Tijdschrift voor Criminologie 1, jrg. 37 1995

Self-report delinquentie-schalen zijn nog steeds betrouwbaar

Een reactie op de studies van Bruinsma[1]

Peter G.M. van der Heijden, Klaas Sijtsma en Harm 't Hart

In twee studies heeft Bruinsma (1991, 1994) op basis van empirisch materiaal uitspraken gedaan over de test-hertest betrouwbaarheid van de *self-report* methode. In beide studies liet hij een groep mensen een *self-report* vragenlijst invullen. Na een korte periode benaderde hij dezelfde mensen met het verzoek dezelfde vragen opnieuw te beantwoorden, 'omdat, zo werd vooraf meegedeeld, er iets mis was gegaan met de vorige' (Bruinsma, 1991: 247). In de eerste studie (1991) was de tussenperiode twee weken, en bestond de onderzochte groep uit 160 eerstejaars studenten bestuurskunde. In de tweede studie was de tussenperiode een maand, en ging het om een aselecte steekproef van 87 jongeren uit Enschede en Hengelo, die ondervraagd werden in het kader van het *International Self-Reported Delinquency-project*. Doordat Bruinsma van alle respondenten twee afnames had, was hij in staat verschillen te constateren tussen de antwoorden op afname 1 en die op afname 2.

Bruinsma analyseert in beide studies de verschillen op itemniveau (afzonderlijke vragen naar delinquentie) en op schaalniveau (somscores). Op itemniveau gebruikt hij het criterium dat er sprake is van onbetrouwbaarheid indien een respondent niet precies hetzelfde op afname 1 als op afname 2 heeft geantwoord. In onderzoek worden meestal niet de individuele items gebruikt maar de schalen die gebaseerd zijn op de individuele items. Wij zullen ons hier daarom concentreren op de geconstrueerde schalen, en op de bruikbaarheid ervan in sociaal-wetenschappelijk onderzoek.

Bruinsma's constructie van delinquentieschalen

In de studie uit 1991 zijn twee delinquentieschalen geconstrueerd, die ook in de studie van 1994 gebruikt worden. De eerste schaal is een frequentieschaal, waarbij elk item antwoordscores 0, 1, 2, 3 heeft (= aantal keer afgelopen jaar gepleegd, drie betekent drie of meermalen gepleegd), en waarbij een som berekend is over 29 items. De tweede schaal is een variatieschaal, waarbij elk item antwoordscores 0 en 1 heeft (= afgelopen jaar wel of niet gepleegd), en waarbij de som over 29 items is genomen.

Bruinsma definieert ook op schaalniveau betrouwbaarheid op een strenge manier: de schaalwaarde dient voor de eerste afname precies gelijk te zijn aan die voor de tweede afname. In de studie uit 1991 wordt dit voor de frequentieschaal slechts in 5.3% van alle respondenten[2] gehaald, en voor de variatieschaal slechts in 8.7%. Voor de frequentieschaal levert een correlatie

1. De auteurs bedanken Marianne Junger, Hans Landsheer en Gert-Jan Veerman voor hun commentaar op een eerdere versie.
2. Deze analyses zijn gebaseerd op 150 van de 160 personen. Bruinsma (1991) vermeldt niet waarom de gegevens van 10 personen niet zijn gebruikt.

tussen de schaalwaarden op de twee afnames een waarde van -.14 op, en voor de variatieschaal een waarde van -.13.
In 1994 wordt naast de twee schalen uit 1991 ook nog een variatie-ooit-schaal berekend waarbij elk item antwoordscores 0, 1 (= nooit, ooit gedaan) heeft. Voor de frequentieschaal heeft 23.7 % van de respondenten[3] hetzelfde antwoord, voor de variatieschaal is dit percentage 27.1 % en voor de variatie-ooit-schaal is het 40.7 %. Alhoewel de verschillen minder groot zijn dan in 1991, lijken zij nog steeds aanzienlijk.
Dit brengt Bruinsma (1994) tot de conclusie dat het met de test-hertest betrouwbaarheid van de methode van zelfrapportage niet goed gesteld is. Hij beveelt aan uiterst voorzichtig te zijn met de uitkomsten van onderzoeken naar jeugdcriminaliteit die met deze methode verkregen zijn. Wij zullen hieronder aantonen dat zijn studies weinig reden geven om aan de betrouwbaarheid van *self-report* studies te twijfelen.
Naast de hierboven genoemde gegevens waarop Bruinsma zijn pessimistische conclusies baseert, rapporteert hij voor de frequentieschalen ook Cronbach's alfa's. De waarden van alfa zijn, respectievelijk, .75 (eerste afname in studie 1991), .76 (tweede afname in studie 1991), .63 (eerste afname in studie 1994) en .75 (tweede afname in studie 1994). Bruinsma lijkt voor de studie uit 1994 vergeten te zijn de correlaties tussen de somscores van de drie schalen te rapporteren. Wij hebben deze zelf berekend[4]. Voor de drie schalen zijn de correlaties tussen de eerste en tweede afname .82, .84 en .83. De gegevens zijn samengevat in tabel 1.

