruikersvriendelijk en kan gemakkelijk leiden tot fouten.

Desiderata

Het bestaande analyseprogramma heeft alle kenmerken van een bètaversie van een serieus programma. Het dient professioneler te worden gemaakt en er zijn zinvolle uitbreidingen mogelijk.

Daarnaast is het wenselijk dat het SPSS programma, dat nu ontwikkeld is door en eigendom is van IOPS-UU, in het publieke domein komt. Een van de zaken die hiervoor dienen te worden geregeld is een adequate handleiding van het programma. Dit programma zou zonder veel moeite te gebruiken moeten zijn door mensen die om kunnen gaan met SPSS.

Op deze wijze kan voortaan in elk onderzoek waarin gebruik wordt gemaakt van RR de relatie gelegd worden tussen verklarende variabelen en de RR-variabele(n). Tot nu toe is dit niet het geval: alleen in onderzoek waar IOPS-UU bij betrokken is, worden dergelijke logistische regressies uitgevoerd. Dit lijkt ons voor het Ministerie van Justitie een onwenselijke situatie. Het komt ons voor dat het Ministerie baat heeft bij een situatie waarbij elke marktpartij die RR onderzoek in haar opdracht uitvoert, ook logistische regressies kan uitvoeren.

Voorstellen voor onderzoek

We onderscheiden voorstellen voor onderzoek ten behoeve van professionalisering van de bestaande bètaversie van het analyseprogramma en voorstellen voor onderzoek ten behoeve van uitbreiding met nieuwe opties.

Professionalisering.

Wij denken hier aan:

1. meer gebruikersgemak bij de bediening, bijvoorbeeld m.b.t. de behandeling van missing data.
2. Nu moeten er in het programma bij elke nieuwe analyse enkele zaken in het programma worden veranderd om het programma goed te laten lopen. Dit kost tijd, maar kan ook tot fouten leiden.
3. Bij een te groot aantal verklarende variabelen loopt het programma soms vast. Waarschijnlijk zit dit in de aard van het probleem, d.w.z. het is geen programmafout maar een gevolg van kenmerken van de likelihood. Het programma zou hiervan wel de oorzaak dienen te rapporteren.
4. Het maken van een gebruikersvriendelijke handleiding..

Zinvolle uitbreidingen

Wij denken hier aan:

1. het door het programma laten uitrekenen van effectmaten voor van tevoren door de gebruiker op te geven waarden van de verklarende variabelen
2. De mogelijkheid om gewichten in te voeren. Momenteel speelt het volgende probleem in het SoZaWe onderzoek. In de berekening van kans op regelovertreding wordt gebruik gemaakt van gewichten die, op basis van populatiekenmerken, aan de respondenten in de steekproef zijn toegekend. Hierdoor krijg je zuivere schattingen van kansen op regelovertreding. Echter, in het analyseprogramma voor logistische regressieanalyse kan nog niet worden gewogen. Hierdoor levert dit programma geen zuivere schattingen op in de tabel met effectsizes. Door gebruik te maken van gewichten levert dit programma wel zuivere schattingen op, maar deze mogelijkheid biedt dit programma tot dusverre niet.
3. Zinvolle uitbreidingen die ook beschikbaar zijn in gewone programma's voor logistische regressie, bijvoorbeeld classificatietabellen.

3.4. Telvariabelen

Stand van zaken

Er is in het kader van onderzoek voor het Ministerie van SoZaWe een statistische procedure ontwikkeld voor het tellen hoeveel regels een bepaalde persoon heeft overtreden. Een dergelijke telvariabele is beleidsmatig van belang (hoeveel personen overtreden 1 regel, hoeveel 2 regels, hoeveel 3, enzovoort; maar ook: hoeveel personen overtreden *minimaal* 1 regel) maar ook om psychometrische redenen (een somscore heeft een hogere betrouwbaarheid dan losse scores indien de scores een zelfde onderliggend concept reflecteren).

Hoe wordt zo'n telvariabele geconstrueerd? Omdat dit nog nergens is gedocumenteerd besteden we er hier enige aandacht aan.

We werken dit uit voor het design dat wordt gebruikt bij SoZaWe, maar het Monitor design van het Ministerie van Justitie kent een vergelijkbare uitwerking. Voor een RR variabele over fraude geldt dat

$$\pi (r) = P(r|f)\, \pi(f) + P(r|n)\, \pi(n)$$

d.w.z. de kans op 'rood' ($\pi (r)$) is gelijk aan de kans op een rode kaart gegeven de fraude stapel (dus $P(r|f)$) maal de kans dat de kleur van de kaart uit de fraudestapel dient te worden gerapporteerd ($\pi(f)$, d.w.z. de kans op fraude), plus de kans op een rode kaart gegeven de geen-fraude stapel (dus $P(r|n)$) maal de kans dat de kleur van de kaart uit de geen-fraudestapel dient te worden gerapporteerd ($\pi(n)$). De kansen $P(r|f)$ en $P(r|n)$ zijn bekend, namelijk $P(r|f) = .8$ en $P(r|n) = .2$. Hieruit volgt dat we deze vergelijking eenvoudig kunnen herschrijven als

$$\pi (f) = [\pi(r) - .2] / .6.$$

Indien we voor $\pi(r)$ de proportie gerapporteerde 'rood'-antwoorden invullen, vinden we de schatting van de kans op fraude. Merk echter op dat dit niet goed gaat als de proportie gerapporteerde 'rood'-antwoorden kleiner is dan .2, want dan is de schatting voor kans op fraude negatief! Voor afzonderlijke vragen komt dit bij het SoZaWe-onderzoek niet voor. Indien dit wel voor zou komen, zetten we de negatieve schatting op nul. De aldus verkregen schatting is de zgn. maximum likelihood schatter.

We gaan nu laten zien hoe we komen tot een telvariabele indien er meer dan een RR-variabele is. We doen dit voor twee randomized respons vragen. We noteren de kans op ('rood', 'rood') als $\pi (rr)$, en evenzo $\pi (rz)$, $\pi (zr)$ en $\pi (zz)$. Indien de randomized reponse vragen betrekking hebben op twee vormen van fraude, dan kunnen we de kansen op al dan niet fraude noteren als $\pi (ff)$, $\pi (fn)$, $\pi (nf)$ en $\pi (nn)$. De kansen op ('rood', 'zwart') zijn gerelateerd aan de kansen op fraude/ niet fraude door middel van kansen als $P(rr|ff)$, d.w.z. de kansen op een antwoord ('rood', 'rood') indien iemand heeft gefraudeerd voor beide vragen. De volgende vier vergelijkingen kunnen worden opgesteld:

$$\pi (rr) = P(rr|ff)\, \pi(ff) + P(rr|nf)\, \pi(nf) + P(rr|fn)\, \pi(fn) + P(rr|nn)\, \pi(nn)$$

$$\pi (rz) = P(rz|ff)\, \pi(ff) + P(rz|nf)\, \pi(nf) + P(rz|fn)\, \pi(fn) + P(rz|nn)\, \pi(nn)$$

$$\pi (zr) = P(zr|ff)\, \pi(ff) + P(zr|nf)\, \pi(nf) + P(zr|fn)\, \pi(fn) + P(zr|nn)\, \pi(nn)$$

$$\pi (zz) = P(zz|ff)\, \pi(ff) + P(zz|nf)\, \pi(nf) + P(zz|fn)\, \pi(fn) + P(zz|nn)\, \pi(nn)$$

De kansen als $P(rr|ff)$ zijn bekend, want voor 1 RR vraag geldt bijvoorbeeld in het onderzoek van SoZaWe dat $P(r|f) = .8$, en voor twee vragen geldt dat $P(rr|ff) = P(r|f)P(r|f) = .8*.8 = .64$. Wij kunnen nu de kansen $\pi(ff)$, $\pi(fn)$, $\pi(nf)$ en $\pi(nn)$ schatten door voor de uit de kansen $\pi(rr)$, $\pi(rz)$, $\pi(zr)$ en $\pi(zz)$ de geobserveerde proporties $p(rr)$, $p(rz)$, $p(zr)$ en $p(zz)$ in te vullen. Dit komt neer op het oplossen van een stelsel van vier onbekenden uit vier vergelijkingen, onder de restrictie dat de schattingen van $\pi(ff)$, $\pi(fn)$, $\pi(nf)$ en $\pi(nn)$ groter of gelijk zijn aan nul.

Uit deze kansen $\pi(ff)$, $\pi(fn)$, $\pi(nf)$ en $\pi(nn)$ kunnen we nu de kansen op twee keer fraude, een keer fraude en nul keer fraude afleiden als $\pi_f(2) = \pi(ff)$, $\pi_f(1) = \pi(fn) + \pi(nf)$ en $\pi_f(0) = \pi(nn)$. Deze methode is gemakkelijk te veralgemeniseren naar de situatie van meer dan twee randomized response items.

Hieronder staan twee tabellen afkomstig uit een notitie van Van der Heijden voor de begeleidingscommissie voor het SoZaWe-onderzoek, gedateerd 12-12-2000.

4| WAO, Omvang regelovertreding per categorie (n=1308).

Kleinere klusjes voor vergoeding niet doorgegeven (afgelopen 12 maanden)	14%
Baan of uitzendwerk niet doorgegeven (afgelopen 12 maanden)	2%
Zwart werk naast uitkering (afgelopen 12 maanden)	6%

5 | Aantal overtreden regels in de WAO die betrekking hebben op inkomen

Aantal overtreden regels:	Regels mbt inkomen uit arbeid
Géén	86%
1 regel	3%
2 regels	2%
3 regels	8%

In de bovenste tabel staat voor elk van drie gevoelige gedragingen, de kansschattingen, bijvoorbeeld om iemand aan te treffen die 'zwart' werkt naast een uitkering (6%). In de onderste tabel staat de kansschatting om iemand aan te treffen die 0, 1, 2 of 3 regels overtreedt. De kansschatting op regelovertreding in de bovenste tabel zijn echter niet te rijmen met de kansschattingen van het aantal overtreden regels. Immers, hoe kan bijvoorbeeld 8 % van de mensen alle drie de regels m.b.t. inkomen uit arbeid overtreden, terwijl zwarte fraude (de derde regel) slechts door 6 % wordt overtreden? Je zou verwachten dat als voor alle drie de regels 8 % wordt gevonden, elk van de eerste drie percentages in de bovenste tabel minimaal 8 % is.

De oorzaak van dit probleem zit in de schattingsmethode. Hierboven gaven we al aan dat het schatten neerkomt op het oplossen van een stelsel van vier onbekenden uit vier vergelijkingen, onder de restrictie dat de schattingen van $\pi(ff)$, $\pi(fn)$, $\pi(nf)$ en $\pi(nn)$ groter of gelijk zijn aan nul. Juist deze restrictie veroorzaakt deze problemen, en zonder deze restrictie zou dit probleem niet bestaan. Toch kan men niet zomaar deze restrictie laten varen, omdat dan kansschattingen gaan ontstaan van $\pi(ff)$, $\pi(fn)$, $\pi(nf)$ en $\pi(nn)$ die negatief kunnen zijn.

Het gevolg is dat, bij het gebruik van momentschatters negatieve schattingen ontstaan, en dat bij het gebruik van maximum likelihood schatters multivariate kansschattingen ontstaan die niet te rijmen zijn met univariate kansschattingen. Er zijn verschillende oorzaken te noemen voor het verkrijgen van schattingen die negatief zouden worden als je er niets aan zou doen. Een eerste mogelijke oorzaak is toeval. Voor het univariate geval is dit al gemakkelijk uit te leggen. Het kan bijvoorbeeld dat men vrijwel steeds een zwarte kaart trekt, ongeacht de stapel waaruit men trekt. Hierdoor wordt de kans op 'rood' kleiner dan .2 (zie boven). Indien de steekproef echter voldoende groot is, kan toeval niet z'n grote rol spelen, al dient gezegd te worden dat we te doen hebben met gevoelig gedrag en de kans op dit gevoelig gedrag vaak dicht bij .00 zal liggen. Met andere woorden, gevoelig gedrag is per definitie weinig voorkomend, en bij weinig voorkomend gedrag is er weinig toevallige fluctuatie nodig om de kansen onder de .00 te brengen. Een tweede mogelijke oorzaak is dat er een groep personen is die de randomized respons procedure niet volgt (*'cheaters')*. Hierbij kun je denken aan de mogelijkheid dat een groep in de gaten heeft dat 'rood' vaker 'ja' impliceert dan 'zwart', en men daarom 'zwart' zegt, ook al is de kleur die men eigenlijk dient op te geven 'rood'.

Voor deze laatste mogelijkheid is er enige evidentie in de data van de SoZaWe-studie. Dit probleem doet zich niet voor bij univariate RR gegevens, maar wel wanneer we meer dan een RR variabele tegelijkertijd bestuderen, zoals we doen bij het construeren van de tel variabelen. De onderstaande tabel illustreert dit. We hebben daar gegevens voor de combinatie van drie RR-items. Stel we hebben de schattingen van kansen π(fff), π(ffn), π(fnf), π(fnn), π(nff), π(nfn), π(nnf) en π(nnn) op al dan niet fraude voor de combinatie van deze drie RR-items. Sommige van deze schattingen zullen nul zijn gemaakt in de schattingsprocedure. Je kunt dan met formules zoals hierboven schattingen van kansen op ('rood' /'zwart') combinaties uitrekenen, dus schattingen van π(rrr), π(rrz), π(rzr), π(rzz), π(zrr), π(zrz), π(zzr) en π(zzz). Indien je deze kansen met de steekproefomvang vermenigvuldigt, dan krijg je de geschatte frequenties in onderstaande tabel. Je kunt die vergelijken met de frequenties die je hebt geobserveerd. Je ziet dan dat meer mensen 'zzz' hebben geantwoord dan geschat op basis van het model (geobserveerd 668, geschat slechts 587). Indien er *inderdaad* een groep is die systematisch 'zwart' antwoordt, ongeacht wat geantwoord zou moeten worden volgens de spelregels, dan zou dit het patroon van frequenties in tabel 6 kunnen opleveren.

