



## VASTSTELLEN VAN DE VALIDITEIT VAN REGISTERVARIABLEN<sup>1</sup>

**Bart F.M. Bakker<sup>2</sup>**

*Summary: Data from administrative sources and surveys have measurement and representation errors. One of the measurement errors is deficiency in validity. We apply the classical test theory to determine the validity of administrative measures of age, gender, educational attainment and wages. A Simultaneous Equations Model is applied to combined data from the OSA-panelsurvey and administrative data from the Social Statistical Database. The results show that age and gender have high and educational attainment and wages medium levels of validity. The method is useful for determining the validity of administrative variables as long as the true associations between the variables are roughly known.*

*Keywords: Validity, administrative data, registers, data linkage, data quality, Simultaneous Equations Model.*

### **1. Inleiding**

Het gebruik van administratieve gegevens, ook wel registraties genoemd, voor sociaalwetenschappelijk onderzoek neemt toe. Daarvoor zijn vele redenen aan te

---

<sup>1</sup> Dit paper is gepresenteerd op de conferentie “Registers voor sociaalwetenschappelijk onderzoek. Mogelijkheden en valkuilen” dat gehouden is op 26 juni 2010 te Leiden.

<sup>2</sup> De auteur is bijzonder hoogleraar Methodologie van registerdata aan de Vrije Universiteit te Amsterdam en manager van de taakgroep Sociaal Economisch totaalbeeld van het CBS. Dit artikel is op persoonlijke titel geschreven.

voeren. De kwaliteit van de informatie die verkregen wordt uit enquêtes is minder geworden, voornamelijk vanwege de toename en selectiviteit van non-response. Enquêtes zijn bovendien duur in vergelijking met registraties. Door ontwikkelingen in de informatietechnologie zijn steeds meer registraties beschikbaar gekomen. Bovendien is door de invoering van het Sofinummer, tegenwoordig BurgerServiceNummer geheten, de mogelijkheden vergroot om registraties onderling te koppelen waardoor meer bruikbare databestanden ontstaan. De belangrijkste oorzaak van de toename is echter gelegen in de nieuwe onderzoeksmogelijkheden die registraties bieden (Bakker, 2009b).

Ondanks de toename van het gebruik van registratieve gegevens in sociaalwetenschappelijk onderzoek, is nog maar weinig bekend over mogelijke kwaliteitsproblemen ervan. Een eerste poging om systematisch in kaart te brengen welke soorten fouten onderscheiden zouden moeten worden onderscheiden is opgenomen in Bakker (2009b). In de volgende paragraaf wordt daarvan een beknopte samenvatting gegeven.

## **2. Meetfouten**

Registraties worden dikwijls verkregen via een of andere vorm van elektronische, schriftelijke of mondelinge enquêtering. Of u zich nu inschrijft als werkzoekende, een belastingformulier invult of uw kind aanmeldt bij een school, steeds moeten formulieren ingevuld worden door uzelf of door een vertegenwoordiger van de registratiehouder. Dit betekent dat de fouten die in enquêtes voorkomen ook voor kunnen komen in registraties. Die fouten worden uitgebreid beschreven in de literatuur (zie bijvoorbeeld Groves et al. 2004, Bethlehem, 2009, 178-248). Daarnaast zijn er fouten die specifiek gelden voor registraties (Bakker, 2009b).

Op de meting van variabelen in registers heeft de onderzoeker in de praktijk vrijwel geen invloed. Dat betekent dat het administratieve begrip het uitgangspunt is en dat slechts door gebruik te maken van verschillende variabelen die in gecombineerde registraties aanwezig zijn, het sociaalwetenschappelijke concept kan worden benaderd. Ontbreken die gegevens, dan moet noodgedwongen met het administratieve concept worden gewerkt. Voor de validiteit heeft dat negatieve gevolgen.

De omvang van de fouten in registers en de vertekening wordt mede bepaald door de mate waarin de fouten door de registratiehouder worden signaleerd en gecorrigeerd. Uiteraard worden fouten in registraties zoveel mogelijk voorkomen door het inbouwen van mechanismen in het administratieve proces. Zo kunnen belastingambtenaren uw aangifte controleren door een koppeling uit te voeren met andere registratieve data van bijvoorbeeld de UWV of de banken. Evenzo kan de arbeidsbemiddelaar aan u vragen om diploma's te overleggen wanneer hij of zij uw hoogste succesvol afgeronde opleiding invoert. Juist de variabelen die belangrijk zijn voor de registerhouder worden gecontroleerd. Deze zullen daarom gemiddeld

beter van kwaliteit zijn dan data die niet door de registerhouder worden gecontroleerd.

De omvang van de fouten en de vertekening in registratieve metingen worden ook bepaald door de belangen die de geregistreerden hebben bij een bepaalde registratie. Een voorbeeld daarvan is het ingeschreven zijn als uitwonend voor het verkrijgen van een hogere studiebeurs. Op basis van de belangen van de registerhouder en de geregistreerden kunnen hypothesen worden geformuleerd over de vertekening die er in registers is opgenomen. Dat is een terrein dat bij mijn weten nog volledig onontgonnen is, maar het wel verdient om verder te exploreren.

Administratieve vertraging is een belangrijke foutenbron in registers. Daarvan is sprake als een gebeurtenis later wordt geregistreerd dan hij heeft plaatsgevonden. Het duidelijkste voorbeeld is dat van de vertraging in de registratie van huwelijken van allochtonen die trouwen in hun land van herkomst. Vaak wordt pas als de huwelijkspartner overkomt het huwelijk in de GBA geregistreerd. Soms gaat daar een periode van 1 à 2 jaar overheen.