Tabel 1. De door ons besproken gegevens uit Bruinsma (1991, 1994)

	Bruinsma (1991)	Bruinsma (1994)
Frequentieschaal		
% gelijk afname 1 en 2	5.3	23.7
correlatie tussen		
afname 1 en 2	-.14	.82
alfa afname 1	.75	.63
alfa afname 2	.76	.75
Variatieschaal		
% gelijk afname 1 en 2	8.7	27.1
correlatie tussen		
afname 1 en 2	-.13	.84
Variatie-ooit-schaal		
% gelijk afname 1 en 2	n.v.t.	40.7
correlatie tussen		
afname 1 en 2	n.v.t.	.83

Wij zullen hieronder allereerst aangeven dat Bruinsma (1991, 1994) zich

3. Deze analyses zijn gebaseerd op 59 van de 87 personen. Bruinsma (1994) vermeldt dat de vragenlijsten van de andere personen niet bruikbaar waren, maar geeft de reden hiervan niet aan.
4. Het bleek dat in de gerapporteerde tabellen 4, 5 en 6 (TvC, 1994: 229-132) fouten zaten. Bruinsma heeft ons de gecorrigeerde gegevens geleverd.

voor de bepaling van betrouwbaarheid op de verkeerde maten baseert, en vervolgens beargumenteren dat er geen sprake is van een bepaling van de test-hertest betrouwbaarheid, maar dat het in deze situatie juister is conclusies inzake de betrouwbaarheid te baseren op Cronbach's alfa.

Kritiek op de berekening van de betrouwbaarheid

Bij gegevens van een nominaal meetniveau is het gebruikelijk om een betrouwbaarheidscoëfficiënt te baseren op het aantal keren dat respondenten niet precies hetzelfde hebben geantwoord. Voor gegevens op een ordinaal meetniveau, of gegevens op een interval meetniveau, dient men andere maten te kiezen, omdat bijvoorbeeld een verschil in schaalwaarde van 1 tussen de twee afnames (bijvoorbeeld: in eerste afname worden 2 delicten gemeld en in tweede afname 3) minder erg is dan een verschil van bijvoorbeeld 15 (bijvoorbeeld in eerste afname worden 2 delicten gemeld en in tweede afname 17). Voor delinquentieschalen zoals geconstrueerd door Bruinsma is het gebruikelijk een interval meetniveau aan te nemen. In test-hertest betrouwbaarheidsonderzoek neemt men dan de correlatie als 'coefficient of stability'[5] (zie b.v. Crocker and Algina, 1986: 133-134).
Bruinsma baseert zijn conclusies op het aantal personen dat precies dezelfde somscores heeft op de twee afnames. Dit is niet gangbaar. Het heeft ertoe geleid dat hij de verkeerde conclusies heeft getrokken. Wij baseren ons hieronder op de correlaties tussen de schalen op de eerste en de tweede afname, en op de door Bruinsma gerapporteerde Cronbach's alfa's. Maar eerst is het nodig iets te zeggen over de bepaling van de betrouwbaarheid van een schaal.[6]

Een theoretische beschouwing over betrouwbaarheid

Bruinsma's claim dat hij met zijn studies de test-hertest betrouwbaarheid heeft gemeten, is niet gerechtvaardigd. Er is slechts sprake van test-hertest betrouwbaarheid indien de test twee maal *onder identieke omstandigheden* is afgenomen. Men bestudeert dan de invloed van toevallige meetfouten, zoals verkeerd coderen, gokken en per ongeluk verkeerd invullen. Ook is het mogelijk dat de gemeten eigenschap in de tijd aan verandering onderhevig is, bijvoorbeeld, in de context van dit artikel, indien een respondent tussen de twee afnames delicten pleegt. Dit is een verandering die men wenst te meten met de schaal, maar dergelijke effecten zullen de test-hertest betrouwbaarheid verlagen. In dit geval is de test-hertest betrouwbaarheid een nuttige indicatie van de stabiliteit van de schaalwaarden, maar is het geen schatting van de klassieke betrouwbaarheid (Lord & Novick, 1968: 61) meer.
Vaak is het moeilijk om tot de schatting van een test-hertest betrouwbaarheid te komen, omdat ook andere, *ongewenste*, storingsbronnen een rol spelen,

5. Mits er aan enkele methodologische voorwaarden is voldaan die wij verderop zullen bespreken.
6. De onderstaande beschouwing is niet bedoeld als inleiding in het bepalen van de betrouwbaarheid van een schaal (zie daarvoor de genoemde literatuur), maar zij beoogt Bruinsma's conclusies te weerleggen.

