Meetequivalentie in internationaal vergelijkend onderzoek[1]

Bart Meuleman[*1], *Eldad Davidov*[2], *Jan Cieciuch*[2,3], *Jaak Billiet*[1] & *Peter Schmidt*[4,5]

Samenvatting

Vergelijkend onderzoek verschaft inzicht in verschillen tussen nationale en culturele contexten en levert een belangrijke bijdrage aan onze sociologische kennis. Een geldige vergelijking vereist echter dat theoretische constructen op equivalente wijze gemeten worden over landen heen. Vooral bij abstracte concepten, zoals waarden, attitudes en opinies, is deze veronderstelling van vergelijkbaarheid verre van evident. Meetequivalentie mag dan ook niet zomaar voor waar aangenomen worden, maar is een assumptie die empirische toetsing vereist.

Dit artikel schetst een overzicht van de literatuur die rond cross-nationale vergelijkbaarheid verschenen is. Na een conceptuele verkenning van het begrip 'meetequivalentie' worden bronnen van en preventieve maatregelen tegen inequivalentie besproken. Vervolgens gaan we in op statistische modellen die toelaten om de assumptie van meetequivalentie empirisch te testen. De aandacht gaat hierbij vooral uit naar de populairste techniek, namelijk multiple groep confirmatorische factoranalyse (MGCFA). Bij wijze van illustratie toetsen we of de ESS-meting van steun voor de welvaartsstaat vergelijkbaar is voor Nederlandse, Vlaamse en Waalse respondenten. Tot slot behandelt dit artikel de vraag hoe best omgegaan kan worden met metingen die niet (volledig) vergelijkbaar zijn.

Kernwoorden

Comparatief onderzoek; operationalisering; vergelijkbaarheid; meetfout

* bart.meuleman@soc.kuleuven.be

1 KU Leuven

2 University of Zurich

3 University of Finance and Management in Warsaw

4 National Research University Higher School of Economics, Moscow

5 University of Giessen

Inleiding

De jongste twee decennia heeft internationaal vergelijkend onderzoek een hoge vlucht genomen. Een belangrijke drijvende kracht achter deze ontwikkeling is de toenemende beschikbaarheid van data afkomstig uit grootschalige internationale surveys, zoals de *European Social Survey* (ESS), de *Statistics on Income and Living Conditions* (SILC) survey, het *Program for the International Assessment of Student Achievement* (PISA) of de *Survey of Health, Ageing and Retirement in Europe* (SHARE). In al deze studies worden theoretische constructen – zoals attitudes ten aanzien van minderheidsgroepen, armoederisico's, leerresultaten en mentale gezondheid – gemeten met als expliciet doel deze over landen heen te vergelijken.

Dergelijke internationale vergelijkingen zijn bijzonder nuttig gebleken voor het ontwikkelen van nieuwe en het testen van bestaande sociologische denkkaders. Comparatief onderzoek verschaft inzicht in verschillen tussen nationale en culturele contexten en legt op die manier aspecten bloot die onder de radar blijven in onderzoek dat zich beperkt tot één enkel land. "Comparative sociology is not a special branch of sociology: It is sociology itself", zo wist Durkheim (1982 [Orig. 1895], p. 157) ons meer dan een eeuw geleden al te vertellen. Voor de sociologie als discipline is het dan ook van cruciaal belang om over geldige en betrouwbare internationale surveydata te beschikken.

Het verzamelen van dergelijke data is geen evidente bezigheid. Uit de uitgebreide surveyliteratuur is geweten dat surveymeting binnen één enkele context een hele resem aan valkuilen bevat (zie bv. Billiet, 1993; Groves *et al.*, 2009; Tourangeau, Rips & Rasinski, 2000). Maar het verzamelen van internationale surveydata brengt obstakels mee die zo mogelijk nog uitdagender zijn (Berry *et al.*, 1992; Harkness, van de Vijver & Mohler, 2003; Harkness, Villar & Edwards, 2010b; van de Vijver & Leung, 1997). Zo verschillen beschikbare steekproefkaders (Häder & Gabler, 2003; Heeringa & O'Muircheartaigh, 2010) en non-responsmechanismen (Billiet, Koch & Philippens, 2007; Couper & De Leeuw, 2003) vaak sterk van land tot land, met onvergelijkbare steekproeven tot gevolg. Internationale surveys bevragen respondenten met verschillende moedertalen en vereisen dus een adequate vertaling van de vragenlijsten (Harkness & Schoua-Glusberg, 1998; Harkness *et al.*, 2010a). Respondenten zijn bovendien gesocialiseerd binnen uiteenlopende economische contexten en culturele achtergronden. Bijgevolg bestaat de kans dat ze bepaalde concepten of survey-items op uiteenlopende wijzen interpreteren. Respondenten uit verschillende landen associëren het woord 'immigrant' bijvoorbeeld met uiteenlopende groepen en herkomstlanden, afhankelijk van de migratiegeschiedenis. En wat we precies onder het begrip 'welvaartsstaat' begrijpen is onlosmakelijk verbonden met het gevoerde sociaal beleid. Het spreekt voor zich dat deze fenomenen de internationale vergelijkbaarheid van data ernstig in gevaar brengen.