6 | Geobserveerde frequenties, en frequenties geschat op basis van de drie afzonderlijke variabelen gecombineerd; Regels m.b.t. inkomen uit arbeid, WAO

Patroon	Geobserveerd	Geschat
Rrr	66	66
Rrz	67	61
Rzr	68	65
Rzz	169	175
Zrr	52	51
Zrz	95	151
Zzr	123	152
Zzz	668	587

Tot slot een opmerking over de validiteit van RR-data. Indien grote groepen individuen systematisch 'zwart' zouden zeggen, tast dat natuurlijk de validiteit van de RR-methode aan. Het effect in tabel 6 lijkt echter niet dusdanig groot dat we die vrees dienen te hebben.

Een tweede opmerking is dat in het onderzoek van SoZaWe steeds is aangegeven dat de univariate kansschattingen die worden gerapporteerd waarschijnlijk onderschattingen zijn van de werkelijke hoeveelheid regelovertreding. De resultaten in deze paragraaf laten niet alleen zien wat de oorzaak van deze onderschatting kan zijn, mogelijk kan die onderschatting ook worden gekwantificeerd en/ of geëlimineerd (zie hieronder). Zonder in modieuze termen te willen vervallen is dit systematisch 'zwart' zeggen dus niet zozeer een bedreiging als wel een kans om een probleem aan te pakken!

Desiderata

Om een aantal redenen dient onderzoek te worden gedaan naar het probleem van de telvariabele:

1. Telvariabelen zijn beleidsmatig van groot belang voor het beantwoorden van de vraag hoeveel regels per persoon zijn overtreden, en, hierdoor ook hoeveel mensen minimaal één regel overtreden.
2. Losse vragen hebben over het algemeen een lage betrouwbaarheid (in psychometrische zin). Hierdoor is het soms lastig een relatie tussen zo'n losse vraag en een andere vraag significant te krijgen, terwijl die relatie er in de populatie wel is (lage power). Indien vragen hetzelfde meten, neemt de betrouwbaarheid van de somscore van die variabele (hier: telling) toe. Hierdoor kun je bij telvariabelen vaker relaties met andere variabelen vaststellen. Men kan bij andere variabelen bijvoorbeeld denken aan variabelen uit de T11.
3. De problematiek rondom telvariabelen biedt de mogelijkheid om de onderschatting in RR-onderzoek te kwantificeren en/ of te elimineren. Door dit onderzoek kan RR-onderzoek mogelijk meer valide uitslagen gaan opleveren.

Onderzoeksvoorstellen

De problematiek van de telvariabele leidt tot een groot aantal vragen waarvan de onderzoekbaarheid geëxploreerd is in Van der Heijden en Van den Hout (2001, lezing voor capaciteitsgroep M&S). Te onderzoeken vragen zijn:

1. Eerste analyse van de problematiek: in hoeverre speelt dit een rol bij andere randomized response methoden (m.n. ‘forced response methoden’) die tot nu toe zijn gebruikt? Dit kan een criterium zijn om een bepaalde methode te prefereren boven een andere. Er dient een vergelijking gemaakt te worden met methoden gehanteerd in onderzoek begeleid door het Ministerie van Justitie, en ander onderzoek uitgevoerd bij SoZaWe. Hiertoe dienen data opnieuw te worden geanalyseerd.
2. Tweede analyse van de problematiek: bepaling van significantie van afwijking van designkansen. Momenteel kunnen wij constateren dat de waargenomen frequenties van (‘zwart’-‘zwart’-‘zwart’) hoger zijn dan de designkansen voorschrijven. D.w.z. stel dat er geen enkele regelovertreder is. Je verwacht dan een proportie van .2 antwoorden ‘rood’. Door het toeval kan deze proportie lager zijn, want het kan dat door het toeval er vaker een rode kaart is getrokken in de linkerstapel. De relevante vraag is hier: wanneer besluit je nu dat de afwijking zo groot is dat deze de designkansen overstijgt? We hebben zo een norm waarmee we kunnen bepalen of de afwijking zo groot is, dat er wel sprake van ‘cheaters’ moet zijn.
3. Derde analyse van de problematiek: uit de psychometrie is bekend dat het maken van telvariabelen leidt tot een verhoogde betrouwbaarheid (en daardoor een verhoogde power) indien de variabelen die de grondslag voor de telvariabele vormen, hetzelfde meten. Anderzijds, indien de onderliggende variabelen niet hetzelfde meten, dan zal het creëren van een telvariabele niet leiden tot verhoging van de betrouwbaarheid. In deze context zou een mogelijke invulling van dit “hetzelfde meten” iets kunnen zijn als “neiging tot regelovertreding”. Een model waarmee je dit zou kunnen onderzoeken bij gewone (d.w.z. niet –RR) variabelen is latente klasse analyse. In dit project zou dan ook een RR-variant van latente klasse analyse ontwikkeld dienen te worden. De modelspecificatie hiervoor is als volgt. Stel X is de latente variabele, A,B en C zijn RR-variabelen, en A*, B* en C* zijn de hiermee corresponderende werkelijke regelovertredingen. Het latente klasse model (geformuleerd als een loglineair model met een latente variabele) is dan (XA*)(XB*)(XC*)(A*A)(B*B)(C*C), en dit model is te schatten met het EM-algoritme. Daarnaast dient onderzocht te worden of een maat voor betrouwbaarheid van de telvariabele in de context van RR ontwikkeld kan worden.
   Met andere woorden: deze derde analyse levert op:
   a. Een RR-variant van latente klassenanalyse om te onderzoeken in hoeverre de variabelen die de telvariabele vormen, hetzelfde meten;
   b. Een maat voorde betrouwbaarheid van die telvariabele en het gemeten ‘construct’.
4. Vierde analyse van de problematiek: regressiemodellen ontwikkelen waarbij de telvariabele de afhankelijke variabele is. Zoals hierboven aangegeven, een telvariabele is beleidsmatig en psychometrisch interessant. Een volgende logische vraag is hoe verklarende variabelen (zoals factoren uit de ‘Tafel van 11’) gerelateerd zijn aan de telvariabele. Hiertoe dient een regressiemodel ontwikkeld te worden, dat ermee rekening houdt dat de afhankelijke variabele verwijst naar randomized response variabelen. Het zgn. ordinale regressiemodel lijkt hiertoe het meest geëigend. Dit dient bekeken te worden voor verschillende randomized response methoden, bijvoorbeeld dobbelsteenmethode (‘forced response’), en kaartmethode (‘Kuk’).
5. Vijfde analyse van de problematiek: het probleem omzeilen: Ontwikkeling van een andere statistische methode voor het creëren van telvariabelen. Momenteel worden telvariabelen gemaakt via het aantal gegeven ‘zwart’-antwoorden in plaats van alle combinaties van ‘rood’-zwart antwoorden. Dus, bij 3 regels zijn er 8 antwoorden. Nu worden er 8 regelnalevingscombinaties geschat (bijvoorbeeld “wel-niet-niet’ is een van de acht). Omdat de interesse naar een telling uitgaat, lijkt het echter beter om uitsluitend op basis van het aantal ‘rood’ antwoorden het aantal “wellen” te schatten. Bij drie regels zijn er vier kansen (0 keer “wel” tot drie keer “wel”), en dit vermindert het probleem dat kansen negatief zouden kunnen worden. Dit is als volgt aan te pakken. We doen dit weer voor twee RR-vragen. Door het invullen van de designkansen in bovenstaande vergelijkingen vind je

$$\pi(rr) = .8^2\,\pi(ff) + .2*.8\,\pi(nf) + .8*.2\,\pi(fn) + .2^2\,\pi(nn)$$

$$\pi(rz) = .8*.2\,\pi(ff) + .2^2\,\pi(nf) + .8^2\,\pi(fn) + .2*.8\,\pi(nn)$$

$$\pi(zr) = .2*.8\,\pi(ff) + .8^2\,\pi(nf) + .2^2\,\pi(fn) + .8*.2\,\pi(nn)$$

$$\pi(zz) = .2^2\,\pi(ff) + .8*.2\,\pi(nf) + .2*.8\,\pi(fn) + .8^2\,\pi(nn)$$

Dit is als probleem te reduceren. Definieer hiertoe $\pi_r(k)$ als de kans op k keer 'rood' en $\pi_f(k)$ als de kans op k keer 'fraude'. Dan geldt dat

$$\pi_r(2) = .8^2\,\pi_f(2) + .8*.2\,\pi_f(1) + .2^2\,\pi_f(0)$$

$$\pi_r(1) = 2*.8*.2\;\pi_f(2) + (.8^2 + .2^2)\;\pi_f(1) + 2*.8*.2\;\pi_f(0)$$

$$\pi_r(0) = .2^2\;\pi_f(2) + .8*.2\;\pi_f(1) + .8^2\,\pi_f(0)$$

Te bekijken voor verschillende randomized response methoden, bijvoorbeeld dobbelsteenmethode ('forced response'), en kaartmethode ('Kuk').

6. Het percentage cheaters schatten: Het ontwikkelen van een zgn. 'mengsel benadering' die rekening houdt met cheating. In een 'mengselbenadering' neem je aan dat de steekproef die je ziet eigenlijk bestaat uit twee deelsteekproeven. We nemen hier aan dat er twee groepen mensen zijn: zij die louter 'zwart' zeggen en zij die eerlijk de RR-procedure volgen. Dit werkt als volgt: we gaan te werk net als in onderzoeksvoorstel 3, maar schatten nu ook π_{ob}, d.w.z. de kans op obstructie van de RR-spelregels,

$$\pi_r(2) = (1 - \pi_{ob}) * \{.8^2\,\pi_f(2) + .8*.2\,\pi_f(1) + .2^2\,\pi_f(0)\}$$

$$\pi_r(1) = (1 - \pi_{ob}) * \{2*.8*.2\;\pi_f(2) + (.8^2 + .2^2)\;\pi_f(1) + 2*.8*.2\;\pi_f(0)\}$$

$$\pi_r(0) = \pi_{ob} + (1 - \pi_{ob}) * \{.2^2\;\pi_f(2) + .8*.2\;\pi_f(1) + .8^2\,\pi_f(0)\}$$

en het is te verifiëren dat deze kansen optellen tot een. We merken nog op dat dit probleem niet geïdentificeerd is, omdat we vier parameters schatten uit drie datapunten. Echter, doordat de parameters gerestricteerd (groter dan nul en optellen tot 1) zijn, kan dit wel zinvol te interpreteren onder- en bovengrenzen van de parameter π_{ob} opleveren. Dergelijke onder- en bovengrenzen geven aan hoe goed de RR-procedure is gevolgd, en geven hiermee een indicatie van de kwaliteit van de verzamelde data. Daarnaast krijgen we betere schattingen van $\pi_f(0)$, $\pi_f(1)$ en $\pi_f(2)$, namelijk schattingen die ontdaan zijn van 'cheaters'.

4. De T11 in surveyonderzoek. Wat kan er beter?

Henk Elffers,
NSCR

4.1. Inleiding

Deze beschouwing richt zich op onderzoek met behulp van de Tafel-van-Elf (Inspectie voor de Rechtshandhaving, 1994, 1997; Van Reenen, 2000)[7] onder de doelgroep van een wet of regel, dus onder hen wier gedrag wordt voorgeschreven door die regel, en die zich daar al of niet iets aan gelegen kunnen laten liggen. In het algemeen tracht zulk onderzoek door middel van een inventarisatie met behulp van de T11 na te gaan welke factoren van de perceptie van de regelgeving onder de beregelden samenhangen met (de neiging tot) overtreden. Ik wil hier nagaan wat er kan en moet worden verbeterd aan het instrument van de T11, of aan de wijze waarop het wordt ingezet, teneinde zulk onderzoek te optimaliseren. Door middel van ordening en kritische bespreking van bestaande kennis en inzichten en met aandacht voor standaardisatie van terminologie en presentatiemethoden beschouw ik twee vragen:

1. wat zijn sterke en zwakke eigenschappen van het instrument voor het indiceren van regelnaleving?
2. op welke punten is nader onderzoek geboden om de T11 te kunnen verbeteren, of nadere informatie te verkrijgen om het inzetten van de T11 te kunnen optimaliseren? We zullen daarbij onderscheid maken tussen aanbevelingen tot nadere analyse van bestaand materiaal enerzijds, en nieuw empirisch onderzoek anderzijds

Bestaande empirische studies

Alvorens in te gaan op deze vragen is het goed de aanwezige empirische doelgroepsurveys te rubriceren. Allereerst gaat het natuurlijk om het Monitorrapport (1997), met de achterliggende rapportages

Veldkamp/Intomart (1997a, 1997b, 1997c). Deze studies beschrijven een doelgroepsurvey onder beregelden van de individuele huursubsidiewet, de bestrijdingsmiddelenwet en de precursorenwet. Aan deze rapportages zijn secundaire analyses gewijd in Ruimschotel, Elffers & Hezemans (1997), en Elffers, Van der Heijden & Hezemans (aangeboden voor publicatie). Daarnaast zijn er twee doelgroepsurveys betreffende de warenwet en de taxiwet (Veldkamp/Intomart, 2000a, 2000b; Prinsen & Visser , 2000a, 2000b), met secundaire analyses door Elffers & Van Giels (2000a, 2000b), en Elffers (2000b). Voorts is er een onderzoek van regelovertreding in de ABW, WAO en WW (Van Gils, Van der Heijden en Rosebeek, 2001). Tenslotte noem ik een studie van Elffers, Ruimschotel & Barwegen (1997; vergelijk ook Elffers & Ruimschotel, 1997) onder een beperkte groep door de wet melding oneigenlijke transacties beregelden.