die zorgen dat de situatie op afname 1 niet identiek is aan die op afname 2. Crocker en Algina (1986: 134-135) wijzen hier vooral op de mogelijkheid dat de respondenten zijn veranderd ten gevolge van de eerste afname. In het standaardwerk *Educational Measurement* schrijven Feldt en Brennan (1991: 110) hierover dat er aan twee eisen moet zijn voldaan voordat er sprake is van een test-hertest betrouwbaarheid. In de *eerste* plaats mag er geen verandering te verwachten zijn tussen de twee afnames in de vaardigheid of karakteristieken van de respondenten[7]. In de *tweede* plaats mag de herinnering van de antwoorden op vragen van de eerste afname niet de antwoorden beïnvloeden op de tweede afname. Zij concluderen dat de vaststelling van de test-hertest betrouwbaarheid nauwelijks realiseerbaar is voor 'paper-and-pencil instruments' zoals gebruikt in Bruinsma's studies (vgl. Stanley, 1971: 406-407).

Voor de bepaling van de betrouwbaarheid van delinquentieschalen verkregen uit *self-report* gegevens is er dus behoefte aan een andere maat. Zo'n alternatieve maat voor betrouwbaarheid is dan de bekende Cronbach's alfa (Cronbach, 1951), die een ondergrens aangeeft voor de betrouwbaarheid gebaseerd op een enkele afname. Deze ondergrens ligt doorgaans hooguit enkele honderdsten onder de theoretische betrouwbaarheid[8], en dit verschil kan voor praktisch gebruik dus verwaarloosd worden. Om deze reden en omdat alfa slechts op een enkele afneming gebaseerd is, is het veruit de meest gebruikte schattingsmethode voor de betrouwbaarheid.

Vergeleken met een adequaat bepaalde test-hertest betrouwbaarheid veronachtzaamt Cronbach's alfa de fluctuaties van dag tot dag in het invullen van de vragenlijst. Dit lijkt de reden te zijn dat Bruinsma (1991, 1994) de test-hertest betrouwbaarheid prefereert boven de betrouwbaarheid zoals geschat met alfa. Maar evenals Stanley (1971) zijn wij van mening dat de test-hertest betrouwbaarheid voor deze schalen op deze wijze niet vastgesteld kan worden, en baseren wij ons voor de schatting van de betrouwbaarheid van *self-report* delinquentieschalen verder op Cronbach's alfa.

Onze conclusie op basis van Bruinsma's resultaten

Bruinsma vermeldt in beide studies voor de frequentieschaal schattingen van de alfacoëfficiënt voor de beide vragenlijstafnemingen in 1991 en in 1994. Deze waarden zijn .75, .76 , .63 en .75 (zie tabel 1). In de literatuur waagt men zich zelden aan uitspraken over wat een aanvaardbare hoogte is van de betrouwbaarheid van een test. Wij hebben slechts twee bronnen kunnen vinden, die in de praktijk door de Nederlandse COTAN[9] zijn overgenomen. Feldt en Brennan (1991: 106) noemen schalen met betrouwbaarheden van .70 of lager ongeschikt voor de beoordeling van studenten. Nunnally and Bernstein (1994: 264-265) noemen een betrouwbaarheid van .70 'modest',

7. Tenzij die verandering inherent is aan de gemeten eigenschap, bijvoorbeeld het plegen van delicten tussen de twee afnames.
8. Simulatiestudies van de tweede auteur (Sijtsma en Molenaar, 1987) wijzen hier op.
9. Dit is ons meegedeeld door dr J.J.F. ter Laak, vakgroep Ontwikkelingspsychologie, UU, die lid is geweest van de COTAN. De COTAN is een commissie van de beroepsvereniging van psychologen, het NIP, die schalen beoordeelt op betrouwbaarheid en validiteit.

maar maken vervolgens een onderscheid tussen schalen die zijn geconstrueerd voor verder sociaal-wetenschappelijk onderzoek, en schalen die zijn gemaakt voor individuele classificatie van personen. Voor het eerste gebruik stellen zij dat:

'in basic research, the concern is with the size of correlations and with the differences in means for different experimental treatments, for which purposes a reliability of .80 for the different measures involved is adequate'.

Hieruit concluderen wij dat de betrouwbaarheid van het *self-report* meetinstrument voldoende is voor verder gebruik in sociaal-wetenschappelijk onderzoek.