Comparatieve surveyonderzoekers zijn vertrouwd met deze uitdagingen (zie bv. Jowell, 1998) en hebben een waaier aan maatregelen ontwikkeld om problemen van vergelijkbaarheid zoveel mogelijk te voorkomen (voor een overzicht, zie: Harkness *et al.*, 2003; Harkness *et al.*, 2010a). Maar zelfs het strikt naleven van de strengste voorschriften tijdens de veldwerkfase biedt geen garantie op vergelijkbare metingen. Er is

nood aan een methodologisch kader om, eens de data verzameld zijn, te testen tot op welke hoogte surveyvragen erin slagen nationale en culturele grenzen te overstijgen.

Deze bijdrage geeft een overzicht van de literatuur die recentelijk verschenen is rond het testen van de internationale en cross-culturele vergelijkbaarheid van surveymetingen. Een eerste deel werkt het centrale begrip in deze literatuur – namelijk meetequivalentie – uit. Vervolgens bespreken we de voornaamste bronnen van *in*equivalentie evenals mogelijke preventieve maatregelen om onvergelijkbaarheid te voorkomen. Een derde deel zoomt in op statistische modellen die toelaten om meetequivalentie empirisch te toetsen. Een vierde deel behandelt de vraag hoe best omgegaan kan worden met metingen die niet (volledig) vergelijkbaar zijn.

Meetequivalentie: een conceptuele verkenning

Meetequivalentie (*measurement equivalence*) – of het synoniem meetinvariantie (*measurement invariance*) – is hét centrale concept in de literatuur rond de internationale vergelijkbaarheid van metingen. Desalniettemin ontbreekt unanimiteit over hoe dit begrip best ingevuld kan worden. In een literatuuroverzicht stelt Johnson (1998) vast er meer dan vijftig definities van meetequivalentie in omloop zijn. De grootste gemene deler van deze definities is een verwijzing naar de vergelijkbaarheid van scores die vastgesteld werden bij verschillende populaties onderzoekseenheden. Meetequivalentie is een kenmerk van een set observaties – bij surveyonderzoek: antwoorden op een vragenlijst. Equivalentie impliceert dat deze observaties in verschillende subgroepen van respondenten eenzelfde theoretisch construct meten (Chen, 2008; Meredith, 1993; Vandenberg & Lance, 2000; van de Vijver & Leung, 1997). Of, zoals Horn en McArdle (1992, p. 117) het gevat stellen: "The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute".

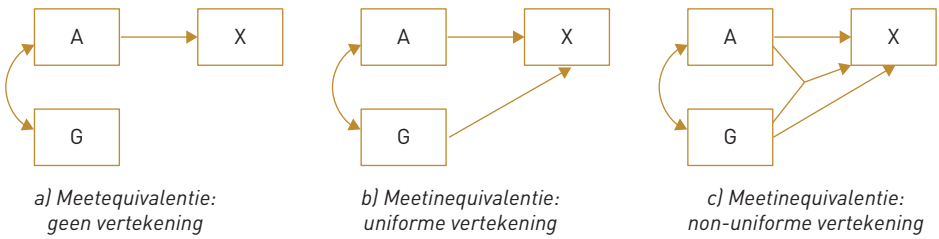
Een meer geformaliseerde definitie vinden we terug bij Mellenbergh (1989), die meetequivalentie ziet als het tegendeel van vertekening met betrekking tot een bepaalde groepsvariabele. Neem meetinstrument X (bijvoorbeeld een IQ-test) dat ontwikkeld werd om latent kenmerk A te meten (in dit geval: intelligentie) in verschillende culturele groepen (groepsvariabele G). De meting is equivalent als en slechts als de verdeling van instrument X, conditioneel op kenmerk A, gelijk is over de groepen G. Deze vereiste kan als volgt geschreven worden:

$$f_1 (X|A = a, G = g) = f_2 (X|A = a, G = h) \text{ voor alle g, h} \in G \quad (1)$$

Volgens deze definitie is de meting equivalent met betrekking tot G wanneer respondenten die dezelfde waarde hebben op kenmerk A maar tot verschillende groepen behoren toch dezelfde score op X krijgen. In het voorbeeld: individuen die even intelligent zijn moeten dezelfde score behalen op de test, onafhankelijk van hun culturele achtergrond. Onder controle van kenmerk A mag groepslidmaatschap met andere

woorden niet gerelateerd zijn met meting *X*. Deze situatie staat in luik a van Figuur 1 afgebeeld. Merk op dat meetequivalentie niet betekent dat de verdeling van meting *X* gelijk moet zijn over de groepen. Equivalentie impliceert daarentegen wel dat groepsverschillen in score *X* enkel en alleen een gevolg mogen zijn van verschillen op het kenmerk dat gemeten wordt, namelijk *A*.