Voor de goede orde: in veel van boven aangehaalde studies wordt meer gedaan dan alleen een doelgroep-survey, en daarnaast zijn nog meer empirische studies waarin de T11 is gebruikt voorhanden, maar dan gaat het niet om doelgroepsurveys, maar om inventarisatie van nalevingkennis onder handhavingexperts (o.a. Inspectie voor de Rechtshandhaving, 1994; Elffers & Klaasen, 1997; Ruimschotel & Damaisjah, 1998; Van Dijk, Elffers, Hessing & Hoogenboom, 1999; Van Dijk en Vaneman, 2000; Klaasen, 2000). Deze studies komen hier slechts zijdelings aan de orde. Tenslotte noem ik OECD (1999) als een publicatie die de kern van de T11 in beschrijvende zin behandelt.

Ik constateer wel dat het niet gemakkelijk is toegang te krijgen tot deze literatuur, omdat het deels om rapporten gaat die bij de evaluatie van bepaalde wetten zijn gebruikt, zonder openbaar gepubliceerd te zijn. Het moet ook niet uitgesloten worden geacht dat er meer T11 –doelgroepstudies zijn verricht, die zich aan het zicht hebben onttrokken. Dit leidt tot de **eerste aanbeveling** dat er ergens, bijvoorbeeld bij het Expertisecentrum Rechtshandhaving ERH, een literatuurbank wordt onderhouden, waar alle betrokken rapporten, liefst ook elektronisch, voorhanden zijn en een up-to-date literatuurlijst wordt bijgehouden. Daarbij is ook een beschikbaar komen van de databestanden van de betreffende studies, in ieder geval van de doelgroepstudies, bijvoorbeeld bij het Steinmetzarchief of bij het ERH, alleszins aan te raden.

Wat is eigenlijk een T11 –survey-studie?

Opvallend in de hierboven aangehaalde literatuur is dat er eigenlijk helemaal geen sprake is van een gestandaardiseerd instrument. Zelfs de namen van de elf dimensies liggen niet vast: er zijn heel wat verschillende benamingen en varianten in omloop, en in sommige artikelen wordt, al of niet met enige toelichting, soms een hele dimensie weggelaten. De meeste rapporten gebruiken een eigen, aan de betreffende regelgeving aangepaste, variant van de T11 en het is lang niet altijd duidelijk op welke gronden en hoe die specifieke versie tot stand is gekomen. Tafel-van-Elf is eigenlijk meer de aanduiding van een gedachteschema om over handhaving en naleving te denken, dan dat het een concreet meetinstrument is. Teneinde in ieder geval binnen deze rapportage daaraan geen onhelderheid toe te voegen geef ik hier de door mij gebruikte benamingen weer:

De TAFEL van ELF
T_1: bekendheid, helderheid
T_2: kosten en baten van naleven en overtreden
T_3: acceptatie specifiek
T_4: acceptatie generaal (normconformisme)
T_5: sociale norm / informele controle
T_6: kans op spontaan uitkomen
T_7: controlekans
T_8: controlediepgang (detectiekans)
T_9: controleselectiviteit (targeting)
T_{10}: sanctiekans
T_{11}: sanctie-zwaarte

De **tweede aanbeveling** voor het verantwoord inzetten van de T11 is om een degelijke standaardbeschrijving te vervaardigen van wat exact verstaan moet worden onder “Tafel-van-Elf-onderzoek” , hetgeen in twee delen uiteenvalt: een beschrijving van het meetinstrument enerzijds, en een beschrijving van de analysemethoden anderzijds.

Teneinde het vervolg van het betoog te structureren zal ik hier een voorzet doen, en de hoofdlijnen van een zodanige standaardbeschrijving proberen te geven.

Voor zo'n handleiding lijkt het goed een onderscheid te maken tussen twee typen analyses die men aantreft bij T11 - onderzoek in de literatuur. Allereerst gaat het dan om *'kale T11 – inventarisatie'* ofwel het opstellen van een nalevingprofiel: hoe dient men dat precies te doen (er zijn nogal wat varianten in omloop), en hoe moet men het resulterende profiel interpreteren. Het tweede type analyse is de *'correlationele T11 – inventarisatie'*, waarbij het verband tussen T11 –scores uit het profiel en een maat voor feitelijke naleving wordt bestudeerd. Ik wijs er overigens vast op dat aan de zinvolheid van 'kale' studies door sommigen getwijfeld wordt.

De structuur van een T11 - studie

In het navolgende stel ik een raamwerk voor om het gebruik van de T11 mee te structureren, dat voor zo'n handleiding als leidraad kan dienen. Dat raamwerk is deels gedestilleerd uit wat men in de verschillende rapporten tegenkomt, maar ook deels op een 'reconstructie' van wat er mogelijkerwijs bedoeld is, of –in mijn ogen– verstandig is.

Het opzetten van een T11 –survey zou moeten bestaan uit twee fases, die elk weer in drie onderdelen uiteenvallen.

Stap 1. OPZET EN DATAVERZAMELING

a. specificatie of uitvouwing
b. operationalisatie (bij een kale T11 –inventarisatie gaat het alleen om de operationalisering van de T11 –dimensies, bij een correlationele T11 –inventarisatie daarnaast ook om operationalisatie van de afhankelijke variabele, de nalevingrapportage
c. dataverzameling

Stap 2. ANALYSE EN INTERPRETATIE

a. condensatie
b. het maken van een T11 –profiel
c. interpretatie
- bij een kale T11 –inventarisatie vooral van het T11 –profiel
- bij een correlationele T11 –inventarisatie ook van de samenhang T11 –naleving

Laat ons eerst eens zien wat de literatuur ten aanzien van opzet en dataverzameling in T11 –surveys biedt. De in omloop zijnde beschrijvingen concentreren zich op de theoretische claim dat de elf dimensies van de tafel de gehele ruimte opspannen die de instigaties en remmingen tot regelovertreding zouden vormen (o.a. Ruimschotel, Van Reenen & Klaasen, 1995, 1996, 2000). Beslist zwakker is de uitwerking in termen van een concrete operationalisering. De enige studie die hier systematisch aandacht aan besteedt is Bronner&Emmering (2000). Vragen die in dit verband nog betrekkelijk weinig onder ogen zijn gezien zijn:
- hoe moeten de dimensies van de tafel in concrete vragen worden omgezet?
- welke vormen van operationalisering verdienen de voorkeur?
- zijn er gegevens over betrouwbaarheid en validiteit van de mogelijke vormen van operationalisatie?

We kunnen ons in dit verband ook afvragen of het mogelijk is te komen tot een set standaardregels om tot een operationalisatie te komen, die onafhankelijk van de te bestuderen regel met succes kan worden ingezet.

Een tweede observatie naar aanleiding van de literatuur is dat er ook geen door de verschillende auteurs gedeelde vorm schijnt te bestaan waarin men T11 –onderzoek presenteert en interpreteert. Dit onderdeel van het gebruik van de T11 wordt nogal stiefmoederlijk behandeld in de literatuur. Eigenlijk geven alleen Inspectie voor de Rechthandhaving (1997) en Ruimschotel (2000) een aanzet, vergelijk ook Elffers (2000a). Toch vind ik dat dergelijke handleidingen nog te veel een insiders-optiek hebben, om niet te zeggen een ontwikkelaaroptiek. Elffers & Van Giels (2000a, 2000b) geven een methode om, secundair, binnen de doelgroep subgroepen te identificeren die een afwijkend T11 – profiel hebben, maar ook deze rapporten zijn niet als handleiding geschreven.
De **derde aanbeveling** is dan ook om een heldere handleiding te ontwikkelen die analyse en interpretatie van de uitslagen van Tafel-van-Elf-onderzoek weergeeft.

In de volgende paragrafen zal ik kort ingaan op enkele van de hier boven onderscheiden stappen van een T11-surveystudie. Aan stap 1.3, dataverzameling ga ik voorbij, aangezien dat in het eerste hoofdstuk van dit rapport wordt behandeld. Sommige onderdelen komen in meer dan één subparagraaf aan de orde.

4.2. Opzet en dataverzameling

Specificatie, uitvouwing

Een aanzet om enige helderheid te brengen in wat precies moet worden verstaan onder een T11 –doelgroepsurvey, en hoe daarmee om te gaan, wordt gegeven door Elffers, Van der Heijden & Hezemans (aangeboden voor publicatie). Ik volg hier grotendeels hun beschouwingen.

De T11 –theorie pretendeert dat de hierboven gegeven elf dimensies het hele bereik van instigaties en remmingen opspannen dat naleving kan verklaren. In de geest van de rationele keuzetheorie kunnen we regelnaleving zien als gedreven door een evaluatiefunctie van de verschillende gedragsalternatieven (naleving of overtreding): wie onder de regel valt, houdt zich eraan of overtreedt, al naar gelang de verwachte uitkomst van naleving of overtreding het gunstigst is. We stellen ons daarbij voor dat die verwachte uitkomst gegeven wordt door een evaluatiefunctie, die alleen van de elf kenmerken van de T11 afhangt, ofschoon we de exacte vorm van die functie niet kennen. Nu is die claim minder verreikend dan het op het eerste gezicht lijkt, omdat alle dimensies eigenlijk niet meer zijn dan labels voor op zich alweer ruime gebieden, en in veel T11 -toepassingen een of meer concepten weer op verscheidene manieren worden geoperationaliseerd. Als voorbeeld: de kosten/ baten factor T_2 kan heel wel een viervoudige operationalisatie krijgen als:

- T_{2a}: *kosten van naleving*, dat wil zeggen de hoeveelheid werk of geld die zijn verbonden aan het naleven van de regels, zoals bijvoorbeeld administratieve lasten;
- T_{2b}: *kosten van overtreding* (dit concept wordt soms aangeduid met de term: "overtredingdrempel")
- T_{2c}: *baten van overtreden*;
- T_{2d}: *baten van naleving*.

Het is geenszins gezegd dat alle vier 'subconcepten' nodig zijn, en ik benadruk dat een T11 –studie opzetten begint met het maken van een verdedigbare keus ten aanzien van welke dimensies en subdimensies zullen worden geoperationaliseerd, en in welke mate van diepgang. Ik wijs er hierbij op dat de claim dat de T11 voor alle regels voldoende is om het geheel aan instigerende en remmende factoren te beschrijven allerminst betekent dat die beschrijving voor elke regel identiek zal zijn. Voor sommige regels zullen bepaalde dimensies of subdimensies wel relevant zijn, voor andere niet. Als voorbeeld noem ik onderzoek naar witwassen door grote reguliere financiële instituties zoals banken. Kosten van naleving van de anti-witwasregels zijn geducht, omdat men in het algemeen zogeheten 'compliance officers' moet aanstellen. Daarentegen is overtreden van die regel gratis: dat betekent immers dat men verzuimt bepaalde procedures in werking te zetten. In dat geval is T_{2b} niet van belang. Ook de sanctiezwaartedimensie, T_{11},wordt vaak opgedeeld, in financiële boetes enerzijds, en de reputatieschade die men door openlijke bestraffing oploopt. Bij sommige regels is zelfs een hele dimensie niet van belang. Zo is T_8 onbelangrijk in het geval van de verplichting bij donker op de fiets een brandend achterlicht te voeren: *als* een agent controleert, is er geen enkele twijfel aan dat hij in voorkomende gevallen kan vaststellen dat men zich niet aan die regel houdt.

Ik noem het in een bepaald onderzoek naar bepaalde regels maken van een keus uit de T11 –(sub)– dimensies de fase van de **SPECIFICATIE** en, omdat sommige dimensies van de T11 als het ware worden uitgepakt, **UITVOUWING VAN T11**. Zo bestaat een T11 studie uit het allereerst vaststellen welke (sub)-dimensies, al of niet ontvouwd, meegenomen worden, onmiddellijk gevolgd door een operationalisatie van de gekozen subdimensies. Zo'n operationalisatie kan uiteraard van verschillende diepgang zijn, van een enkel Likert-schaaltje tot een complete schaal met verschillende items waarop een schaalanalyse wordt toegepast.

De 'specifikatie-uitvouwing'– gedachte leidt tot de **vierde aanbeveling**: in een handleiding voor de methode van de T11 dient men expliciet aandacht te besteden aan de noodzaak dat elke toepassing aanvangt met reflectie op de voor die specifieke toepassing op te nemen dimensies, en op de mate waarin binnen dimensies nader onderscheid wordt gemaakt

Operationalisatie: betrouwbaarheid

Aan het vraagstuk hoe T11 –dimensies betrouwbaar –in psychometrische zin– te operationaliseren is betrekkelijk weinig aandacht besteed. Bij veel van de bestaande toepassingen van de T11 wordt elk van de (sub-) dimensies met meestal slechts één, of hoogstens enkele items geoperationaliseerd. De betreffende items zijn vaak vijfpunt-Likertschalen, soms ook kansschattingen op een 100%-schaal. Vanuit het oogpunt van een betrouwbare meting is het uiteraard niet op voorhand zeker dat een enkele-item-benadering betrouwbaar genoeg is. Of, scherper uitgedrukt, er is gerede twijfel of operationalisatie met slechts een enkele vraag wel betrouwbaar genoeg is. Er is dan ook alle aanleiding nauwkeurig na te gaan of gebrek aan betrouwbaarheid bij enkele item-operationalisatie zich in concrete gevallen voordoet, en of operationalisatie met meer dan een item voldoende winst oplevert. In de Monitor-studie was in die gevallen waar wel sprake was van meerdere items (meestal gevallen waarin men tot condensatie in de zin van bovenstaande paragraaf overging) soms wel, soms geen sprake van schaalbaarheid. Ontbreken van schaalbaarheid deed zich daar vooral voor in de T_2 –hoek (Elffers, van der Heijden & Hezemans, aangeboden voor publicatie). Deze auteurs vervolgden hun analyses in de T_2-hoek daarom met meerdere subdimensies.