Een ander kwaliteitsprobleem treedt op wanneer koppeling van registraties teveel gemiste of miskoppelingen oplevert. Van gemiste koppelingen is sprake wanneer een deel van de records niet gekoppeld kan worden. Dat kan leiden tot een vergelijkbare selectiviteit als selectieve non-respons bij enquêtes. Van miskoppelingen is sprake als de informatie van twee eenheden ten onrechte aan elkaar gekoppeld wordt. Dit kan leiden tot een fout die vergelijkbaar is met een meetfout in enquêtes.

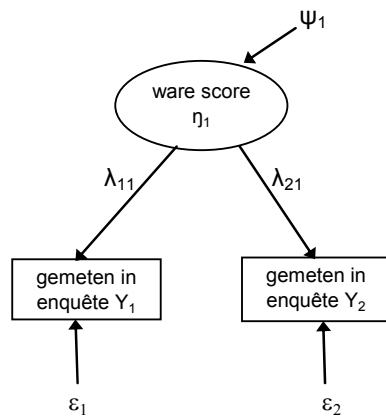
### **3. Validiteit en betrouwbaarheid**

De kwaliteit van de meting van een variabele kan met verschillende indicatoren worden bepaald. De belangrijkste daarvan zijn betrouwbaarheid en validiteit. Bij betrouwbaarheid gaat het om willekeurige fouten, dat wil zeggen dat als we verschillende steekproeven zouden trekken uit de populatie de spreiding van de schattingen rondom het werkelijke gemiddelde zijn gespreid. De betrouwbaarheid van een meetinstrument kunnen we vaststellen door deze (ten minste) twee maal te gebruiken bij dezelfde (steekproef)populatie. In figuur 1 is het model weergegeven dat dan wordt geschat. In het ovaal is de ware score  $\eta_1$  opgenomen. Dit wordt ook wel de latente variabele genoemd, waarmee het gewenste concept wordt weergegeven. Deze wordt gemeten door twee maal het meetinstrument af te nemen. Het product van de  $\lambda_{11}$  en de  $\lambda_{21}$  is de waargenomen correlatie tussen de metingen op de beide tijdstippen en wordt ook wel de test-hertest-correlatie genoemd. De  $\varepsilon_1$  en de  $\varepsilon_2$  zijn de meetfouten. Dit model kan trouwens niet worden uitgerekend, omdat er meer onbekenden zijn dan vergelijkingen. Door aan te nemen dat  $\lambda_{11}$  en  $\lambda_{21}$  aan elkaar gelijk zijn, kan dat wel.

Bij validiteit gaat het om systematische fouten, dat wil zeggen dat als we verschillende steekproeven zouden trekken het gemiddelde van de schattingen niet

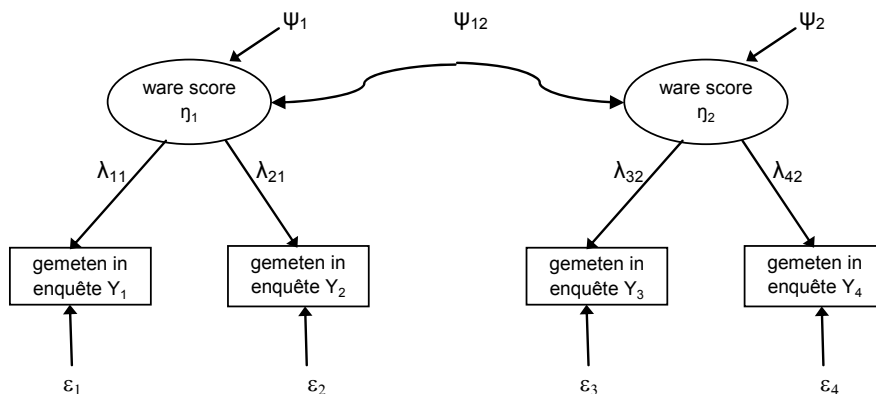
het ware gemiddelde opleveren. We spreken dan ook wel van vertekening. Er zijn verschillende soorten validiteit van een variabele te onderscheiden. In de eerste plaats gaat het dan om de zogenaamde kruisvaliditeit. Daaronder wordt de samenhang met een verwante variabele verstaan. Als we bijvoorbeeld weten dat een variabele goed gemeten is met een bepaald meetinstrument in een enquête, dan kan

**Figuur 1.**  
**Vaststellen van de betrouwbaarheid door herhaalde meting**



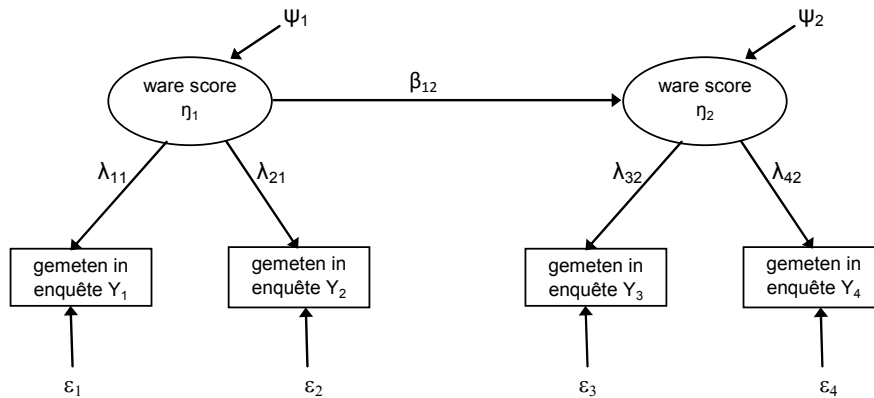
de kruisvaliditeit van een variabele met een ander meetinstrument, gecorrigeerd voor de mate van betrouwbaarheid van beide meetinstrumenten, met behulp van het model in figuur 2 worden geschat. Daarbij wordt bij dezelfde (steekproef)populatie beide meetinstrumenten twee keer afgenomen. De  $\psi_{12}$  is dan de kruisvaliditeit: de voor onbetrouwbaarheid gecorrigeerde correlatie tussen de ware scores  $\eta_1$  en  $\eta_2$ . De producten van  $\lambda_{11}$  en  $\lambda_{21}$  en van  $\lambda_{32}$  en  $\lambda_{42}$  zijn weer de test-hertest-correlaties van respectievelijk de meetinstrumenten 1 en 2. De  $\varepsilon_1$  en  $\varepsilon_2$  en de  $\varepsilon_3$  en  $\varepsilon_4$  zijn de metingen voor onbetrouwbaarheid.