Nogmaals test-hertest betrouwbaarheid en Bruinsma's studies

De gevonden correlaties tussen de eerste afname en de tweede afname in Bruinsma's onderzoeken vragen om een nadere verklaring: in 1991 waren de correlaties -.14 en -.13, terwijl zij in 1994 .82, .84 en .83 zijn. Hoe kan dit? Door de opzet van Bruinsma's onderzoeken kunnen de correlaties volgens ons niet geïnterpreteerd worden als maten voor test-hertest betrouwbaarheid. Respondenten kunnen zich bij afname 2 proberen te herinneren wat zij bij afname 1 hebben geantwoord, en in dit geval is de test niet onder identieke omstandigheden afgenomen. Daarnaast is het mogelijk dat men langer over het onderwerp is gaan nadenken, en tot de conclusie is gekomen dat men bij afname 1 teveel of te weinig heeft gerapporteerd. Daarom alleen al zijn de gevonden correlaties niet te interpreteren als maten van test-hertest betrouwbaarheid.
Een ander aspect van de studieopzet vormt een bijzondere bedreiging van de test-hertest betrouwbaarheidsbepaling. De respondenten is de indruk gegeven dat hun eerdere formulier is weggeraakt. Wat voor invloed zou het hebben als een respondent dit niet gelooft? Hier kunnen we alleen maar naar raden. Het is niet uitgesloten dat er respondenten zijn die het onderzoek zullen saboteren door maar wat in te vullen, of in ieder geval niet al te zeer hun best doen om de juiste antwoorden te geven.
Vooral ongeloof bij de eerstejaars studenten (de onderzoeksgroep van de studie in 1991) lijkt ons niet denkbeeldig, omdat studenten toch een speciale onderzoeksgroep zijn. De lage negatieve correlaties die gerapporteerd worden in Bruinsma (1991) kunnen er volgens ons op wijzen dat er nogal wat studenten zijn geweest die de vragenlijst niet serieus hebben ingevuld. Wat te denken van studenten die op de frequentieschaal op de eerste afname tussen 17 of hoger scoren en op de tweede afname tussen 0 en 8 (dit zijn er 9), of studenten die op de eerste afname tussen 1 en 4 delicten bekennen en op het tweede afname 17 of meer? Het lijkt ons zeer onwaarschijnlijk dat hier sprake is van een toevallige meetfout. Wij hebben de indruk dat in 1991 de studenten op elk van de twee afnames een consistent beeld van zichzelf hebben willen schetsen (leidend tot de alfa's van .75 en .76), maar dat zij niet allen op een serieuze manier hebben meegewerkt aan het onderzoek door bij de eerste afname een ander consistent beeld van zichzelf te schetsen dan bij

de tweede afname (leidend tot de negatieve correlatie tussen afname 1 en 2). Als onze interpretatie juist is, dan is er niets mis met de betrouwbaarheid van het meetinstrument, maar heeft de gekozen onderzoeksopzet geleid tot het uit balorigheid niet serieus invullen van de vragen.
In Bruinsma (1994) vinden wij minder van dergelijke extreme verschillen tussen afname 1 en 2. Dit komt vermoedelijk doordat in deze studie niet is gekozen voor studenten als onderzoeksgroep. Maar voor deze studie vinden wij wel het vreemde[10] resultaat dat de correlaties (.82, .84 en .83) hoger zijn dan de Cronbach's alfa's (.63 en .75). Theoretisch gezien, hebben er immers meer verschillende toevalsfouten invloed op de correlaties (namelijk ook de verandering door de tijd) dan op de Cronbach's alfa's (die leiden tot een maat bepaald op 1 afname). Wij denken dat de respondenten in 1994 hun best hebben gedaan om bij de tweede afname hetzelfde in te vullen als bij de eerste afname, wat er toe heeft geleid dat de correlaties een overschatting geven van de betrouwbaarheid.

Geen twijfel aan betrouwbaarheid van self-report studies

Op basis van de gerapporteerde alfa's blijkt de betrouwbaarheid van de door Bruinsma gebruikte schalen voldoende. Daarbij dient vermeld te worden dat Cronbach's alfa doorgaans relatief laag uitvalt indien de verzameling van vragen heterogeen van inhoud is. De alfa kan soms worden verhoogd door inhoudelijk homogene clusters van vragen te identificeren en een schatting per cluster te maken, of door minder goed bij de vragenlijst passende vragen weg te laten en alfa voor de resterende vragen te schatten. Bruinsma heeft hierover niets gerapporteerd, maar er mag verondersteld worden dat de alfa-waarde van de schaal verder verhoogd kan worden door specifieke vragen weg te laten. Het kan ook zijn dat delinquent gedrag multidimensioneel is, en dat het mogelijk is subschalen te construeren met veel hogere alfa-waarden. Dit zou dan tot nog positievere conclusies kunnen leiden over de betrouwbaarheid van het *self-report* meetinstrument.
De resultaten van Bruinsma (1991, 1994) geven volgens ons dus geen enkele aanleiding te twijfelen aan de gangbare opvatting in de literatuur (zie bijvoorbeeld Hindelang e.a. 1981) dat de betrouwbaarheid van *self-report* studies voldoende is.

Literatuur

Bruinsma, G.J.N., 'De test-hertest betrouwbaarheid van de self-report methode', *Tijdschrift voor Criminologie*, 33/3 1991: 245-255.