Vijfde aanbeveling:. Bij toekomstige toepassingen waar mogelijk meervoudige operationalisatie van (sub-) dimensies toe te passen (om te analyseren in welke mate daarbij betrouwbaarheidswinst geboekt wordt). Men kan denken aan operationalisatie met meervoudige equivalente vragen, maar ook aan de vorm van een operationalisering waarbij men meerdere items die een zelfde dimensie indiceren probeert op te stellen, waarvoor het, gegeven de mening van een respondent, steeds onwaarschijnlijk wordt dat hij het met een volgend item eens zal zijn (item-response-operationalisering). Uiteraard dienen zulke gegevens dan ook voor de verdere methodeontwikkeling te worden geanalyseerd.

Zesde aanbeveling: Systematische heranalyse van alle beschikbare data om te bezien in welke mate multipele operationalisatie van een dimensie aanwijzing geeft over de mate van betrouwbaarheid van operationalisatie. Dit zal vermoedelijk een betrekkelijk geringe opbrengst hebben, omdat er niet veel multipele operationalisaties voorhanden zijn.

Operationalisatie: kansvragen

In de Monitorstudie werd ook geëxperimenteerd met meerdere soorten vraagformaten als het ging om kansvragen, bijvoorbeeld bij dimensie T_6, T_7 etc. (‘hoe groot is die kans in procenten?’ respectievelijk ‘is die kans zo groot dat u daar rekening mee houdt?’). Deze bleken (Ruimschotel, Elffers & Hezemans, 1997:34) duidelijke samenhang te vertonen, maar geenszins identiek te zijn. Er was sprake van een duidelijk formaateffect, in de zin dat een factoranalyse van een reeks items (n (sub-)dimensies, 2 formaten) twee formaatfactoren identificeerde, en niet n (sub-)dimensiefactoren (Ruimschotel, Elffers & Hezemans, 1997:36 e.v.). Bronner & Emmering (2000) rapporteerden dat de ‘rekening houden met’ – vragen voor de respondenten erg onduidelijk waren.

Ook bij de taxiwet en de warenwet is met dubbele kansvragen in verschillend formaat gewerkt. Hier betrof het categorische kansmeting (vijfpunts-schaal, heel klein – heel groot) en percentuele kansmeting (hoeveel procent). In Elffers & Van Giels (2000a; 2000b) is nagegaan in hoeverre deze formaten sterk gerelateerd zijn. Dit bleek vooral voor de wat grotere kansen problematisch te zijn: respondenten die aangaven dergelijke kansen ‘heel groot’ te achten, bleken daar –in procenten uitgedrukt– toch heel verschillende waarden mee te bedoelen, met een interkwartielafstand van tientallen procenten Bij de meeste andere studies in de literatuur is slechts één antwoordformaat gebruikt.

Deze studies overziend moeten we toch vrezen dat er twijfel bestaat aan de kwaliteit van de operationalisatie van de kansvragen. De ‘easy-way-out’ is de verschillen te accepteren, en een bepaalde formulering bij voorkeur gebruiken. Toch lijkt er wel aanleiding tot wat nader spitten in deze materie, hetgeen leidt tot de **zevende aanbeveling**: bezien of een Multitrait-Multimethod (MTMM) heranalyse van de monitorstudie-kansvragen meer inzicht in het formaat-versus-dimensie-effect kan worden verkregen, om zo te bezien of het advies bij de meting van T11 -kansdimensies tot één vraagformaat te beperken gerechtvaardigd is.

Operationalisatie: conditionele vragen

Nogal wat dimensies van de T11 adresseren conditionele gebeurtenissen. Bijna altijd gaan de vragen over wat er zou gebeuren indien iemand, of de respondent zelf, de regels wel of niet zou hebben overtreden. Soms is die conditionaliteit nog ingewikkelder. Zo wordt onder de T_8 de detectiekans bedoeld, gegeven dat

er gecontroleerd wordt en gegeven dat er wordt overtreden, bij de T_{10} de kans op een sanctie, gegeven dat een overtreding is geconstateerd, en impliciet dus ook gegeven dat er een overtreding is begaan. Dat zijn een nogal ingewikkelde concepten, die niet gemakkelijk te meten zijn. Als voorbeeld noem ik hier de T_{10}-vraag uit de warenwet-survey naar de kans op een sanctie:

> "Stel u heeft een overtreding begaan, hoe groot schat u de kans dat u gestraft zou worden of een waarschuwing zou krijgen als de overheid ook echt ontdekt dat u een overtreding zou hebben gepleegd van de regels over de bewaartemperatuur van slaatjes en gekoelde snacks."

Het is mijns inziens buitengewoon onwaarschijnlijk dat zulke formuleringen ook 'aankomen' bij de respondent.

Er zijn twee problemen te onderscheiden: het eerste betreft de vraagformulering, het tweede gaat in op de mate waarin nalevende respondenten inzicht (kunnen) hebben in de sterkte van controle-activiteit.

Eerstens is het de vraag hoe je een surveyvraag zo formuleert dat dit conditionele karakter tot de respondent doordringt. Omschrijving is al gauw erg ingewikkeld, zoals uit bovenstaand voorbeeld blijkt: *stel dat u de regels zou hebben overtreden, en stel dat de …-inspectie bij u zou komen controleren, hoe groot acht u dan de kans dat ze die overtreding ook werkelijk zouden ontdekken?* Een vraag als deze overtreedt het eerste gebod van de vragenstellerij, dat we slechts één topic tegelijk aan de orde behoren te stellen. Men mag gevoegelijk aannemen dat deze vraag voor een substantieel deel van de respondenten nagenoeg identiek is aan de vraag '*wat is de kans dat men een overtreding bij u zou aantreffen'*, een interpretatie die pertinent niet bedoeld is.

Er is zeker evidentie voor deze vrees. Allereerst is het mijn persoonlijke ervaring bij het afnemen van T11 –vragen in expertpanels van handhavingexperts dat ook binnen zo'n groep van mensen die regelmatig over diverse aspecten van handhaving nadenken, zulke vragen multi-interpretabel blijken, en vaak anders worden geïnterpreteerd dan bij de opstellers van de vragenlijst de bedoeling was. Bovendien blijkt regelmatig dat T_7, T_8, T_{10} gecorreleerd zijn in de perceptie van de respondenten. Daar is op voorhand helemaal geen reden voor. Voor sommige regels zal trouwens T_8 100% zijn, bijvoorbeeld bij snelheidsovertredingen: als er gecontroleerd wordt, is het vrijwel zeker dat te snel rijden wordt ontdekt. Toch noteren respondenten dat zelden of nooit. Een derde aanwijzing dat er hier iets wringt vinden we bij Ruimschotel, Elffers & Hezemans (1997), die concluderen dat het uiteindelijk nalevingniveau in regressieanalyses goed verklaarbaar blijkt, maar het nalevingniveau van de tussenstappen veel minder. Dat kan erop wijzen dat die tussenstappen onvoldoende worden onderscheiden door de respondenten.

Het tweede probleem is dat we ons kunnen afvragen of mensen die zich zelf zien als alvers wel in staat zijn om conditionele vragen, op de voorwaarde dat ze zouden hebben overtreden, koel te overwegen: wellicht vinden zij het onder ogen moeten zien van de conditie dat ze de regels zouden hebben overtreden –om dan te antwoorden wat er dan zou gebeuren– wel zodanig beschuldigend, dat ze er niet meer open en fris tegenaan kunnen of willen kijken. Bovendien is het niet evident wat we aanmoeten met het antwoord van zulke mensen: zij hebben immers helemaal geen ervaring met controle-effectiviteit, dus hun mening kan nauwelijks informatie geven over reële controle-effectiviteit, en hun antwoord kan wellicht veeleer als een attitudemeting ten aanzien van de regels worden gezien. Het punt is daarbij niet dat zulke respondenten geen inzicht in de door hen ervaren afschrikkende werking van het controlecomplex als zodanig kunnen geven, maar dat zij daarbinnen, bij gebrek aan ervaring, geen onderscheid kunnen maken.

Achtste aanbeveling: Ik vind dat naar de kwaliteit van het meten van conditionele kansen nader onderzoek nodig is, allereerst om na te gaan hoe erg het probleem is. Zulk onderzoek zal expliciet tot doel moeten hebben de vraagkwaliteit door te lichten, met behulp van methoden zoals van Belson (1981). Zulk onderzoek kan profijtelijk gebruik maken van een cognitief laboratorium (Snijkers, 2002), en zodoende empirisch ontleden of en in welke mate respondenten in staat zijn de verscheidene stappen in het controleproces te onderscheiden, te interpreteren zoals bedoeld, en daar een gefundeerde mening over kunnen geven. Als de diagnose uit zulk onderzoek zou blijken te zijn dat er beduidende verbeteringen nodig zijn, dan zijn twee opties het overwegen waard. Een is om de kwaliteit van de conditionele vragen belangrijk te verbeteren. Dat zal mijns inziens onvermijdelijk uitdraaien op omvangrijker vragen-batterijen, en mogelijk zelfs op vraagmethoden die de simpele surveyvraag te boven gaan. De andere optie is afzien van de pretentie om het controleproces zo sterk in stappen uiteen te gooien. Waarom willen we eigenlijk weten hoe de doelgroep het controleproces geleed in de trits [controlekans / pakkans bij controle /

sanctiekans bij betrapping]? Voor een helder beeld van het controleproces is zo'n ontleding wel nuttig, maar dan veeleer vanuit het 'objectieve' perspectief van de controle-experts. Ten aanzien van de doelgroep is het, meen ik, vaak wel genoeg te weten of de controle bedreigend is, en daarbij is een 'pakket-schatting' wel genoeg. Dat sluit ook aan bij de elders gemaakte observatie dat condensatie veelal tot clustering van de hele formele controle en sanctiebatterij leidt. Dus dan kan men afzien van een opsplitsing van T_7, T_8, T_{10} en soms zelf T_{11}. Ik denk dat deze laatste optie veruit de voorkeur verdient. Deze aanbeveling komt uiteraard neer op een doorbreken van de parallel tussen expertsessie-T11's en doelgroepsessie T11's, en een erkenning dat voor doelgroepsurveys de T11 meestentijds met meer vrucht door een 'Twat minder' kan worden vervangen. Als **negende aanbeveling** stel ik voor over deze optie in een studieconferentie van T11 – experts serieus van gedachten te wisselen. Uiteraard gaat het dan om een majeure herstructurering van de T11 –benadering voor doelgroepensurveys.

Operationalisatie: vermenigvuldiging van kansen en zwaarte

Het neoklassieke rationele keuzemodel stelt dat afschrikking gemeten kan worden als verwachte strafzwaarte. Als iemand een kansschatting van controle p geeft, en een zwaartescoring van de sanctie z, dan hoort de afschrikking gelijk te zijn aan, of een functie te zijn van het product p*z . Kan deze redenering in T11 –context worden toegepast? En in het bijzonder, kan door een passende vermenigvuldiging van de conditionele kansen en de gevreesde strafgroottes en sommatie van onafhankelijk optredende gevolgen een soort overall afschrikkingsfunctie worden geconstrueerd?
Op voorhand lijkt twijfel op zijn plaats, zowel vanwege de kwaliteit van de metingen als vanwege het conceptueel model.
Allereerst breng ik in herinnering dat hierboven al twijfels zijn geuit aan de kwaliteit van de kansschattingen in het T11 –kader, omdat alternatieve kansmetingen niet geheel blijken te sporen, wat voor deze benadering wel een conditio-sine-qua-non is: het blijft dan immers onduidelijk welke meting men zou moeten prefereren. Maar ook conceptueel lijkt het neoklassieke model te simpel te zijn, en blijken mensen in hun omgang met kansen zich niet conform dit model te gedragen. Herbert Simon verdiende met deze observatie een Nobelprijs, en de economische psychologie als geheel heeft afstand genomen van dit te eenvoudige concept (cf. Kahneman & Tversky, 1979).

Moerland (1991) heeft deze benadering in Nederlandse criminologische context uitgeprobeerd op winkeldiefstal, en concludeerde dat het geen succes was. Een goed overzicht van het feit dat p*z-benaderingen niet werken in criminologische (en waarschijnlijk ook niet in andere) context is Carroll&Weaver (1986), en overigens ook diverse andere artikelen in de bundel van Cornish&Clarke (1986).

Het is uiteraard mogelijk op de bestaande data in het T11 –kader na te gaan hoe de vermenigvuldigings-methode zich daar houdt. Mijn **tiende aanbeveling** is echter om zich deze moeite te besparen.

4.3. Analyse en interpretatie

Condensatie

De claim dat naleving van elke regel in termen van de T11 kan worden beschreven betekent allerminst dat alle dimensies ook bij elk onderzoek een eigen, welonderscheiden bijdrage leveren. Meestal zullen voor een zekere regel sommige dimensies empirisch gecorreleerd blijken te zijn, over een groep respondenten. Dat betekent dat het veelal mogelijk zal zijn verscheidene T11 –dimensies achteraf samen te nemen, omdat ze sterk verwant blijken in de perceptie van de doelgroep, zoals dan zou kunnen blijken uit schaalanalyse of factoranalyse, of op basis van a a-priori redenen. Het op deze wijze samennemen van dimensies noem ik *condensatie* van de T11. Er is dus sprake van een tegengestelde beweging: eerst wordt de T11 uitgevouwen in subdimensies, teneinde de geldigheid van de operationalisatie te verhogen, dan weer gecondenseerd om de betrouwbaarheid van de resulterende meetschalen te verhogen. Of dat laatste verantwoord is, blijkt meestal uit de doelgroepsurvey, op grond van hoge correlaties. Zo kan het voorkomen dat verschillende kosten/ baten subdimensies die eerst uit elkaar getrokken zijn, omdat dat conceptueel voor de hand ligt, en daarmee de meting voor de respondenten aantrekkelijker maakt, op basis van geobserveerde samenhang weer worden samengenomen. In een aantal studies bleken bijvoorbeeld heel wat controle en sanctiedimensies (T_6 tot T_{11}) sterk geassocieerd te zijn in de perceptie van respondenten.

Een gecondenseerde T11 biedt vaker een veel spaarzamere zowel als betrouwbaarder beschrijving van naleving en handhaving.
Merk op dat, in theorie, de wijze van condensatie, dus welke dimensies blijken samengenomen te kunnen worden, weer niet voor elke regeling identiek hoeft te zijn. Ik kom daar in de volgende paragraaf op terug.