**Figuur 2.**  
**Kruisvaliditeit van een meetinstrument, gecorrigeerd voor meetonbetrouwbaarheid**



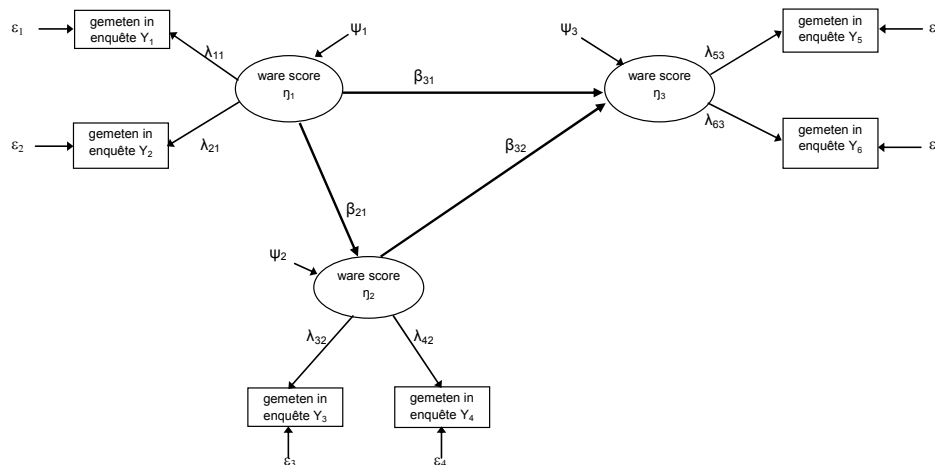
Een tweede soort validiteit is de predictieve validiteit. Daarmee wordt bedoeld dat het meetinstrument de variabele zo moet meten dat hij een andere variabele goed voorspelt. Als we weer corrigeren voor de onbetrouwbaarheden van de beide meetinstrumenten, ziet het model eruit als in figuur 3.

**Figuur 3.**  
**Predictieve validiteit van een meetinstrument, gecorrigeerd voor de meetonbetrouwbaarheid**



De laatste vorm van validiteit die ik bespreek is die van de constructvaliditeit. Daarbij gaat het erom dat meer dan twee variabelen onderling samenhangen volgens verwachting. Hierbij gaat het dus om ten minste drie variabelen. Als we weer corrigeren voor de meetonbetrouwbaarheid van de variabelen, ziet het model er uit zoals weergegeven in figuur 4.

**Figuur 4**  
**Constructvaliditeit van meetinstrumenten gecorrigeerd voor de meetonbetrouwbaarheid**



Uiteraard is een noodzakelijke voorwaarde voor het vaststellen van de validiteit van een meetinstrument dat we *weten* op welke wijze de variabelen met elkaar samenhangen. Dat geldt voor de kruisvaliditeit, de predictieve en de constructvaliditeit. Dit referentiekader is lang niet altijd aanwezig. Vaak is er niet meer dan een theoretisch idee op welke wijze variabelen samenhangen. Als dat het geval is, kunnen meerdere meetinstrumenten worden vergeleken. Die meetinstrumenten die het meest in overeenstemming zijn met de theoretisch verwachte samenhangen hebben dan de hoogste validiteit.

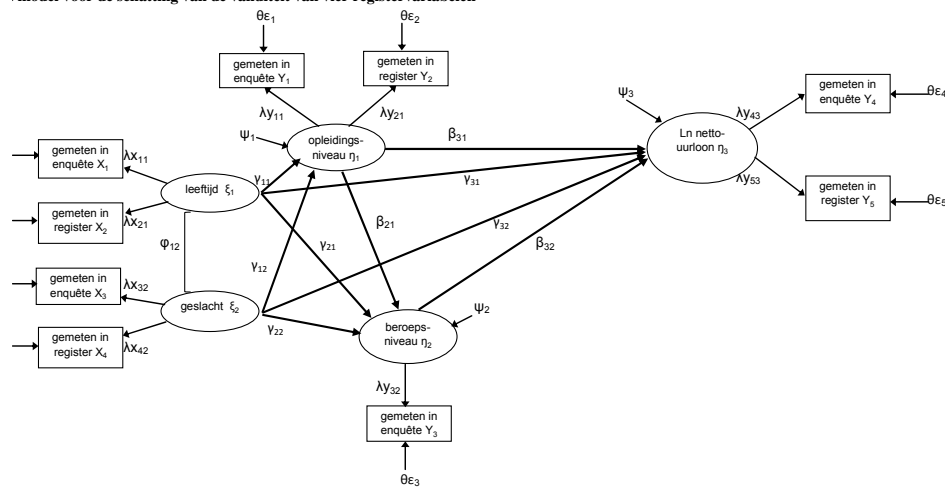
Tot nu toe gingen we ervan uit dat we alle variabelen twee maal via dezelfde vraagstelling in een enquête hebben gemeten. Daarmee kan worden vastgesteld of de vraagstelling betrouwbare resultaten oplevert. Het ligt dan voor de hand om aan te nemen dat de  $\lambda$ 's die bij dezelfde vraagstelling behoren aan elkaar gelijk te stellen.