Bruinsma, G.J.N., 'De test-hertest betrouwbaarheid van het meten van jeugdcriminaliteit', *Tijdschrift voor Criminologie*, 36/3, 1994: 218-235.

10. Dit is vreemd onder de aanname dat er sprake is van een delinquentieschaal die een-dimensioneel is. Indien er sprake is van een verzameling vragen die meer-dimensioneel is (bijvoorbeeld, als delinquentie meer-dimensioneel is, wat het geval is indien er sprake is van specialisatie), dan verlaagt dit de alfa-waarden terwijl de stabiliteit, zoals gemeten met de correlaties, groot kan zijn. In dit laatste geval is het overigens niet zinvol uitspraken te doen over een test-hertest betrouwbaarheid.

Crocker, L. & Algina, J., *Introduction to classical and modern test theory*, London: Holt, Rinehart and Winston, 1986.
Cronbach, L.J., 'Coefficient alpha and the internal structure of tests', *Psychometrika*, 16, 1951: 297-334.
Feldt, L.S. and Brennan, R.L., 'Reliability', in: R.L. Linn (ed.), *Educational measurement* (3rd edition), New York: Macmillan, 1991.
Hindelang, M.J., Hirschi, T., & Weis, J.G., *Measuring delinquency*, Beverly Hills: Sage, 1981.
Lord, F.M. & Novick, M.R., *Statistical theories of mental test scores*, Reading, MA: Addison-Wesley, 1968.
Nunnally, J.C. and Burnstein, *Psychometric theory* (3rd edition), New York: McGraw Hill, 1994.
Stanley, J.C., 'Reliability', in: R.L. Thorndike (Ed.) *Educational Measurement* (2nd edition), New York: Macmillan, 1971.
Sijtsma, K., Molenaar, I.W., 'Reliability of test scores in nonparametric item response theory', *Psychometrika*, 52, 1987: 79-97.

Repliek: de onbetrouwbaarheid van de self-report methode

G.J.N. Bruinsma

In hun reactie op mijn empirische studies naar de (test-hertest) betrouwbaarheid van de methode van zelfrapportage, stellen Van der Heijden, Sijtsma en 't Hart dat op grond van deze studies de test-hertestbetrouwbaarheid *niet* is vast te stellen en dat aan de Cronbach's alfa als maat voor betrouwbaarheid de voorkeur moet worden gegeven. Volgens hen valt het allemaal wel mee met de betrouwbaarheid van de *self-report*methode. In mijn reactie zal ik met name beargumenteren dat aan de betrouwbaarheid van het meten van gedrag andere eisen moeten worden gesteld dan aan het meten van houdingen, attitudes, cognities met behulp van testbatterijen waarvan zij uitgaan.

Kunnen mijn studies de test-hertest betrouwbaarheid vaststellen?

Wat betreft hun kritiek dat mijn onderzoeken de test-hertest betrouwbaarheid niet kunnen vaststellen, kan ik het volgende opmerken. In de *eerste* plaats volg ik de omschrijving van betrouwbaarheid van Swanborn[11] letterlijk in mijn onderzoek: het gaat om de stabiliteit in de scores bij herhaalde metingen! Ik heb derhalve om de betrouwbaarheid vast te stellen de metingen binnen een korte periode herhaald (2 weken (1991) en 1 maand (1994)). Daarna heb ik de verschillen tussen beide metingen op itemniveau en op schaalniveau vastgesteld. De vergelijking op itemniveau maakte ik om na te gaan welke delicten meer last hebben van toevallige meetfouten dan andere. Ik ben in mijn analyses vrij streng. Van der Heijden c.s. beweren dat deze strenge vergelijkingsmethode gebruikelijk is op nominaal meetniveau en niet voor metingen op intervalniveau. De vraag is echter of bij het meten van crimineel gedrag zo vanzelfsprekend van het intervalniveau mag worden uitgegaan. Om zo'n meetniveau te bereiken is het van belang dat respondenten exact kunnen en willen aangeven hoe vaak zij in de afgelopen periode bepaalde vormen van crimineel gedrag hebben gepleegd en waarbij het verschil tussen

11. En die vaker in de methodologische literatuur is te vinden.

1 en 2 keer net zo groot moet zijn als tussen 12 en 13. Dit stelt hoge eisen aan het geheugen. Wie weet bij voorbeeld, als hij daarover onverwachts wordt ondervraagd, exact hoe vaak hij het afgelopen jaar naar een feestje is geweest? Ik niet en naar het zich laat aanzien kunnen mijn respondenten dat ook niet. Het intervalniveau dat de drie kritici veronderstellen is discutabel, zeker wanneer de wijze waarop mensen schattingen van hun eigen gedrag maken in ogenschouw wordt genomen. Aan het schatten van het feitelijk uitvoeren van handelingen door mensen zit echter meer vast dan Van der Heijden c.s. zich lijken te realiseren. De antwoordscores zijn, in tegenstelling tot attitudemetingen, *niet normaal verdeeld.* In de regel zijn de antwoordscores bij *self-reports* scheef verdeeld, waarbij de modus in de lage scores ligt. Het is evenwel ook mogelijk dat de schattingen van respondenten over het voorkomen van hun feitelijke gedrag polymodaal of misschien wel logaritmisch zijn. Vooralsnog laten mijn studies zien dat een nominaal meetniveau (wel of niet een delict te hebben gepleegd) tot meer betrouwbare metingen leidt dan schattingen van het werkelijke aantal delicten door de respondenten. Bij een index waarbij gebruik wordt gemaakt van 'ooit-vragen' neemt de betrouwbaarheid verder toe.