Deze gedachten over condensatie als een wezenlijke stap in een T11 –studie leidt tot **uitbreiding van de vierde aanbeveling:** bij het opstellen van een handleiding voor de T11 dient expliciet aandacht te worden besteed aan de noodzaak of mogelijkheid tot condensatie, die mogelijk per te evalueren wet een andere groepering van dimensies met zich meebrengt.

Condensatie: wetonafhankelijke factoren?

Eerder is al uiteengezet dat er geen a a-priori reden is dat er een over wetten heen geldende T11 –structuur zou behoren te gelden. Toch blijkt in de praktijk van de condensatieanalyse (Ruimschotel, Elffers & Hezemans, 1997; Elffers, Van der Heijden & Hezemans, aangeboden voor publicatie; Elffers & Van Giels, 2000a, 2000b) vaak een weliswaar niet identieke, maar toch wel min of meer overeenkomende groepering van dimensies plaats te vinden. Zo werd voor alle drie gevallen uit de monitorstudie (twee overtredingen binnen de bestrijdingsmiddelenwet, en een overtreding bij de wet huursubsidie) telkens een kennisfactor, een sociale normfactor (met verschillende varianten), en een controle&sanctiefactor aangetroffen.
Interessant is ook te zien dat een der T11 –dimensies, T_9, op geen enkele factor laadt. Ook bij de taxi- en de warenwet werden weer niet identieke, maar toch deels parallelle complexen van factoren aangetroffen (controlekans, controlediepgang, verankering regels, sociale norm). De gevonden complexen van factoren weerspiegelen zo de herkomst van de T11 uit het Fishbein-Ajzen-paradigma
Het is daarom te overwegen (**elfde aanbeveling)** te bezien of een nadere secundaire analyse (van de Monitorstudie-datasets en de daarna beschikbaar gekomen datasets) meer inzicht kan geven of een kern van gecondenseerde T11 –dimensies kan worden aangegeven die voor een grote groep regelingen nagenoeg identiek is. Het is verstandig zich daarbij te realiseren dat het aantal vragen dat bij een bepaalde groep dimensies behoort mede bepaalt of een bepaalde factor geprononceerd naar voren komt. Bij een dergelijke secundaire analyse kan ook worden nagegaan in welke mate de ontvouwing van dimensies de eindstreep heeft gehaald, of weer 'weggecondenseerd' is. Als methodiek voor dit type analyse kan men denken aan het schatten van factormodellen met restricties over verschillende studies heen, door middel van covariantiestructuuranalyse.

Constructie van het T11 – profiel

Eerder al wees ik op het ontbreken van een communis opinio hoe uitkomsten van een T11 te interpreteren. In het algemeen wordt een "T11 –profiel" opgesteld, dat ook wel "risicoprofiel" of "nalevingsprofiel" wordt genoemd. In het rapport 'De eerste stap' (Inspectie voor de Rechtshandhaving, 1994) zien we een eerste poging om met zo'n profiel naar voren te komen, in de vorm van het aangeven welke categorie voor elke T11 –dimensie door een groep experts als van toepassing is aangewezen, via consensus, of als de het vaakst onderschreven categorie. Vrij gebruikelijk werd het vervolgens –de toon is hier gezet in het Monitorrapport (1998)– om een grafiek te vervaardigen van de gemiddelde score[8] voor elke dimensie van de T11. Maar omdat soms ook kansvragen in de T11 worden opgenomen ("hoe groot acht u de kans dat ..") die niet op een vijfpuntsschaal worden gemeten, maar met een vrij in te vullen percentageschaal, moet hier een beslissing worden genomen. Niet altijd is duidelijk hoe die kwestie is opgelost. Elffers, Ruimschotel & Barwegen (1997) geven aparte deelprofielen voor kansvragen en 5-puntsschalen, anderen transformeren de kansvragen, bijvoorbeeld door de 100-puntsschaal in vijf gelijke delen te categoriseren. Van Dijk en Vaneman (2000) projecteren ook de standaarddeviaties in de grafiek. Soms –ook hier heeft het Monitorrapport de toon gezet– wordt in plaats van een lijngrafiek van gemiddelden een gestapeld staafdiagram weergegeven met percentages van de respondenten die bepaalde antwoordopties hebben gekozen. Bronner & Emmering (2000) stellen daarbij voor om eerst enkele categorieën samen te nemen (bij een Likert vijfpuntschaal kiezen zij er meestal, maar niet altijd, voor om de opties 1 en 2 tezamen te nemen, de optie 3 en 'weet niet/ geen antwoord' tezamen, en de optie 4 en 5 tezamen; bij kansschalen nemen zij achtereenvolgens: 'tot 49%', '50%' en 'weet niet/geen antwoord', en 'meer dan 50%'). In de taxi- en warenwetrapporten (Prinsen en Visser, 2000a, 2000b) wordt het profiel in twee stukken opgedeeld, voor de remmende en de instigerende dimensies apart, die vervolgens in een nogal complex samenstel worden afgebeeld. Die rapporten nemen van Elffers & Van Giels (2000a, 2000b) ook een profiel voor gecondenseerde factoren over, met als spreidingsindicatie een kwart standaarddeviatie.

Om de zaak te compliceren worden soms ook zelfgerapporteerde naleving in hetzelfde profiel geprojecteerd, door een schaal van procenten (0-100%) op de kwetsbaarheidsschaal (1-5) af te beelden.

Interpretatie van het kale T11–profiel: het probleem

De volgende vraag is dan hoe moet men nu een T11 - profiel interpreteren? In de literatuur wordt daar nauwelijks expliciet aandacht aan besteed, het wordt kennelijk vanzelfsprekend geacht. Toch zijn hier wel degelijk problemen te onderkennen.

Naast een veelheid van presentatiestijlen komen we ook verschillende standaarden tegen van welke scores men nu als wel of niet kwetsbaar moet beschouwen. Het Monitorrapport(1998) spreekt, uitgaand van scoring van T11 –vragen op vijfpunt Likertschalen met scores 1 (erg kwetsbaar) tot 5 (erg veilig), van kwetsbaar als de gemiddelde score onder de 2.5 zakt. Elffers, Ruimschotel & Barwegen (1997) spreken van zeer kwetsbaar bij gemiddelde scores onder de 2, van enigszins kwetsbaar als de score tussen de 2 en de 2.5 ligt. Elffers & Van Giels (2000a,b) houden ook de 2.5-punt kwetsbaarheidgrens aan, maar stellen daarnaast ook een grens van 3.5 voor, waarboven men de dimensie als een veilig aspect representerend mag beschouwen. Ruimschotel, Van Reenen en Klaasen (2000) noemen de 2.5-grens arbitrair, maar miskennen daarbij mijns inziens de logica van het argument dat, gegeven dat ervaring leert dat de standaarddeviatie van Likertscores veelal in de orde van 1 ligt, een waarde van 2.5 of 3.5 een halve standaarddeviatie verwijderd is van de neutrale score ‘3’, en er dan dus sprake is van een verschil met dit neutrale punt dat men als ‘redelijk groot’ kan duiden in de termen van Cohen (1977).

Het staat iedereen uiteraard vrij om met de T11 om te gaan op een wijze die hij passend vindt, toch zou ik de tweede en derde **aanbeveling** willen herhalen om een handleiding te schrijven waarin een standaardisering van terminologie wordt voorgesteld, zodat een ieder daarnaar kan verwijzen, of er zich desgewenst tegen kan afzetten. Zo’n stuk zou dan niet alleen op de terminologie moeten ingaan, maar ook een kloek voorstel moeten bevatten omtrent wanneer men wel of niet van kwetsbaarheid behoort te spreken.

In deze paragraaf ga ik niet zelf dat kloeke voorstel doen, maar ik geef een aantal overwegingen en observaties die laten zien wat voor soort problemen er eigenlijk aan de orde komen bij interpretatie van een nalevingprofiel.

Op grond van een T11 –profiel hoopt men een indruk te krijgen van de sterke en zwakke punten van de naleving en handhaving van de regels. Als we kijken naar het profiel als geheel, proberen we een geaggregeerd oordeel over een grote groep beregelden te geven. De naïeve interpretatie is ongeveer: punten op het T11 – profiel die in het kwetsbare gebied liggen zijn problematisch, andere niet. Toch is dat een beetje een al te naïef beeld. Bezien wij nogmaals de claim van de T11, die erop neerkomt dat iemand tot overtreding zal overgaan als er *voldoende* instigationele dimensies kwetsbaar genoeg zijn, en *geen* der remmende dimensies zo sterk is dat alsnog de overtredingsgeneigdheid wordt gestopt. Valt dat aan het profiel te onderscheiden? Eigenlijk is er geen statistische methode voorhanden die dit probleem oplost.

Het is daarom zo moeilijk, omdat de onderlinge interactie tussen de factoren van belang kan zijn. Elffers (2000a) liet dat aan het volgende voorbeeld zien. Een regeling waarbij een zeer scherpe controle en zware sancties bestaan (T_7 en T_{11} zijn niet kwetsbaar), zal nauwelijks of geen overtreding laten zien, *onafhankelijk* van andere dimensies, bijvoorbeeld onafhankelijk van de kosten en baten van naleven (T_2). In een dergelijke regeling zal daarom het samenhangpatroon op twee manieren afwijken van de naïeve veronderstelling dat alle dimensies die ‘ertoe doen’ samenhang vertonen met regelovertreding. Allereerst zal vermoedelijk het scorepatroon op de buitengewoon sterke dimensies T_7 en T_{11} nauwelijks spreiding laten zien: alle betrokkenen percipiëren deze dimensies gelijkelijk als zeer sterk. Daarnaast zal het zo zijn dat er geen verband zal kunnen worden aangetoond tussen T_2 en naleving. Het is evenwel onjuist dit in *absolute* zin te interpreteren als: “bij deze regeling is het onbelangrijk hoe hoog kosten en baten zijn”, slechts een conditionele interpretatie is op zijn plaats: “gegeven de krachtige controle en sancties is de hoogte van kosten en baten niet relevant voor naleving”. Bij het huidige handhavingniveau zit er in de gepercipieerde kosten zo weinig variatie, dat er geen ruimte is voor differentiatie van de invloed van kosten op het nalevingniveau.

Een andere manier om ditzelfde probleem te verkennen is na te denken over de absolute en relatieve positie van personen binnen de puntenwolk van alle respondenten. Het kan immers zijn dat een persoon weliswaar een erg lage positie ten opzichte van anderen inneemt in zijn perceptie van een dimensie, zonder dat hij die dimensie zo laag schat dat hij kwetsbaar is. Bijvoorbeeld, als bijna iedereen de sancties zeer zwaar acht, maar deze respondent vindt ze slechts redelijk zwaar, dan is hij relatief gezien extreem negatief, toch hoeft dat niet te betekenen dat hij tot overtreding overgaat, waar het immers mogelijk is dat voor hem de sancties altijd nog zwaar genoeg zijn. In een simpele correlatieanalyse met de afhankelijke 'overtredingrapportage' vermoedt men dat hij stevig bijdraagt aan de correlatie. Evenwel, als het zo is dat een dimensie pas bijdraagt aan de overtredingsgeneigdheid indien personen onder een kwetsbaarheiddrempel zakken, dan zal men in dit geval helemaal geen bijdrage aan de overtredingsgeneigdheid verwachten; de respondent bevindt zich nog duidelijk in het veilige gebied. Men mag bij zo'n stand van zaken ook geen correlatie tussen deze dimensie en de overtredingrapportage verwachten. Een en ander geldt des te sterker als voor een groep dimensies nagenoeg geen variantie mag worden verwacht.

Bovenstaande probeert aannemelijk te maken dat het bestuderen van een kaal T11 –profiel misschien toch nog iets anders bijdraagt, of zou moeten bijdragen, dan dat wat gevonden kan worden door het bestuderen van het samenhangpatroon tussen *T11* –dimensies en overtredingrapportages, dus datgene wat we een correlationele T11 –inventarisatie hebben genoemd.
Men begrijpe mij goed: ik probeer niet te stellen dat correlationele T11 –inventarisaties niet wenselijk zouden zijn, of niet wenselijker dan kale T11 –inventarisaties. In tegendeel, ik zal op hun verdiensten in de volgende paragraaf nog nader ingaan. Maar ik stel wel dat zelfs binnen het onderzoekdesign van een correlationele T11 –inventarisatie toch aandacht moet blijven voor de interpretatie van het T11 –profiel als zodanig. En wie dat met mij vindt, kan zich dan niet onttrekken aan de vraag wat dan precies de bijdrage van die studie van dat profiel is. Het vervelende is evenwel dat daar tot op heden niet zo'n duidelijke methodologie voor is uitgekristalliseerd, en het in de bestaande rapporten dan ook wordt onderbelicht. Eigenlijk wordt zulks vooral aan de lezer overgelaten. Wellicht zou het goed zijn als onderzoekers die de T11 toepassen hun verantwoordelijkheid nemen, en expliciet verwoorden wat zij vinden dat een profiel onthult. Dat is, bij gebreke van een standaardmethode, tot op zekere hoogte een 'art'. Allicht kunnen daarbij deskundigen in het veld van de betrokken regels van dienst zijn.

Net zoals bij de expert –T11–sessies na het vaststellen van het profiel plaats is voor een evaluatiefase, is dat ook in doelgroep-survey het geval, waarbij deze taak meestal toevalt aan de onderzoekers. Het is immers moeilijk zich voor te stellen dat er een soort 'absolute interpretatie' van een T11 –profiel bestaat. Deze evaluatietaak is niet eenvoudig. Als men om de betekenis van het niveau van een score in de T11 te kunnen interpreteren tegelijkertijd het niveau en de spreiding van alle andere scores in het oog moet houden, staan we eigenlijk voor een niet goed doenlijke taak. Om die reden verdient het zeker aanbeveling om, waar men in het kader van een T11 –studie zowel een doelgroepstudie als ook een expertonderzoek is voorzien, de doelgroepstudie eerst te laten plaatsvinden, om vervolgens gebruik te maken van de expertise van de experts door hen in hun evaluatiefase ook te laten spreken over eventuele verschillen die zich tussen het doelgroepprofiel en het expertprofiel vertonen.