Dat hoeft overigens niet. Een vraagstelling kan bij herhaling toch een ander resultaat opleveren. Dit kan bijvoorbeeld veroorzaakt worden door een verschillende context of door geheugeneffecten. Dit is bijvoorbeeld het geval als de vragen binnen hetzelfde vraaggesprek wordt herhaald, de respondent gedurende het vraaggesprek nog nagedacht heeft over het antwoord en bij de herhaling tot een ander antwoord komt. In het model zoals dat in figuur 6 is weergegeven wordt dat geschat: de  $\varepsilon_3$  en de  $\varepsilon_4$  kunnen een andere waarde aannemen.

Het is ook mogelijk dat er verschillende vraagstellingen worden gebruikt waarvan verondersteld wordt dat beide het gewenste concept meten. Het ligt dan veel minder voor de hand dat de twee verschillende vraagstellingen het concept even goed meten. Dit komt tot uitdrukking in een verschil in de  $\lambda$ 's die bij hetzelfde latente variabele behoren en uiteraard ook bij een verschil in de  $\varepsilon$ 's. Die krijgen dan echter een andere betekenis dan een meting van de meetonbetrouwbaarheid: het wordt dan een meting van de mate waarin het latente concept wordt gemeten en kan als zodanig dus als een meting van de validiteit worden beschouwd. Idealiter zou je ieder van de vraagstellingen twee keer moeten afnemen om weer de betrouwbaarheid en validiteit te kunnen onderscheiden.

Hoe kunnen we nu gebruik maken van deze modellen om de validiteit van registervariabelen vast te stellen? In plaats van twee verschillende vraagstellingen in een enquête op te nemen, kun je ook een enquête koppelen aan een register en zo een meting uit een enquête vergelijken met de meting uit een register. Daarvoor is bijvoorbeeld het loonverwervingsmodel bruikbaar zoals weergegeven in figuur 5. In het model zijn vijf latente variabelen opgenomen: leeftijd, geslacht, opleidingsniveau, beroepsniveau en (ln) netto-uurloon. Alle variabelen met uitzondering van het beroepsniveau zijn gemeten in een enquête en in een register. Aangetekend moet worden dat de variabele opleidingsniveau niet in zijn geheel een registervariabele is. Het betreft een combinatie van register- en enquête-informatie. Het grootste deel van de mensen heeft echter een opleidingsniveau vanuit een administratief bestand toegekend gekregen.

uur 5.  
model voor de schatting van de validiteit van vier registervariabelen



We verwachten dat de effecten van de meest valide meting van opleidingsniveau op uurloon het grootst is, gecontroleerd voor leeftijd, geslacht en beroepsniveau. Meetfouten leiden immers in de meeste gevallen tot lagere correlaties en daarmee tot lagere effecten. Aangezien hier het effect niet voor de afzonderlijke variabelen wordt vastgesteld, maar van de gemeenschappelijke variantie van de twee metingen, vertaalt dit zich in een lagere meetfout voor de meest valide metingen. Het model wordt uitgerekend voor mensen die in loondienst werkzaam zijn voor 12 uur of meer per week. Daarmee worden kleine baantjes waarvoor de samenhangen wel eens anders kunnen liggen, uitgesloten.

## 4. De data

### 4.1 Enquêtedata

We maken voor wat betreft de enquêtedata gebruik van het zogenaamde OSA-Aanbodpanel 2004. In het databestand zijn 4785 respondenten opgenomen. De steekproef is een huishoudensteekproef gestratificeerd naar leeftijd, geslacht, regio en huishoudensamenstelling. Personen die jonger zijn dan 16 of ouder dan 64 jaar of personen die volledig dagonderwijs volgen zijn uitgesloten. De eerder deelnemende huishoudens zijn opnieuw benaderd. Als een persoon een huishouden verlaat wordt het gehele nieuwe huishouden benaderd. Personen die buiten de doelpopulatie vallen worden niet langer gevolgd. Personen die bij een eerdere golf niet maar inmiddels wel tot de doelpopulatie vallen, worden wel ondervraagd. De data zijn vervolgens zodanig herwogen dat de steekproef de verdeling van bovengenoemde variabelen representeert zoals die in de Enquête beroepsbevolking worden gemeten (zie verder NIWI, 2010). De vraaggelbesprekken zijn gehouden rond de datum van 1 oktober 2004. Dat wordt beschouwd als de peildatum van deze OSA-golf.

*Leeftijd* is in de enquête gemeten door de geboortedatum te vragen en deze vervolgens om te rekenen naar een leeftijd op de enquêtedatum. *Geslacht* is gemeten met de vraag: “Wat is uw geslacht?”

*Opleidingsniveau* is gemeten met de vraag: “Wat is de hoogste opleiding die u heeft voltooid, waar u een diploma voor heeft behaald?” De respondent krijgt dan een kaart voorgelegd waarop 40 opleidingen genoemd worden. Deze opleidingen stammen uit een lange periode om alle generaties voldoende keuzemogelijkheden te bieden. Enkele voorbeelden daarvan zijn: “lagere school, basisschool”, “LTS: Lagere Technische School”, “MBO: Middelbaar Beroepsonderwijs”, “Gymnasium”, “Primair leerlingwezen”, “BOL: assistent-opleiding”, “Dagopleiding of deeltijdopleiding in het buitenland”. Als het beroepsonderwijs of een buitenlandse opleiding betreft, wordt gevraagd: “Welk type school en studierichting volgde u in het laatste jaar daarvan?”. Deze informatie is later omgezet in een code van de Standaard Onderwijs Indeling.