Identieke omstandigheden bij beide metingen

Van der Heijden c.s. stellen als eis aan een test-hertest onderzoek dat *alle* omstandigheden bij de eerste en de tweede meting *identiek* aan elkaar moeten zijn. Dit is een beetje flauw omdat daaraan strikt genomen natuurlijk nooit kan worden voldaan. Bekend met de literatuur op dit gebied heb ik er wel voor gezorgd dat in beide studies de meeste, relevante omstandigheden waarin de beide interviews zijn gehouden, zoveel mogelijk dezelfde zijn. Uit een nadere analyse komt naar voren dat diverse mogelijke interview- en interviewerkenmerken *niet* significant samenhangen met de verschillen in antwoorden[12].
Tevens mogen er van Van der Heijden c.s. geen veranderingen in de vaardigheid of karakteristieken van de respondenten zijn te verwachten bij de tweede meting. In mijn studies vielen zulke veranderingen niet te verwachten. Voorts mag volgens hen de herinnering van de antwoorden van de eerste meting niet de antwoorden beïnvloeden van de tweede meting. Ik heb daarop telkens gecontroleerd en vond geen aanwijzingen voor een herinneringseffect. Mocht herinnering wel een rol hebben gespeeld in de tweede meting dan is in werkelijkheid de betrouwbaarheid van de *self-report*methode dus nog lager dan ik heb vastgesteld!

12. Ik heb nagegaan welke achtergrondvariabelen van de respondenten (leeftijd, sekse, e.d.), huiselijke condities (onderhoud van huis, straat, buurt) en interview- en interviewerkenmerken (lengte van interview, thuis of elders geïnterviewd, e.d.) van de twee metingen mogelijk systematisch samenhangen met de 'fouten' van de respondenten. Van zo'n samenhang blijkt in geen van de gevallen sprake te zijn. Deze analyses sterken mij in het oordeel over het toevallige karakter van de 'fouten'(d.i. onbetrouwbaarheid).

Cronbach's alfa

Van der Heijden c.s. geven voor de vaststelling van de betrouwbaarheid van het meten van crimineel gedrag de voorkeur aan de Cronbach's alfa boven de test-hertest betrouwbaarheid. Ik vind hun argumentatie op dit punt niet indrukwekkend. Hun op autoriteit gebaseerde aanbeveling dat de Cronbach's alfa een goede indicator is voor de betrouwbaarheid, heeft betrekking op *psychologische tests en andere schalen die zijn gebaseerd op unidimensionele indices van vragen en stellingen die slechts op één tijdstip worden onderzocht.* Maar een gecombineerd mondeling- en schriftelijk interview over feitelijk gedrag is geen 'paper-and-pencil instrument', zoals zij denken. In de twee studies wilde ik achterhalen of een respondent op de vraag hoe vaak hij/zij het afgelopen jaar een inbraak heeft gepleegd, niet bij voorbeeld eerst 'vijf keer' antwoordt en twee weken later slechts 'één keer'. Zulke betrouwbaarheidsonderzoeken worden vrijwel nooit uitgevoerd in de sociale wetenschappen, en zeker niet in de criminologie. Kennis van het antwoordgedrag in *self-reports* is voor de criminologie van groot belang omdat veel etiologische theorieën en overheidsinterventies op basis van *self-reports* worden ontwikkeld. Het gaat mij dus niet om de gemiddelde samenhang tussen de items[13] noch om te weten of respondenten in één interview consistent antwoorden op verschillende delictvragen.

Cronbach's alfa wordt door Van der Heijden c.s. als de veruit meest gebruikte schattingsmethode van betrouwbaarheid naar voren geschoven. Natuurlijk is dat zo omdat in de sociale wetenschappen (helaas) vrijwel uitsluitend één-moments surveys worden uitgevoerd, waarin de onderzoekers alleen maar de beschikking hebben over de alfa als indicator voor betrouwbaarheid. Met Cronbach's alfa kan, zoals bekend, de interne consistentie en de homogeniteit in antwoorden worden geschat (*wanneer er althans sprake is van unidimensionele schalen of indices*), maar niet de stabiliteit in antwoorden door de tijd. Van der Heijden c.s. beoordelen mijn alfa's als voldoende terwijl zij zouden moeten weten dat met de criminaliteitsmetingen een breed scala van gedragingen (29, resp. 33 delicten) worden gemeten die niet allemaal op één continuüm kunnen worden geplaatst (geweld, druggebruik, diefstal, enz.).