Uiteraard is deze hele stand van zaken nogal onbevredigend: we hebben immers geen standaardinstrument met een standaardinterpretatie, maar zien de T11 nu als een hulpmiddel om het onderzoekers / experts op het betreffende regelhandhavingsgebied mogelijk te maken zinvol met elkaar van gedachten te wisselen, en hun impliciete kennis met elkaar te bespreken. Dat is ongetwijfeld heel nuttig, maar laat meteen zien dat de toegevoegde waarde van een survey –in vergelijking met expert-T11 –inventarisaties– betrekkelijk is.

In de volgende paragraaf wordt een aantal in de literatuur aangetroffen hulpmiddelen voor het interpreteren van het T11 –profiel besproken, terwijl in de daaropvolgende het relateren van een profiel aan nalevingschattingen onder ogen wordt gezien.

Interpretatie van het kale T11–profiel: hulpmiddelen

Van Dijk & Vaneman (2000) doen –voor expertsessiegebruik van de T11 – een voorstel om de nalevingprofielgrafiek aan te vullen met *invloedsscores* en *verbeterbaarheidscores*, resulterend in wat zij de hypothetische handhavingkromme noemen. Ofschoon hun behandeling niet geheel duidelijk maakt hoe zij deze begrippen operationaliseren, is hun opzet interessant, omdat zij trachten een methode te ontwikkelen

die de betekenis van een lage of hoge score op een T11 –dimensie voor de naleving doorgrondt. Hun aanpak verdient echter verduidelijking.

Elffers & Van Giels (2000a,b) presenteren een methode om de kale T11 – studies te gebruiken *om 'extreme groepen'* te bepalen. Hun aanpak is om gecondenseerde profielen te maken, en op een reeks relevante achtergrondvariabelen uitbijtende groepen te achterhalen. Zij noemen een groep uitbijtend als er sprake is van een gemiddeldeverschil (op een –gecondenseerde– dimensie) tussen deze groep en de totale steekproef dat èn significant, èn groot is, waarbij zij voor dat laatste een maatstaf in termen van de standaarddeviatie van de scores op die gecondenseerde dimensies voorstellen, in lijn met Cohen (1977). Bij de taxiwet en de warenwet leidde dit tot interessante inzichten. Met deze methode is het mogelijk deelgroepen in de doelgroep van beregelden te identificeren die de betreffende regeling verrassend veel kwetsbaarder achten dan anderen. Ofschoon dat niet impliceert dat zij ook vaker overtreden, kan men deze stand van zaken zien als een aansporing voor de wetshandhavers te bezien of deze groepen speciale aandacht behoeven.

Enigszins in dezelfde lijn stelt Elffers (2000a) voor om *comparatieve profielen* te maken, waarbij deelgroepen qua T11-profiel worden vergeleken. Als voorbeeld kan men dan denken aan: mensen die ooit gecontroleerd zijn versus mensen die dat niet zijn, of bestraften tegen niet-bestraften, nieuwkomers vs. oude rotten onder de regeling. Dit is vergelijkbaar met wat in expertsessies gedaan wordt, wanneer men regelingen of overheden vergelijkt (bijvoorbeeld in Elffers&Klaasen, 1997; Klaasen, 2000), maar is in de survey-context nog nauwelijks van de grond gekomen. Op deze wijze kan men inzicht krijgen in het effect van controle en bestraffing op de houding tegenover de betreffende regels en hun handhaving, in T11 –termen.

Deze reeks observaties overziend kom ik tot de conclusie dat de onduidelijkheid rond interpretatie van T11 –profielen beslist de achilleshiel van de T11 –aanpak is, en er dringend behoefte aan meer eenduidigheid is.. Dit leidt tot twee aanbevelingen:
Twaalfde aanbeveling De interpretatie van nalevingprofielen verdient nadere bestudering, waarbij specifiek dient te worden ingegaan op de betekenis van een profiel, en gewerkt wordt aan analysemethoden die interpretatie kunnen vergemakkelijken. Het lijkt voor de hand te liggen aan eventuele voorstellen tot interpretatie van de T11 een studieconferentie met T11 –experts te wijden.
Dertiende aanbeveling Er is ruimte voor een secundaire analyse van enkele eerdere studies met behulp van de uitbijtende groepenanalyse en de comparatieve profielenanalyse, om te bezien of die aanpak er inderdaad toe leidt dat interpretatie wordt vergemakkelijkt.

Interpretatie van samenhang tussen T11 en nalevingsmarten

Naast de 'kale' T11 –studies, die alleen een nalevingprofiel opleveren, zijn ook correlationele T11 –inventarisaties populair. Elffers(2000a) beveelt zelfs aan nooit kale, maar alleen correlationele inventarisaties uit te voeren. Zonder af te doen aan de opmerkingen over de aanbevelenswaardigheid van een interpretatie van het kale profiel, ook bij een correlationele inventarisatiestudie, zoals die eerder zijn gegeven, neigen de auteurs van dit rapport eveneens naar dit standpunt..

In een correlationele T11 -inventarisatie wordt het profiel aangevuld met één of meer door de respondenten uit het survey gerapporteerde nalevingsmarten, zoals bijvoorbeeld directe zelfrapportage, of rapportage verkregen door middel van de randomised-responseprocedure. Zonder hier nu in te gaan op de problemen die deze nalevingmaten zelf met zich meebrengen (vergelijk het hoofdstuk omtrent de randomised response-benadering van nalevingrapportage), beschouw ik hier de betekenis van het analyseren van de samenhang tussen T11 –scores en nalevingmaten, dat in de besproken literatuur veelal op grond van correlaties of logistische regressietechnieken gebeurt (met, in het geval van randomised-responsematen, speciale aangepaste methoden). In hoeverre is dat een passende methodiek?

Veel van de problemen die in de vorige paragrafen zijn aangestipt, werken onverkort door wanneer het gaat om correlationele analyse van nalevingmaten. Als een T11 –dimensie nagenoeg geen variantie heeft, kan hij geen correlatie vertonen met naleving, als binnen de doelgroep de spreiding op een dimensie zich vrijwel geheel aan een zijde van de kwetsbaarheidgrens bevindt, zal er geen correlatie optreden. Als dimensies sterk samenhangen, is een correlatiecoëfficiënt van één zo'n dimensie met de afhankelijke vaak een weinig inzichtelijke grootheid, en lijkt multivariate analyse aangewezen. Uiteraard worden deze problemen minder pregnant als men met gecondenseerde profielen werkt, maar ze zijn daarmee geenszins volledig verholpen.

Is er een alternatief voorhanden? Ik opper hier de –niet uitgeteste– mogelijkheid om de onafhankelijke variabelen in dit soort analyses anders te behandelen. In de standaardmethoden, of het nu om correlationele of looglineaire regressiemethodiek gaat, worden de onafhankelijke variabelen (de T11 –dimensies) in relatieve of gestandaardiseerde zin gebruikt. Dat wil zeggen dat we de samenhang tussen zo'n dimensie en een nalevingscore bekijken door de afwijkingen ten opzichte van het gemiddelde van een groep op elkaar te betrekken. Men kan overwegen om in plaats daarvan met *absolute scores* te werken. De scores op de T11 –dimensies worden eerst in absolute klassen ingedeeld ('kwetsbaar', 'tussenin', 'veilig', bijvoorbeeld), en de (logit van de) nalevingrapportage als functie van de cellen in zo'n meerdimensionale (3 x 3 x 3 x …) tabel wordt gemodelleerd. Een scheve of sterk geconcentreerde verdeling van een T11 –dimensie leidt dan tot concentratie van het datamateriaal in een beperkt aantal cellen, die mogelijkerwijs aanleiding is om die dimensie niet mee te nemen in de analyse. Dat loglineaire kruistabellenanalyse snel onoverzichtelijk wordt bij veel dimensies, zou een krachtig extra pleidooi opleveren voor een sterke mate van condensatie van de T11 –dimensies, zoals eerder in deze bijdrage al op andere gronden is betoogd. Daarbij kan ook gedacht worden aan technieken om die combinaties van onafhankelijke factoren op te sporen die de doelvariabele, naleving, in verwachting zo hoog mogelijk maken, zoals AID (automatische interactiedetectie).

Ook de methode van de comparatieve profielen, die in de vorige paragraaf werd aangeroerd, dient hier genoemd te worden. We kunnen twee profielen vergelijken voor enerzijds een groep overtreders, anderzijds een groep nalevers. Die indeling kan gebeuren op basis van zelfrapportage, maar, en dat is wellicht vaak een veel interessantere mogelijkheid, het kan ook op basis van andere informatie, bijvoorbeeld vanuit de handhavingdiensten. Vertonen mensen die als overtreders te boek staan een ander profiel dan mensen die als regelnalevers bekend staan? Uiteraard brengt het combineren van surveygegevens met gegevens van handhavingdiensten andere problemen met zich mee, onder andere die van de vraag of men daarover de respondenten moet informeren, en hoe anonimiteit te garanderen. Zulke problemen zijn echter niet onoplosbaar (Elffers, 1991; Van Gils, Van der Heijden, Rosebeek, 2001).

Ook vanuit het oogpunt van de correlationele T11 –inventarisaties blijft bovengeformuleerde twaalfde **aanbeveling**, nadere bestudering van de interpretatie van nalevingprofielen, onverminderd raadzaam, waarbij dan in dit verband ook de meerwaarde of noodzaak van het bekijken van samenhang met nalevingscores onderwerp van studie hoort te zijn.

Bovendien leidt de beschouwing omtrent het inzetten van modellen gebaseerd op absolute scores tot de **veertiende aanbeveling**: bestudeer op grond van bestaand datamateriaal of loglineaire analyse van het verband tussen nalevingprofiel in absolute termen en nalevingrapportage haalbaar en inzichtgevend is, en bezien of bestaande methoden van interactiedetectie daarbij zinvol zijn.

4.4. Samenvatting van de aanbevelingen

In het bovenstaande is de T11–methodologie kritisch besproken, en uit het grote aantal aanbevelingen kan men afleiden dat ik vind dat er alle aanleiding is om te werken aan een verbetering en uitdieping van de methode. Grootste kritiekpunt is de onduidelijkheid die bestaat over hoe men een T11–profiel dient te interpreteren. In die zin zijn aanbevelingen nummers 12 tot en met 14 de meest wezenlijke.

Ik geef hier een samenvattend overzicht van aanbevelingen.

Aanbeveling 1: Aanleggen literatuurbank en databank van bestaande empirische T11 –studies
Aanleiding..... bestaande onderzoek is slecht ontsloten
Soort werk..... ordening bestaand onderzoek
Hoe..... aanleggen zodanige literatuur en databank, bijvoorbeeld bij ERH / Steinmetz-archief
Resultaat..... verbeterde toegankelijkheid

Aanbeveling 2 tot 4: Standaardbeschrijving van meetinstrument en dataverzameling, handleiding analyse en interpretatie, met een beschrijving van specificatie, uitvouwing en condensatie
Aanleiding..... onhelderheid over wat een T11 –studie is

Soort werk..... ordening bestaand werk / voorstellen standaardterminologie; voor de handleiding interpretatie valt te overwegen te wachten tot de aanbeveling 12 tot 14 zijn uitgevoerd
Hoe..... deskresearch
Resultaat..... handleiding

Aanbevelingen 5 tot 7: Nader onderzoek meervoudige operationalisaties
Aanleiding..... twijfel aan betrouwbaarheid van enkelvoudige operationalisaties
Soort werk..... [1] bij nieuwe studies systematisch meervoudige operationalisaties meenemen, en analyseren
..... [2] inventarisatie en analyse bestaande meervoudige operationalisaties
..... [3] MTMM analyse van meervoudige kansvragen
Hoe..... meervoudige operationalisaties opnemen in nieuwe studies
..... secundaire analyse bestaand materiaal
Resultaat..... inzicht en aanbevelingen ten aanzien van operationalisaties

Aanbeveling 8: Onderzoek vraagmethode conditionele gebeurtenissen
Aanleiding..... twijfel aan begrijpelijkheid meervoudig samengestelde vragen met mogelijk hypothetische gebeurtenissen
Soort werk..... separaat empirisch onderzoek
Hoe..... vraagformuleringsstudie in cognitief laboratorium
Resultaat..... inzicht in ernst problematiek / aanbeveling oplossingsrichting

Aanbeveling 9: Revisie T11 tot Twat minder in surveystudies
Aanleiding..... twijfel aan de mogelijkheid, noodzaak en zinvolheid controle en sanctieproces door beregelden zinvol te laten ontleden (vgl. uitkomsten aanbeveling 8-studie)
Soort werk..... ordening bestaande kennis
Hoe..... workshop met T11 -geïnteresseerden
Resultaat..... mogelijkerwijs aanbevelingen om voor surveygebruik met Tminder te werken

Aanbeveling 10: afzien van de vermenigvuldigingsmethode
Aanleiding..... stelselmatige suggestie dat de vermenigvuldigingsmethode (kans op maal ernst van consequentie) de analyse ten goede zou komen
Soort werk..... geen
Hoe..... begraven van die suggestie
Resultaat..... rust aan het front

Aanbeveling 11: Wetsonafhankelijke condensatie
Aanleiding..... wens tot eenduidigheid ‘over wetten heen’ te komen
Soort werk..... ordening bestaande kennis en secundaire analyse
Hoe..... covariantiestructuuranalyse van bestaande data
Resultaat..... inzicht in de mogelijkheid van wetonafhankelijke condensatie

Aanbevelingen 12 en 13: Studie interpretatie en hulpmiddelen bij interpretatie
Aanleiding..... onhelderheid over hoe een (kaal of correlationeel) T11 –profiel te interpreteren
Soort werk..... conceptuele verheldering en nadere analyse van bestaande interpreterings-hulpmiddelen
Hoe..... theoretisch werk; secundaire analyse van het nut van invloedsscores, comparatieve profielen en uitbijtende groepen; workshop met T11 -ers
Resultaat..... verbeterde interpretatiemiddelen

Aanbeveling 14: Ontwikkeling methode analyse verband T11 in absolute termen – naleving
Aanleiding..... zwakte interpretatie verband T11 - naleving
Soort werk..... methodeontwikkeling
Hoe..... theoretisch werk; heranalyse bestaand datamateriaal, met name bij randomised response methoden
Resultaat..... betere interpretatie verband T11 -naleving

Literatuur

Aquilino, W.S. (1994). Interview mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly, 58*, 210-240.