*Beroepsniveau* is vastgesteld met een uitgebreide vraagstelling waarin onder andere naar de belangrijkste werkzaamheden, het aantal personen waarover leiding wordt gegeven, en de belangrijkste leidinggevende werkzaamheden is gevraagd. De informatie is vervolgens gecodeerd in de Standaard Beroepenclassificatie 1992 (CBS, 1993).

Het *ln netto-uurloon* is vastgesteld door de vraag: “Kunt u mij zeggen hoeveel uw netto maandloon bedraagt?”, waarbij de interviewer eerst noteert of de respondent een bedrag per week, per vier weken, per maand of per jaar noemt en vervolgens het desbetreffende bedrag. Verder is de vraag gesteld naar hoeveel uren men volgens het contract werkt. Deze informatie heeft steeds betrekking op de hoofdbaan. Door de informatie te combineren is het uurloon berekend, waarna dit logaritmisch is getransformeerd.

## 4.2 Registerdata

De registerdata zijn ontleend aan het Sociaal Statistisch Bestand (SSB). Dit is een stelsel van koppelbare registers en enkele (steekproef)enquêtes, waarvan de definitieve versie zodanig is opgeschoond dat tot op grote hoogte consistentie in de data is bereikt (Arts & Hoogteijling, 2002; Bakker, 2008). De methodologie die daarbij gebruikt wordt is micro-integratie (Bakker, 2009a). Door micro-integratie wordt de kwaliteit van de data verhoogd, omdat deze worden geharmoniseerd, gecompleteerd en ontdaan van meetfouten. Daarbij wordt een stelsel van beslisregels toegepast. Daarna worden de bestanden register genoemd. In het onderstaande bespreken we de variabelen en geven daarbij ook in grote lijnen aan welke beslisregels daarop van toepassing zijn geweest.

*Leeftijd* en *geslacht* worden ontleend aan de Gemeentelijke BasisAdministratie (GBA). Deze wordt voor demografische variabelen kwalitatief beter geacht dan dezelfde variabelen uit andere bronnen. Mochten personen niet voorkomen in de GBA dan worden hun leeftijd en geslacht ontleend aan andere bronnen. In het databestand dat voor het schatten van het model wordt gebruikt, worden alleen gekoppelde records gebruikt die daarom per definitie ook in de GBA voorkomen.

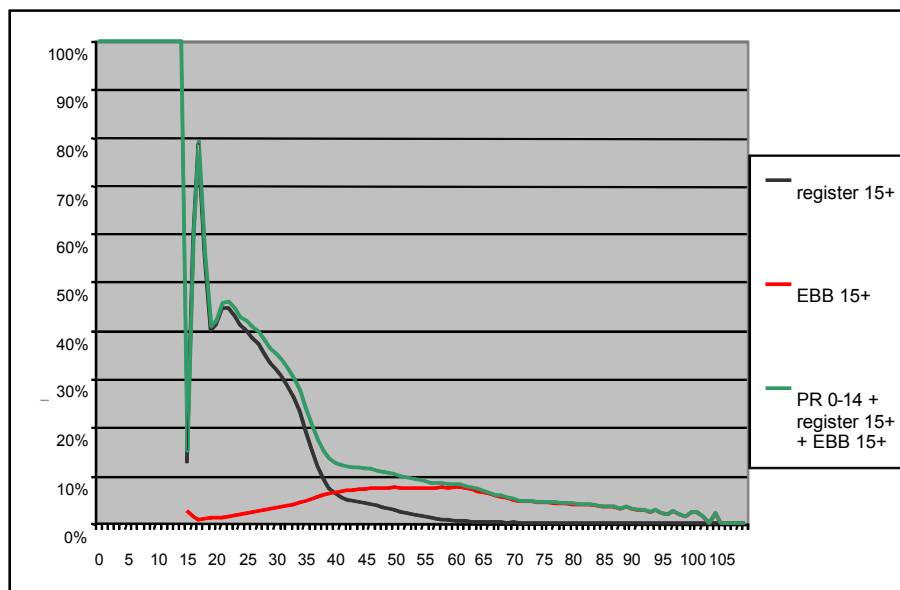
Voor *opleidingsniveau* is geen registratie beschikbaar waarmee de Nederlandse bevolking wordt gedekt. Wel is informatie opgenomen in verschillende registraties van scholieren en studenten zoals de inschrijvingen in het hoger onderwijs (sinds 1983), het Examen Resultaten Register (sinds 1999), de onderwijsnummerbestanden voortgezet onderwijs (vanaf 2004), de onderwijsnummerbestanden middelbaar beroepsonderwijs (vanaf 2005) en enkele kleinere registraties. Naast deze administratieve bronnen worden de gegevens uit de Enquête beroepsbevolking (EBB) gebruikt van de jaargangen 1996-2008. Daarvan wordt getoetst of het opleidingsniveau op het peilmoment nog geldig is. Als de kans groter is dan 5% dat het opleidingsniveau niet geldig is, wordt het record van de EBB verwijderd. Bij opleidingsniveau wordt altijd de hoogste opleiding genomen die in de bronnen is terug te vinden. Als twee bronnen een gelijk opleidingsniveau laten zien, wordt de opleidingsrichting gekozen uit de registraties. Dat laatste is niet van belang voor dit



onderzoek, omdat alleen het opleidingsniveau wordt gebruikt. Verder wordt aangenomen dat kinderen van 12 jaar en jonger nog geen diploma in het voortgezet onderwijs hebben behaald. Voor zover 12-14-jarigen niet in een onderwijsnummerbestand voorkomen, wordt hetzelfde verondersteld. Zo wordt in totaal ongeveer van 7 miljoen mensen het opleidingsniveau vastgesteld. Van deze informatie wordt op verschillende peilmomenten het opleidingsniveau bepaald. Hier wordt gebruik gemaakt van de informatie over het opleidingsniveau op ultimo september 2004. Dat is nagenoeg gelijk aan het peilmoment van het OSA-Aanbodpanel 2004.