Zij zijn van mening dat de studenten uit de eerste studie mij op grote schaal hebben beduveld en dat in de tweede studie de respondenten hun uiterste best hebben gedaan (tevergeefs dus!) identiek aan de eerste meting te antwoorden. Dat zou best zo kunnen zijn. Ik weet echter niet waarop zij deze kennis baseren. Ik zou het niet weten want ik heb het mijn respondenten niet gevraagd. Het zou wel voor de criminologie interessant zijn deze verklaringen te onderzoeken. Maar dan wel in een *validiteitsonderzoek* en niet in een betrouwbaarheidsonderzoek omdat beide genoemde verklaringen betrek-

13. De waarde van de Cronbach's alfa wordt bepaald door de gemiddelde inter-itemcorrelatie en het aantal items van een meetinstrument. Men kan bij gelijkblijvende inter-itemcorrelatie de hoogte van betrouwbaarheid beïnvloeden door het meetinstrument te verlengen (meer vragen te stellen over crimineel gedrag). Bij meer dan 20 items kan de alfa nauwelijks lager worden, ook al is de gemiddelde inter-itemcorrelatie uiterst gering.

king hebben op *systematische* fouten en niet op *toevallige* fouten. En over systematische fouten gingen mijn beide studies dus niet.

Adviezen bij gebruik van de self-report methode

Je kunt, zoals Van der Heijden c.s., ontkennen dat er iets mis is met de betrouwbaarheid van de *self-report* methode. Je kunt mijn beide onderzoeken ook negeren omdat de resultaten daarvan je niet aanstaan. Voor onderzoekers die van de *self-report* methode gebruik maken, is het vervelend te moeten zien dat de 'bodem' in hun eigen empirisch onderzoek ter discussie wordt gesteld. De boodschap die ik met mijn beide publikaties heb willen uitdragen is *niet* dat de methode moet worden afgewezen maar *wel* dat een grote terughoudendheid en een grote voorzichtigheid in acht moet worden genomen bij het gebruik van de *self-report* methode. Als advies zou ik het volgende op grond van mijn artikelen willen aanbevelen:
1. Maak bij de *self-report* methode gebruik van meerdere items omdat bij een klein aantal vragen toevallige fouten een te grote, negatieve rol zullen spelen;
2. Inspecteer in het bijzonder de item-nonresponse bij de *self-report* methode[14]. De meeste toevallige fouten zitten in de nonresponse van de afzonderlijke items. Normaal gesproken wordt nonresponse van afzonderlijke items uit het databestand verwijderd of buiten de berekeningen gehouden. Bij de *self-report* methode zou ik deze praktijk afraden, omdat in de nonresponse juist mensen 'zitten' die om onbekende redenen (aarzeling) de ene keer wel en de andere keer geen antwoord geven op vragen over hun criminaliteit;
3. Ga niet proberen de antwoorden zo precies mogelijk voor te schrijven (Bijv. 1 tot en met 15). Uit mijn onderzoeken komt naar voren dat mensen niet goed in staat zijn een exact aantal delicten te schatten. De betrouwbaarheid van de *self-reports* neemt toe naarmate met categorieën als 'wel of niet het afgelopen jaar een delict gepleegd' wordt gewerkt. De betrouwbaarheid neemt verder toe wanneer met 'ooit-vragen' wordt gewerkt;
4. Maak op basis van zelfrapportage nooit een tweedeling in de onderzochte populatie in 'delinquent' en 'niet-delinquent' omdat zo'n tweedeling is gebaseerd op toevallige scores (fouten) en dus in principe *random* is. Het construeren van afzonderlijke theorieën en verklaringen of beleidsaanbevelingen voor de beide, op deze wijze geconstrueerde categorieën is op drijfzand gebaseerd.

Dupliek: over betrouwbaarheid, stabiliteit en artefacten

Peter G.M. van der Heijden, Klaas Sijtsma en Harm 't Hart

De redactie heeft ons in de gelegenheid gesteld kort te reageren op Bruinsma's reactie. Wij doen dit in een klein aantal punten.

14. Ik laat de nonresponse in de dataverzameling (onbereikbaarheid, niet willen meewerken, enz.) hier gemakshalve buiten beschouwing omdat deze de validiteit van de meting sterk negatief beïnvloedt.