Aquilino, W.S. (1993). Effects of spouse presence during the interview on survey responses. *Public Opinion Quarterly, 57*, 358-376.

Aquilino, W. (1997). Privacy eefects on self reported drug use: Interactions with survey respondent characteristics. In Harrison & Hugehes (eds) The validity of self-reported drug use: Improving the accuracy of survey estimates. NIDA Reseach monograph series, 167, 383-415

Babinski, L. M., Hartsough, C. S., & Lambert, N. M. (2001). A comparison of self-report of criminal involvement and official arrest records. *Aggressive Behavior, 27*, 44-45.

Beckenbach, A. (1995). Computer assisted questioning: the new survey methods in the perception of the respondent. *BMS, 48*, 82-100.

Bethlehem, J. et al. (1996). *Bascula Reference Manual*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

Belson, W.A. (1981) *The design and understanding of survey-questions.* Aldershot: Gower Publishing Corporation.

Billiet, J. en Loosveldt G. (1988). Improvement of the quality of the responses to factual survey questions by interviewer training. *Public Opinion Quarterly*, 1988,52, 190-21.

Bouts, Jan, Ger van Gils en Peter van der Heijden (1998), Randomized response, NIPO, Amsterdam.

Bradburn, N.M. (1983). Response effects. In: P.H. Rossi, J.D. Wright en A.B. Anderson (Eds), *Handbook of survey research*. New York: Academic Press.

Bradburn, N.M. en Sudman, S. (1979). I*mproving interview method and questionnaire design.* San Francisco: Jossey Bass.

Bradburn, N.M., Sudman S., Blair, E., en Stocking C. (1989). Question threat and response bias. In: E. Singer en S. Presser (Eds) *Survey research methods; A reader*. Chicago: University of Chicago Press. (Het betreffende artikel verscheen in eerste instantie in Public Opinion Quarterly, 1978, 42, 2).

Bronner, F.E. & A.J. Emmering (2000). De Tafel van Elf als survey-instrument: van theorie naar praktijk. *In: P. van Reenen (ed). De Tafel van Elf*. p. 65-78. Den Haag: Sdu Uitgevers.

Burt, M. (1995). Handling sensitive information on US homeles surveys. Paper gepresenteerd op de European Conference on Demography, Statistics and Privacy, Parijs, october 1995.

Carroll, J. & F. Weaver (1986), Shoplifters' Perceptions of Crime Opportunities: A Process-Tracing Study. In: D.B.Cornish & R.V Clarke (eds.) *The Reasoning Criminal. Rational Choice Perspectives on Offending. New York etc: Springer-Verlag.* p. 19-38.

Clark, Stephen J. , and Robert A. Desharnais. 1998. "Honest answers to embarrassing questions: detecting cheating in the randomized response model." *Psychological Methods* 3(2):160-168.

CBS. (1993). *De leefsituatie van de Nederlandse bevolking, 1992. Kerncijfers*. 's-Gravenhage: SDU/uitgeverij.

Cohen, J. (1977) *Power Analysis for the Behavioral Sciences, Second Edition*. New York etc.: Academic Press.

Coleman, C., & Moynihan. (1996). *Understanding crime data. Haunted by the dark figure*. Buckingham, UK: Open University Press.

Commissie Kordes (1997). *Privacy-wetgeving en het gebruik van persoonsgegevens voor wetenschappelijk en statistische doeleinden.* Advies aan de minister van Onderwijs, Cultuur en Wetenschappen, Den Haag/Zoetermeer, 1997, hoofdstuk 6.

Cooper, Harris, and Larry V. Hedges (eds.). 1994. *The handbook of research synthesis*. Russell Sage Foundation.

Cornish, D.B.& R.V Clarke (eds.), (1986), *The Reasoning Criminal. Rational Choice Perspectives on Offending*. New York etc: Springer-Verlag.

Dijk, B. van, & J.M. Vaneman (2000). De Tafel van Elf bij de Rijksdienst voor Radiocommunicatie. *In: P. van Reenen (ed). De Tafel van Elf*. p. 87-97. Den Haag: Sdu Uitgevers.

Dijk, Th. van, H. Elffers, D.J. Hessing & A.B. Hoogenboom (1999) *Bewust van de gevaren van criminaliteit. Een inventarisatie van kwetsbaarheden die in de logistieke keten en daarmee ook in de Rotterdamse haven voorkomen.* SI-EUR-reeks nr. 19. Deventer: Gouda Quint..

Dillman, DA (1978). Mail and telephone surveys. New York: Wiley.

Dillman, DA (2000). Mail and internet surveys. New York: Wiley.

Elffers, H. (1991), *Income tax evasion. Theory and measurement*. Reeks Burger en Belastingen nr. 9, Deventer: Kluwer.

Elffers, H. (2000a). Over het nut van survey-onderzoek met de Tafel van Elf. In*: P. van Reenen (ed). De Tafel van Elf*. p. 79-85. Den Haag: Sdu Uitgevers.

Elffers, H. (2000b), *Aanvullende Secundaire Analyse Naleving Taxiwet*. Rapport aan de ERH. Rotterdam: Sanders Instituut.

Elffers, H., P.G.M. van der Heijden and M. Hezewijk. A survey study of rule transgression for two Dutch instrumentel laws, applying the randomized response method. *Aangeboden voor publicatie.*

Elffers, H., P. van der Heijden & M. Hezemans, (aangeboden voor publicatie). Explaining regulatory non-compliance. A survey study of rule transgression for two Dutch instrumental laws, applying the randomised response method. *Aangeboden voor publikatie.*

Elffers, H. & B. Van Giels (2000a) *Secundaire Analyse Naleving Taxiwet*. Rapport aan de ERH. Rotterdam: Sanders Instituut.

Elffers, H. & B. Van Giels (2000b) *Secundaire Analyse Naleving Warenwet*. Rapport aan de ERH. Rotterdam: Sanders Instituut.

Elffers, H. & H.M. Klaasen (1997). *Ex ante evaluatie van de handhaafbaarheid van het Bouwstoffenbesluit bodem- en oppervlaktewaterbescherming aan de hand van de Tafel van Elf*. Den Haag/Rotterdam: IRH/Sanders Instituut.

Elffers, H. & D. Ruimschotel (1997), *The Table of Eleven (T^{11}) as a new content oriented paradigm for evaluation research*. Paper presented at the 2nd European Evaluation Society Conference, Stockholm.

Elffers, H., D. Ruimschotel & F. Barwegen (1997) *Naleving van de Wet MOT door cliëntadviseurs van de banken*. Rotterdam: Sanders Instituut

Elliott, D. S., & Ageton, S. S. (1980). Reconciling race and class differences in self-reported and official estimates of delinquency. *American Sociological Review, 45*, 95-110.

Farrington, D. P., Loeber, R., Stouthammer-Loeber, M., Van Kammen, W. B., & Schmidt, L. (1996). Self-reported delinquency and a combined delinquency seriousness scale based on boys, mothers, and teachers: concurrent and predictive validity for african-americans and caucasians. *Criminology, 34*(4), 493-518.

Fowler, F.J. jr. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage.

Forsyth, B.H. en Lessler, J.T. (1991). Cognitive laboratory methods: A taxonomy. In: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz en S. Sudman (Eds), *Measurement errors in surveys*. New York: Wiley.

Fox, James A., and Paul E. Tracy. 1980a. "A field validation of a quantitative randomized response model." *Proceedings of the survey research methods section, ASA*:299-304.

—. 1980b. "The randomized response approach: applicability to criminal justice research and evaluation." *Evaluation Review* 4(5):601-622.

Gils, Gèr van, Peter van der Heijden, Annemarie Rosebeek (2001), *Randomized response: onderzoek naar regelovertreding, resultaten ABW, WAO en WW*. Den Haag, Ministerie van Sociale Zaken en Werkgelegenheid.

Gils, Ger van Peter van der Heijden en Hans Landsheer (1996), Rapportage van uitkeringsfraude in surveys, Den Haag, Ministerie van Sociale Zaken en Werkgelegenheid.

Greenberg, Bernard G., Roy R. Kuebler, James R. Abernathy, and Daniel G. Horvitz. 1971. "Application of the randomized response technique in obtaining quantitative data." *Journal of the American Statistical Association* 66(334):243-248.

Gibson, D. (1994). Drug injectors may underreport drug risk but not sexual risk behavior. Abstract book tenth international conference on Aids. Yokohama, Japan, 7-12 Aug. 1994.

Groves, R.M. (1989). *Survey errors and survey costs.* New York: Wiley.

Hagan, F. (1997). *Research methods in criminal justice and criminology* (4 ed.). Needham Heights, Massachusets: Viacom.

Himmelfarb, Samuel, and Carl Lickteig. 1982. "Social desirability and the randomized response technique." *Journal of Personality and Social Psychology* 43(4):710-717.

Hippler, H-J, Schwarz, N en Singer, E (1990). Der Einflues von Datenschutzzusagen auf die Teilnamebereitschaft an Umfragen. *ZUMA-Nachrichten, 27*, 54-67.

Heijden, P.G.M. van der, G. van Gils, J. Bouts and J. Hox (2000). A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods and Research, 28*, 505-537.

Heijden, P.G.M. van der, Gils, G. van, Bouts, J., & Hox, J. (1998). A comparison of randomized response, CASAQ, and direct questioning; Eliciting sensitive information in the context of social security fraud. Kwantitatieve Methoden, 59, 15-34.

Heijden, P.G.M. van der, en Gils, G. van (1996). some logistic regression models for randomized response data. In: A. Forcina, G.M. Marchetti, R. Hatzinger, and G. Galmatti (eds). Statistical modeling, Proceedings of the 11th international workshop on statistical modeling, Orvieto, Italy.

Heijden, P., van der, Sijtsma, K., & 't Hart, H. (1995). Self-report delinquentieschalen zijn nog steeds betrouwbaar. Een reactie op de studies van Bruinsma (met repliek en dupliek). *Tijdschrift voor Criminologie, 37*, 71-82.

Hessing, D. J., & Elffers, H. (1995). De validiteit van de self-report methode in onderzoek naar regelovertredend gedrag. *Tijdschrift voor Criminologie*, 54-70.

Hochstim, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association, 62*, 976-989.

Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. Beverly Hills, California: Sage.

Hox, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research, 22*, 3, 300-318.

Inspectie voor de Rechthandhaving (1994). *De eerste stap. Een kosten-effecten analyse van een integrale periodieke rapportage van de mate van naleving van de beleidsinstrumentele wetgeving*. Den Haag: Ministerie van Justitie

Inspectie voor de Rechthandhaving (1997). *De tafel van elf brochure*. Den Haag: IRH.

Inspectie voor de Rechtshandhaving (1998). *Monitoring van beleidsinstrumentele wetgeving. Ontwerp en experiment*. Den Haag: Ministerie van Justitie.

Intomart/ Veldkamp. 2000. "Naleving en handhaving van de beleidsinstrumentele wetgeving: *warenwet*." Veldkamp /Intomart.Junger, M. (1989). Discrepancies between police and self-report data for Dutch racial minorities. *British Journal of Criminology, 29*(3), 273-283.

Junger, M. (1990). *Delinquency and Ethnicity. An investigation on social factors relating to delinquency among Morocan, Turkish, Surinamese and Dutch boys*. Boston,: Kluwer.

Junger, M., & Polder, W. (1991). *Delinquent gedrag bij jongens uit etnische minderheden: achtergronden, II.* (107). Arnhem, Gouda Quint,: Wetenschappelijk Onderzoek en Documentatiecentrum (WODC).

Junger-Tas, J., & Haen-Marshall, I. (1999). The self report methodology in crime research: strength and weaknesses. In M. Tonry & D. P. Farrington (Eds.), *Building a safer society: Strategic approaches to crime prevention* (Vol. 25, pp. 291-367). Chicago Ill: The University of Chicago Press.

Junger-Tas, J., Junger, M., Barendse-Hoornweg, E., & Sampiemon, M. (1983). *Jeugddelinquentie. Achtergronden en justitiele reactie* (Vol. 42). Den Haag: Staatsuitgeverij.

Junger-Tas, J., Kruissink, M., & van der Laan, P. H. (1992). *Ontwikkeling van de jeugcriminaliteit en de justitiële jeugdbescherming: periode 1980-1990* (Vol. 119). Arnhem, NL: Gouda Quint.Lee, R.M. (1993). *Doing research on sensitive topics*. London: Sage.

Kahneman, D. & A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.

Kendall, M. K., and A. Stuart. 1979. *The Advanced Theory of Statistics*. Macmillan.Leeuw, E.D. de (1999). Verzamelen van gevoelige informatie: Een overzicht *Pedagogische Studiën, 76*, 2, 92-103.)

Klaasen, H.M. (2000). “Quick scan” van de handhaafbaarheid: de T^{11} –analyse van het Bouwstoffenbesluit. In: *P. van Reenen (ed). De Tafel van Elf*. p. 99-111. Den Haag: Sdu Uitgevers.

Land, Hedda van 't. 2000. "Similar questions; different meanings." Pp. 144 in *Department of Social Sciences*. Free University Amsterdam.Leeuw, E.D. de (1992). *Data quality in mail, telephone, and face to face surveys*. Amsterdam: TT-publikaties.