Kenmerk van de registraties met onderwijsinformatie is dat ze informatie van recente datum bevatten. Dat betekent dat de registerinformatie selectief is naar leeftijd en alle kenmerken die samenhangen daarmee. Een andere beperking is dat in de onderwijsregistraties alleen het bekostigd onderwijs is opgenomen. Bedrijfsopleidingen en een belangrijk deel van het schriftelijk onderwijs zijn daarin niet opgenomen. Om toch een schatting van het opleidingsniveau van de Nederlandse bevolking te kunnen maken wordt gebruik gemaakt van de EBB waarin informatie over ouderen en niet-bekostigd onderwijs wel is opgenomen. Bij de ouderen zijn de gegevens voornamelijk ontleend aan de EBB. In figuur 8 is de dekkingsgraad weergegeven van de gegevens uit het opleidingsregister. De groene lijn is de totale dekking, de zwarte lijn is de dekking op basis van de registraties en de rode lijn die van de EBB. De totale dekking is tot 14 jaar volledig, valt dan even sterk terug, om dan vanaf 18 jaar geleidelijk van 80% terug te lopen naar minder dan 10% voor de 55-plussers. Om vervolgens de gegevens een goede afspiegeling van de populatie te laten zijn is een weegmodel ontwikkeld voor verschillende peilmomenten (zie verder Bakker, Bouman en Van Toor, 2006; Bakker, Linder en Van Roon, 2008).

**Figuur 6.**  
De dekkingsgraad van het opleidingsregister



Het beroepsniveau is helaas niet in registers opgenomen. Daarom wordt die alleen in de enquête gemeten. Het *In netto-uurloon* wordt als volgt gemeten. Eerst wordt het netto-jaarloon bepaald van de hoofdbaan als het fiscaal loon verminderd met de afgedragen premies en belastingen. Het fiscaal loon is ontleend aan de gegevens die bedrijven leveren aan de Belastingdienst voor het afdragen van belastingen en premies. Het fiscaal loon en de afgedragen premies en belastingen is benaderd voor de maand september 2004 door het jaarloon te delen door het aantal maanden dat de desbetreffende baan duurde die de respondent in september had. Daarmee is het een goede benadering van het begrip zoals dat in de enquête wordt gemeten. Het verschil betreft de bijzondere beloningen voor overwerk e.d. die in het netto-maandloon niet en in het fiscaal maandloon wel worden meegeteld. Voor het bepalen van het uurloon is gebruik gemaakt van de informatie zoals deze in de OSA-enquête voorkwam over het aantal gewerkte uren volgens contract. Dit was noodzakelijk omdat een goede registermeting voor het aantal contracturen voor het verslagjaar 2004 niet beschikbaar was.

### **4.3 Koppeling van de data**

De koppeling van de data hebben we uitgevoerd door aan elk bestand een betekenisloos koppelnummer toe te kennen.

Het OSA-aanbodpanel 2004 is gekoppeld met behulp van de sleutel geboortedatum, geslacht, postcode, en huisnummer. Dit levert een relatief hoog koppelrendement op van 98,9%. Van de 4782 personen die deelnamen aan de enquête konden 4730 personen van een koppelsleutel worden voorzien (Fouarge & Grim, 2007). Inspectie van de gegevens laat zien dat de uitgevallen 1,1% niet erg selectief is. Van deze 4730 gekoppelde personen werken er 3237 personen in loondienst.

Voor de registerinformatie geldt dat ook de afzonderlijke administratieve bestanden van een dergelijk betekenisloos koppelnummer zijn voorzien, waarbij vooral gebruik is gemaakt van het Sofinummer. De koppelrendementen van deze administratieve bestanden verschillen, maar ligt meestal boven de 97%. Bovendien wordt een belangrijk deel van de niet-gekoppelde records terecht niet gekoppeld, vanwege het feit dat deze niet tot de populatie behoren.

Vervolgens wordt gebruik gemaakt van het koppelnummer om het OSA-Aanbodpanel 2004 te koppelen met de informatie uit de registers. Daarbij kan nagenoeg iedereen met een koppelsleutel uit de OSA-enquête worden voorzien van een leeftijd en geslacht uit de registers. Ook zijn er nauwelijks sprake van gemiste koppelingen bij het uurloon. Dat ligt anders bij het opleidingsniveau. Zoals eerder is gemeld wordt het opleidingsniveau vastgesteld door combinatie van gegevens uit registraties en dertien jaargangen EBB. Omdat de registraties vooral de jongere generaties en het bekostigde onderwijs in Nederland beschrijven, kunnen lang niet alle respondenten uit het OSA-Aanbodpanel van een opleidingsniveau worden voorzien. Van de 3237 werknemers uit de OSA-enquête werken er 2873 12 uur of meer. Hiervan kunnen er slechts 953 van een opleidingsniveau worden voorzien.

Om toch de representativiteit te waarborgen is ervoor gekozen om de data te herwegen naar leeftijdsklasse, geslacht en opleidingsniveau vanuit de OSA-enquête.

## 5. Resultaten

We starten eerst met het geven van de correlaties tussen de negen variabelen (Tabel 1). Daaruit kunnen we al veel afleiden. Leeftijd en geslacht worden vrijwel gelijk gemeten in de enquête en het register: de correlaties zijn nagenoeg 1.00. Het opleidingsniveau in het register meet wel iets anders dan dat in de enquête: de correlatie is slechts .80. Het ln netto uurloon heeft een iets grotere overeenkomst tussen het register en de enquêtes dan het opleidingsniveau. Ook hier geldt echter dat niet precies hetzelfde gemeten wordt: de correlatie is .83.