1. Test-hertest betrouwbaarheid versus Cronbach's alfa

Bruinsma stelt dat aan de betrouwbaarheid van het meten van gedrag andere eisen moeten worden gesteld dan aan het meten van houdingen, attitudes en cognities met behulp van testbatterijen. Dit is niet gebruikelijk. Het is goed zich te realiseren dat men niet alleen binnen de criminologie werkt met selfreportgegevens met betrekking tot gedrag, maar ook binnen andere takken van de sociale wetenschappen, zoals de persoonlijkheidspsychologie. Een gecombineerd mondeling- en schriftelijk interview over feitelijk gedrag wordt daar wel degelijk als een 'paper-and-pencil instrument' beschouwd.
Bruinsma beroept zich op Swanborn, die passages aan betrouwbaarheid heeft gewijd in zijn boeken. Wij denken dat het beter is gespecialiseerde standaardwerken op het gebied van betrouwbaarheidsrekening te raadplegen, zoals Lord en Novick (1968), waarin de betrouwbaarheid wordt gedefinieerd als de product-moment correlatie tussen 2 parallelle tests (ibid, 61) en niet als de product-moment correlatie tussen herhaalde metingen met dezelfde test. De laatste definitie zou alleen identiek aan de betrouwbaarheid zijn als er inderdaad sprake is van identieke omstandigheden. Dit houdt bijvoorbeeld in dat de respondenten tussen beide afnemingen niet systematisch mogen veranderen. Ook Swanborn is het hier mee eens (P.J. Swanborn, *Methoden van sociaal-wetenschappelijk onderzoek (nieuwe editie)*, Meppel: Boom, 1987: 180). Het is wel gebruikelijk om de correlatie tussen herhaalde metingen als indicatie van stabiliteit te beschouwen, maar dat is alleen zinvol als er tussen beide metingen geen rare dingen gebeuren (en dit lijkt het geval: zie onze eerste reactie en punt 3).
Bruinsma heeft ook een ongebruikelijke opvatting over Cronbach's alfa. Cronbach's alfa is een goede indicator voor de betrouwbaarheid. Hierbij beroepen wij ons niet op een autoriteit, maar op een wiskundig bewijsbare stelling (Novick and Lewis, 1967) die als resultaat heeft dat alfa kleiner of gelijk is aan de betrouwbaarheid. Verder is de definitie van alfa nergens gebaseerd op welke vooronderstelling dan ook over het aantal dimensies dat men meet met de test of vragenlijst. Wel is het zo dat in een inhoudelijk heterogene test (meerdimensioneel) alfa vaak wat laag kan uitvallen. Dit stond aan het slot van onze eerste reactie.

2. Informativiteit van de gegevens

Bruinsma lijkt vast te willen houden aan zijn strenge eisen bij het meten van betrouwbaarheid: hij wil niet een intervalniveau aannemen. Zoals in onze eerste reactie reeds gesteld: dit is ongebruikelijk. Zijn motieven hiervoor zijn ons niet zo duidelijk. Zijn aanname dat attitudemetingen normaal verdeeld zijn is bijvoorbeeld onjuist, maar ook irrelevant (een intervalniveau veronderstelt niet dat de scores normaal verdeeld zijn). Het is wiskundig eenvoudig aan te tonen dat zijn bevinding 'vooralsnog laten mijn studies zien dat een nominaal meetniveau tot meer betrouwbare metingen leiden dan schattingen van het werkelijke aantal delicten door de respondenten' voor elke studie geldt.

Hoe zinvol is het om stabiliteit over de tijd te bekijken met een nominaal meetniveau? Het zal duidelijk zijn dat de instabiliteit dan veel groter wordt ingeschat dan wanneer men een correlatie gebruikt. De aanname van een nominaal meetniveau lijkt vooral zinvol wanneer men behoefte heeft aan zeer preciese informatie over respondenten. Wanneer bijvoorbeeld het verschil tussen 10 en 12 gerapporteerde delicten van belang is, dan is het nuttig de stabiliteit op nominaal meetniveau te onderzoeken. Dit lijkt ook het geval te zijn wanneer de interesse uitgaat naar antwoorden op itemniveau. Echter, in het meeste surveyonderzoek geldt dit niet, want men is meestal geinteresseerd in de relaties tussen schaalscores en de scores op andere variabelen, en kleine verschillen spelen dan een verwaarloosbare rol.

3. Conclusie

Gezien ons eerste punt blijven wij bij onze stelling dat Bruinsma's studies wijzen op een voldoende betrouwbaarheid van het *self-report*meetinstrument, waarbij wij ons baseren op de alfa's. Indien men de correlatie tussen meting 1 en 2 als betere betrouwbaarheidsmaten zou willen zien, dan moet opgemerkt worden dat deze maat voor de eerste studie negatief (en dus onaanvaardbaar laag) is en voor de tweede studie maar liefst groter dan .80 (en dus goed, in tegenstelling wat Bruinsma in zijn reactie beweert door te stellen dat de studenten 'tevergeefs' geprobeerd hebben identiek te antwoorden) is. Het is opmerkelijk dat Bruinsma op deze nogal verschillende betrouwbaarheidsresultaten niet ingaat. Hij vecht onze interpretatie aan zonder zelf een andere te leveren.