Landsheer, Johannes A, Peter van der Heijden, and Ger van Gils. 1999. "Trust and understanding, two psychological aspects of randomized response." *Quality and Quantity* 33(1):1-12.Leeuw, E.D. de, Hox, J.J. en Snijkers, G.J.M. (1995). The effect of computer assisted interviewing on data quality; A review. *Journal of the Market Research Society, 37*, 325-344.

Leeuw, Edith de, Joop Hox, Sabina Kef, and Marion van Hattum. submitted. "Computer assisted selfinterviewing tailored for special populations; Overcoming the problems of special interviews and sensitive topics." *Fiel Methods*.

Leeuw, E. D. de (1999). Vragen naar gevoelige informatie: een overzicht. *Pedagogische Studiën, 76*, 92-103.

Leeuw, E.D. de (2001). The effect of computer assisted interviewing on data quality: A review of the evidence [CD-Rom]. In: J. Blasius, J. Hox, E. de Leeuw & P. Schmidt (Eds.), *Social science methodology in the new millennium*. Opladen, FRG: Leske + Budrich

Leeuw, E.D. de, Hox, J.J., Kef, S, & Van Hattum, M. (2001). Overcoming the problems of special interviews on sensitive topics: Computerassisted self-interviewing tailored for special populations. [CD-Rom]. In: J. Blasius, J. Hox, E. de Leeuw & P. Schmidt (Eds.), *Social science methodology in the new millennium*. Opladen, FRG: Leske + Budrich

Lensvelt-Mulders, Gerty, Joop Hox, and Peter van der Heijden. submitted. "Meta-analysis of randomized response research: 35 year of validation."

Lessler, J. & O’Reilly, J. (1995). Literacy limitations and solutions for self-administered questionnaires. *Seminar on New Directions in Statistcal Methodology*. Wsahington, DC: US Office of Management and Budget (statistical policy working paper 23)

Lui, P. T., and L. P. Chow. 1976. "A new discrete quantitative randomized response model." *Journal of the American Statistical Association* 71(353):72-73.

Lyberg, L.E. en Dean, P. (1992). Methods for reducing nonresponse rates: A review. Paper gepresenteerd op de jaarlijkse conferentie van AAPOR, St Petersburg, Florida, mei, 1992.

Makkai, T., en McAllister, I. (1992). Measuring social indicators in opinion research; A method to improve accuracy on sensitive questions. *Social Indicators Research, 27*, 2, 169-186.

Malow, R.M., Gustman, S.L., Ziskind, D., McMahon, R., St.Lawrence, J. (1998) Evaluating HIV prevention interventions among drug abusers. Journal of HIV/AIDS Prevention and Education for Adolescents and Children, 2, 21-40

Mangat, N. S., and Ravindra Singh. 1990. "An alternative randomized response procedure." *Biometrica* 77(2):439-442.

Maxfield, M. G., Weiler, B. L., & Widom, C. S. (2000). Comparing self-reports and official records of arrests. *Journal of Quantitative Criminology, 17*, 87-110.

Meeus, W., & 't Hart, H. (1993). *Jongeren in Nederland. Een nationaal survey naar ontwikkeling in de adolescentie en naar intergenerationele overdracht.* Amersfoort: Academische Uitgeverij.

Moerland, H. (1991), *Winkeldiefstal, een te riskante zaak?* Arnhem: Gouda Quint.

Moors, J. J. A. 1971. "Optimization of the unrelated question randomized response model." *Journal of the American Statistical Association* 66(335):627-629.

Nicholls, W.L.II, Baker, R.P., & Martin, J. (1997). The effect of new data collection technologies on survey data quality. In: L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (eds). *Survey Measurement and Proces Quality*. New York: Wiley.

Nijboer, J. (1995). Het meten van delinquentie door middel van self-report. *Tijdschrift voor Criminologie, 37*, 273-280.

OECD (1999). *The state of regulatory compliance: issues, trends and challenges*. Paris: OECD/ODCE Publication PUMA/REG(99)3.

O'Reilly, J.M., Hubbard, M.L., Lessler, J.T., Biemer, P.P., & Turner, C.F. (1994). Audio and video computer assisted self-interviewing: Preliminary tests of new technologies for data collection. *Journal of Official Statistics, 10*, 197-214.

Petersilia, J. (1978). Validity of Criminal Data Derived From Personal Interview. In C. Wellford (Ed.), *Quantitative Studies in Criminology* (pp. 30-47). Beverly Hills, CA: Sage.

Pollock, K. H., and Yuksel Bek. 1976. "A comparison of three randomized response models for quantitative data." *Journal of the American Statistical Association* 71(356):884-886.

Prinsen, H.M. & R.A.Visser (2000a) *Monitoring van beleidsinstrumentele wetgeving. Eindrapport en advies verbetering naleving taxiregels.* Den Haag: Expertisecentrum Rechtshandhaving.

Prinsen, H.M. & R.A.Visser (2000b) *Monitoring van beleidsinstrumentele wetgeving. Eindrapport en advies verbetering naleving warenwetregels.* Den Haag: Expertisecentrum Rechtshandhaving.

Reenen, P. van (red.). (2000). *De Tafel van Elf.* Den Haag: Sdu Uitgevers.

Ruimschotel, D. (2000). Naleving van wetgeving: waarom zou men het willen weten? Schattingsmethoden in theorie en praktijk. In*: P. van Reenen (ed). De Tafel van Elf.* p. 29-50. Den Haag: Sdu Uitgevers.

Ruimschotel, D. & C. Damaisjah, (1998). Handhaving en fraude bij EU-subsidies. Feitelijke situatie, problemen en suggesties voor verbetering. *In: Commissie Bestuursrechtelijke en Privaatrechtelijke Handhaving. Handhaven op niveau.* p. 217-246.

Ruimschotel, D., H. Elffers & M. Hezemans (1997). *Evaluatie van nalevingsmaten en T11-dimensies, nadere analyse van data verzameld in de drie monitoringsstudies naar de bestrijdingsmiddelenwet, de wet individuele huursubsidie, en de precursorenwet*. Rotterdam: Sanders Instituut.

Ruimschotel, D., P. van Reenen & H.M. Klaasen (1995). Tussen norm en feit. *Beleidsanalyse* 2, p.15-23.

Ruimschotel, D., P. van Reenen & H.M. Klaasen (1996). De Tafel van Elf: een conceptueel kader en een instrument bij rechtshandhavingsvraagstukken. *Beleidsanalyse* 3, p.4-13.

Ruimschotel, D., P. van Reenen & H.M. Klaasen (2000). De Tafel van Elf: een conceptueel kader en een instrument bij rechtshandhavingsvraagstukken. <gereviseerde overdruk van: Ruimschotel, Van Reenen & Klaasen 1996>. In*: P. van Reenen (ed). De Tafel van Elf*. p. 11-28. Den Haag: Sdu Uitgevers.Scott, J. (1997). Children as respondents: Methods for improving data quality. In: L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, D. Trewin, C. Dippo, en N. Schwarz (Eds). *Survey measurement and process quality.* New York: Wiley.

Scheers, N.J. en Dayton, C.M. (1988). Covariate randomized response models. *Journal of the American Statistical Association, 83*, 969-974.

Scherpenzeel, Annette. 1995. "A question of quality: Evaluating survey questions by multitrait-multimethod studies." Pp. 193 in *Department of Social Sciences*. Free University Amsterdam.

Schwarz, N., Strack, F., Hippler, H-J, & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*, 193-212.

Short, J. F., & Nye, F. I. (1957). Reported Behavior as a Criterion of Deviant Behavior. *Social problems, 5*, 207-213.

Singh, S. , R. Singh, and N. S. Mangat. 2000. "Some alternative strategies to Moors' model in randomized response sampling." *Journal of Statistical Planning and Inference* 83:243-255.

Singh, S., N. S. Mangat, and R. Singh. 1994. "On the estimation of mean /total of stigmatized quantitative variables." *Statistica* 54:383-386.

Singh, S., R. Singh, and N. S. Mangat. 1995. "Mail survey design for sensitive quantitative variables." *Metron* 53:43-54.

Snijkers, G. (2002), *Cognitive Laboratory Experiences*. Heerlen, NL: CBS.

Soeken, K. L., and G. B. Macready. 1982. "Respondents' perceived protection when using randomized response." *Psychological Bulletin* 92:487-498.

Stem, D. E., and R. K. Steinhorst. 1984. "Telephone interviewing and mail questionnaire applications of the randomized response model." *Journal of the American Statistical Association* 79:555-564.

Sudman, S. en Bradburn, N.M (1982). *Asking questions. A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Sudman, S. en Bradburn, N.M (1974). *Response effects in surveys: A reveiw and synthesis*. Chicago: Aldine.

Siemiatycki, J. (1979). A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health, 69*, 238-245.

Teilmann, K. S. (1976). *Sources of Bias in Self-Reported Delinquency.* University of Southern California, Los Angeles, CA.

Thornberry, J., Bhaskar, B., Krulewitch, C., Wesley, B, Hubbard, M., Adamson, M. (2000). Audio, computerassisted self imterview (A-CASI) with touch screen to detect alcohol consumption in pregnant women: Application of a new technology to an old problem. Paper presented at the Fifth international conference on logic and methodology, Cologne, October 2000.

Thornberry, T. P., & Krohn, M. D. (2000). The Self-Report Method for Measuring Delinquency and Crime. In D. Duffee (Ed.), *Measurement and Analysis of Crime and Justice* (Vol. 4, pp. 33-84). Washington, DC: National Institute of Justice.

Tourangeau, R. & Smith, T.W. (1996). Asking sensitive questions; The impact of data collection, question format, and questionnaire context. *Public Opinion Quarterly, 60*, 275-304

Tracy, Paul E., and James A. Fox. 1981. "The validity of randomized response for sensitive measurements." *American Sociological Review* 46(april):187-200.

Turner, C.F., Ku, L., Rogers, S.M., Lindberg, L.D., Pleck, J.H., Sonenstein, F.L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science, 280*, 867-873.

Turner, C.F., Lessler, J.T., & Devore, J.W. (1992). Effects of mode of administration and wording on reporting of drug use. In: C.F. Turner et al (eds) *Survey measurement of drug use: Methodological studies*. Washington DC: Government printing office.

US Federal Committee on Statistical Methodology (2001). *Measuring and reporting sources of error in surveys*, statisticalpolicy working paper 31. Washington, DC: Executive Office of the President of the United States: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget .

Umesh, U.N. en Peterson, R.A. (1991). A critical evaluation of the randomized response method. Applications, validation, and research agenda. *Sociological Methods and Research, 20*, 1, 104-138.

Veldkamp/Intomart (1997a). Naleving en handhaving van beleidsinstrumentele wetgeving: bestrijdingsmiddelenwet. Project 72808. Amsterdam/Hilversum: Veldkamp / Intomart.

Veldkamp/Intomart (1997b). Naleving en handhaving van beleidsinstrumentele wetgeving: wet individuele huursubsidie. Project 72809. Amsterdam/Hilversum: Veldkamp / Intomart.

Veldkamp/Intomart (1997c). Naleving en handhaving van beleidsinstrumentele wetgeving: precursorenwet. Project 72810. Amsterdam/Hilversum: Veldkamp / Intomart.

Veldkamp/Intomart (2000a). Naleving en handhaving van beleidsinstrumentele wetgeving: warenwet. Project 2602/DV. Amsterdam/Hilversum: Veldkamp / Intomart

Veldkamp/Intomart (2000b). Naleving en handhaving van beleidsinstrumentele wetgeving: taxiwet. Project 2533/DV. Amsterdam/Hilversum: Veldkamp / Intomart

Vereniging van marktonderzoeksbureaus (VMO) Bureau-informatie (1997), *Gedragscode m.b.t marktonderzoek, persoonsgegevens en privacybescherming*.

Vermunt, J. 1993 *LEM: Log-linear and event history analysis with missing data using the EM algorithm.* Tilburg: WORC PAPER 93.09.015/7.

Webb, E.J., Campbell, D.T., Schwarz, R.D. en Sechrest, L. (1966). *Unobtrusive measures: nonreactive research in the social sciences*. Chicago: Rand McNally.

Weisband, S, & Kiesler, S (1996). Self-disclosure on computer forms: Meta analysis and implications. CHI '96 (http://www.al.arizona.edu/~weisband/chi/chi96.html).

Wentland, E. J., & Smith, K. W. (1993). *Survey responses. An evaluation of their validity*. San Diego, CA: Academic press.

Witt, K. J., and S. Bernstein. 1992. "Best practices in Disk-by-mail surveys." in *1989 Sawtooth Conference*. Sawtooth Software.

Wittebrood, K., Michon, J. A., & Ter Voert, M. J. (Eds.). (1997). *Nederlanders over criminaliteit en rechtshandhaving* (Vol. 1). Deventer, The Netherlands: Gouda Quint.

---

[1] De Tafel van Elf is ontworpen door mr.dr. D. Ruimschotel in opdracht van en in samenwerking met het Ministerie van Justitie.

[2] De term ‘p-vraag’ is niet direct van toepassing op de methoden van Warner en Kuk, waar in elke conditie een vraag wordt beantwoord. In die gevallen staat ‘p-vraag’voor de kans op het beantwoorden van de eerste van de twee vragen. Voor de bepaling van de steekproefvariantie is dit onderscheid een detail (Lensvelt-Mulders, Hox & van der Heijden, 2002), en in dit stuk houden wij daarom de term p-vraag aan.

[3] Greenberg et al. 1971

[4] Lui & Chow 1976

[5] Pollock & Bek 1976

[6] Deze tabel is als voorbeeld overgenomen uit: Van Gils, Van der Heijden en Rosebeek, 2001.

[7] De Tafel van Elf is ontworpen door mr.dr. D. Ruimschotel in opdracht van en in samenwerking met het Ministerie van Justitie.

[8] Uiteraard worden vragen zonodig eerst omgepoold, als ze in een veilig-kwetsbaar-richting waren geformuleerd.