**Tabel 1.**  
**De correlaties tussen enquête- en registervariabelen**

	leeftijd		geslacht		opleidingsniveau		beroepsniveau		uurloon	
	enq	reg	enq	reg	enq	reg	enq	enq	reg	
leeftijd uit enquête	1.000									
leeftijd uit register	.998	1.000								
geslacht uit enquête	-.097	-.097	1.000							
geslacht uit register	-.095	-.096	.996	1.000						
opleidingsniveau uit enquête	-.072	-.069	.052	.053	1.000					
opleidingsniveau uit register	-.068	-.066	.075	.070	.797	1.000				
beroepsniveau uit enquête	.083	.084	-.062	-.063	.566	.584	1.000			
ln netto uurloon uit enquête	.420	.422	-.121	-.121	.423	.454	.518	1.000		
ln netto uurloon uit register	.477	.479	-.080	-.081	.391	.404	.495	.827	1.000	

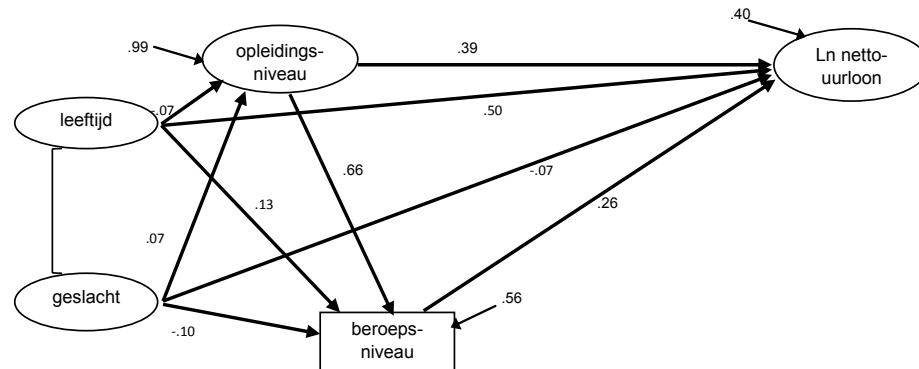
Het opleidingsniveau uit het register hangt hoger samen met het uurloon dan het opleidingsniveau uit de enquête. Daarbij maakt het niet uit of het uurloon uit het register of de enquête wordt genomen. Het opleidingsniveau uit het register hangt ook iets sterker samen met het beroepsniveau dan het opleidingsniveau uit de enquête. De enquêteversie van het uurloon hangt wel sterker samen met beroepsniveau dan de registerversie. De verschillen zijn echter zonder uitzondering vrij klein. Een voorlopige conclusie is dan ook dat de validiteit van de register- en enquêtevariabelen elkaar niet veel ontlopen.

De proef op de som wordt geleverd door de toepassing van het model zoals dat is weergegeven in figuur 5. Met behulp van LISREL 8 wordt de passendheid van dit model berekend (Jöreskog & Sörböm, 1996; Kline, 2005). Het model past met een  $\chi^2$  van 69 bij 18 vrijheidsgraden. Er blijken geen residuele correlaties te zijn van enige omvang, dat wil zeggen dat de correlaties tussen de negen variabelen goed worden gereproduceerd door het model. Ook blijkt daarmee dat er geen sprake is van gecorreleerde meetfouten tussen de variabelen die in het model zijn opgenomen. Dat wil zeggen dat puur het feit dat variabelen uit een enquête of uit een register komt niet leidt tot een grotere samenhang tussen de meetfouten.

Vervolgens is van belang te beoordelen of het model plausibele uitkomsten oplevert (figuur 7). Dat doet het. Het opleidingsniveau heeft een sterk positief effect op het beroepsniveau (.66) en een redelijk sterk direct effect op het uurloon (.39) los van het beroepsniveau. Daarnaast heeft opleidingsniveau nog een indirect effect op het uurloon via het beroepsniveau (.66 x .26 = .17). Het beroepsniveau heeft een positief

effect op het uurloon, maar de correlatie tussen beroepsniveau en loon wordt voor een deel verklaard door een gemeenschappelijke samenhang met het opleidingsniveau.

**Figuur 7.**  
**Schatting van de validiteit van de registervariabelen leeftijd, geslacht, opleidingsniveau en ln netto-uurloon**



Leeftijd heeft een sterk positief effect (.50) op het loon, los van het opleidings- en beroepsniveau en een eveneens positief effect op het beroepsniveau. Het effect is ook erg groot, groter dan verwacht en groter dan de correlaties tussen de metingen voor leeftijd en uurloon doen vermoeden. Blijkbaar is de correctie voor meetfouten die ervoor zorgt dat het effect van leeftijd wordt versterkt. Aangezien de meetfouten van leeftijd te verwaarlozen zijn, wordt dit veroorzaakt door de correctie voor meetfouten in het uurloon. Om dit te onderzoeken is het model toegepast op alleen op de OSA-data. Het effect van leeftijd is dan .40, hetgeen redelijk in de buurt ligt van wat normaal wordt gevonden. Geslacht heeft kleine effecten op opleidingsniveau, beroepsniveau en uurloon. Vrouwen met een baan van meer dan 12 uur zijn hoger opgeleid dan hun mannelijke collega's, hebben een lager beroepsniveau en verdienen minder.

Uiteindelijk beoordelen we de omvang van de meetfouten (tabel 2). De meetfouten in leeftijd en geslacht zijn nihil en niet significant. De meetfouten in het opleidingsniveau en uurloon zijn dat wel. Het opleidingsniveau in de enquête heeft een grotere meetfout dan dezelfde variabele uit het register. Voor het uurloon is het verschil in meetfout gering.

**Tabel 2.**  
**De meetfouten in de enquête- en registervariabelen**

	enquête	register
leeftijd	.00	.00
geslacht	-.01	.01
opleidingsniveau	.24 **	.17 **
beroepsniveau	----	---
ln netto-uurloon	.18 **	.17 **

Significant  $p < .01$

## 6. Conclusies en discussie

Ondanks de toename van het gebruik van registratieve gegevens in sociaalwetenschappelijk onderzoek, is nog maar weinig bekend over mogelijke kwaliteitsproblemen ervan. In dit artikel wordt een methode gepresenteerd waarmee de validiteit van registervariabelen vastgesteld kan worden. Met behulp van de klassieke testtheorie en structurele modellen kan de validiteit van de registervariabelen worden vergeleken met enquêtevariabelen. Tevens kan de meetfout van de registervariabelen onder enkele aannames worden gekwantificeerd.

Dit wordt toegepast op een loonverwervingsmodel waarin leeftijd, geslacht, opleidingsniveau, beroepsniveau en uurloon zijn opgenomen. Met uitzondering van het beroepsniveau beschikken we over een meting uit een enquête en een uit een register. Uit de resultaten blijkt dat de metingen in enquêtes en registers elkaar weinig ontlopen in de omvang van de meetfouten. De uitzondering wordt gevormd door de variabele opleidingsniveau, die beter in het register is gemeten dan in de enquête.

Enkele kanttekeningen zijn op zijn plaats. Ten eerste betreft dit een eerste onderzoek waarin deze methode wordt toegepast. Het zou te vroeg zijn om de algemene conclusie te trekken dat registervariabelen meer valide worden gemeten dan enquêtevariabelen. Dit onderzoek zal meerdere keren moeten worden herhaald om tot meer generieke uitspraken te komen. Bovendien zal het beeld naar alle waarschijnlijkheid gemengd zijn. Sommige variabelen kun je heel goed meten in enquêtes met vraagstellingen die hun validiteit hebben bewezen, en sommige variabelen zijn veel lastiger in enquêtes te meten. Ook voor administratieve data geldt dat de ene variabele beter het concept meet dan andere, onder meer omdat het inhoudelijke concept dichter ligt bij het administratieve concept.

Ten tweede is de meting van het opleidingsniveau aan de registerkant hybride: het betreft registermetingen aangevuld met enquêtegegevens. Dit nunanceert het beeld nog meer: de beste meting voor opleidingsniveau is een combinatie van registers en enquêtes. Het laat ook een nadeel van registers zien: soms ontbreekt een variabele geheel of gedeeltelijk en dan moet je wel enquêteren om de variabele (bij) te schatten.

Ten derde is het alleen toegestaan om de meetfouten te interpreteren als een maat voor de validiteit als met de latente variabelen inderdaad de ware scores worden gemeten. Dat wordt beoordeeld op grond van de parameters in het model waarmee de samenhangen tussen de latente variabelen worden beschreven. In ons voorbeeld is veel kennis over deze samenhangen, zowel van de theorieën als uit empirisch onderzoek. Dat is echter lang niet altijd het geval. De bruikbaarheid van de besproken methode wordt vergroot als deze kennis groter is.

## Referenties

- Arts, C.H. en E.M.J. Hoogteijling (2002). Het Sociaal Statistisch Bestand 1998 en 1999. *Sociaal Economische Maandstatistiek 2002* (12), 66-71.
- Bakker, B.F.M. (2008). De stand van het Sociaal Statistisch Bestand. *Bevolkingstrends*, 56 (2), 14-18.
- Bakker, B.F.M. (2009a). *Micro-integratie* (Den Haag/ Heerlen: CBS).
- Bakker, B.F.M. (2009b). *Trek alle registers open!* (Amsterdam: Vrije Universiteit).
- Bakker, B.F.M., A.M. Bouman en L. van Toor (2006). Opleidingsniveau uit registers: nieuwe bronnen maar nog niet compleet, L. Engberts, F. Linder en F. Bastiaans (red.), *Sociale Samenhang in Beeld, het SSB nu en straks* (Voorburg/Heerlen: CBS), 141-162.
- Bakker, B.F.M., F. Linder en D. van Roon (2008). Could that be true? Methodological issues when deriving educational attainment from administrative datasources and surveys (*Shanghai: Paper prepared for the IAOS Conference on Reshaping Official Statistics, 14-16 October 2008*).
- Bakker, B.F.M. en L. Kuijvenhoven (2010). Registers en sociaalwetenschappelijk onderzoek: een geslaagde combinatie? B.F.M. Bakker en L. Kuijvenhoven, *Registers in sociaalwetenschappelijk onderzoek. Mogelijkheden en valkuilen* (Den Haag / Heerlen: CBS).
- Bethlehem, J. (2009). *Applied survey methods. A statistical perspective* (Hoboken New Jersey: Wiley).
- Borghans, L. (2007). Anders denken, *Edata&research*, 2 (2), 8.
- CBS (1993), *De Standaard beroepenclassificatie 1992* (Voorburg/Heerlen: CBS).
- Fouarge, D. & R. Grim (2007). *Koppeling van het OSA-Arbeidsaanbodpanel aan administratieve gegevens: verslag en documentatie* (Tilburg: OSA).
- Groves, R.M., F.J. Fowler jr., M.P. Couper, J.M. Lepkowski, E. Singer, en R. Tourangeau (2004). *Survey Methodology* (New York: Wiley Interscience).
- Jöreskog, K., en D. Sörbom (1996). *LISREL: 8. User's reference guide* (Chicago: Scientific Software International).
- Kline, R.B. (2005). *Psychological testing: a practical approach to design and evaluation* (New York: SAGE).
- NIWI (2010). *OSA Arbeidsmarktpanel 2000 voorlopige versie. P1462 Steinmetz Archive documentation set version 1.0* (Amsterdam: NIWI).