Wanneer verschillen in de verdeling van instrument *X* niet enkel toegeschreven kunnen worden aan kenmerk *A* maar daarentegen ook van *G* afhangen, dan is het instrument vertekend met betrekking tot *G* en bijgevolg niet equivalent. De psychometrische literatuur spreekt in dit verband over *differential item functioning*, waarbij groepslidmaatschap als *violator* optreedt (Welkenhuysen-Gybels, 2003). Dit doet zich bijvoorbeeld voor wanneer intelligentietesten bepaalde culturele groepen systematisch benadelen – een fenomeen dat gedocumenteerd is (Cronshaw *et al.*, 2005). De vertekening kan twee vormen aannemen. Wanneer *G* een rechtstreeks effect op *X* uitoefent, is de vertekening even groot voor alle waarden van *A* en spreken we over uniforme vertekening (zie luik b van Figuur 1). Wanneer de vertekening daarentegen afhankelijk is van kenmerk *A*, is de vertekening niet-uniform. Dit laatste komt neer op een interactie-effect tussen *G* en *A* (zie luik c van Figuur 1; Welkenhuysen-Gybels, 2003). In het voorbeeld van de IQ-test houdt uniforme vertekening in dat de test de intelligentie van alle leden van een bepaalde groep in dezelfde mate onder- of overschat. Indien de 'werkelijke' intelligentie bijvoorbeeld vooral bij groepsleden met een lage intelligentie onderschat wordt, is niet-uniforme vertekening aanwezig.



a) Meetequivalentie: geen vertekening

b) Meetinequivalentie: uniforme vertekening

c) Meetinequivalentie: non-uniforme vertekening

Figuur 1. Grafische voorstelling van meetequivalentie en vertekening.

Deze definitie maakt duidelijk dat een gebrek aan meetequivalentie een ernstige bedreiging vormt zowel voor vergelijkingen over groepen heen als voor conclusies gebaseerd op gepoolde data die meerdere subpopulaties omvatten (Chen, 2008; Poortinga, 1989). Wanneer metingen equivalentie missen, zijn geobserveerde verschillen tussen groepen of landen mogelijks niet meer dan methodologische artefacten. Omgekeerd kan een gebrek aan equivalentie ook landenverschillen maskeren. Kortom, meetequivalentie is een cruciale assumptie in vergelijkend onderzoek en moet dus empirisch getest worden.

Toetsen voor meetequivalentie is meer dan een *methodologische spielerei*. Gebrekkige equivalentie kan onderzoeksresultaten wel degelijk grondig beïnvloeden, zo

illustreren talrijke voorbeelden uit de literatuur. Billiet (2013) toont aan hoe cross-nationale vergelijkingen van religieuze betrokkenheid dramatisch vertekend kunnen worden door meetinequivalentie. Eén van de ESS-indicatoren van religiositeit peilt hoe vaak respondenten religieuze erediensten bijwonen. Deze indicator is ontworpen vanuit een westers, christelijk perspectief en functioneert op een radicaal andere wijze in landen waar moslims de meerderheid van de bevolking uitmaken, zoals Turkije. In de islam is het namelijk niet gebruikelijk dat vrouwelijke gelovigen de reguliere publieke erediensten bijwonen. Wie dit verschil in betekenis over het hoofd ziet, komt verkeerdelijk tot de conclusie dat Turkije het enige land in het ESS is waar vrouwen minder religieus zijn dan mannen.

Ter afsluiting van deze conceptuele verkenning valt op te merken dat de idee van meetequivalentie een erg groot toepassingsgebied heeft. Ten eerste is meetequivalentie niet enkel in internationaal vergelijkend onderzoek een bekommernis. Vergelijkbaarheid is evenzeer van belang wanneer metingen over verschillende tijdspunten of over sociale categorieën vergeleken worden. Interpretaties van items kunnen ten slotte doorheen de tijd wijzigen (Poznyak *et al.*, 2013) of variëren in functie van bijvoorbeeld gender (Van de Velde *et al.*, 2010) of onderwijsniveau (Steinmetz *et al.*, 2009). Ten tweede is de notie van equivalentie relevant voor alle types gemeten kenmerken. Zowel objectieve karakteristieken – zoals demografische variabelen of socio-economische posities (Ganzeboom, de Graaf & Treiman, 1992; Schneider, 2009; Warner & Hoffmeyer-Zlotnik, 2005) – als meer subjectieve en abstracte disposities – zoals opinies, attitudes en waarden (Davidov, Schmidt & Schwartz, 2008; Kankaras & Moors 2009) – dienen op vergelijkbare wijze gemeten te worden. Desalniettemin besteedt de equivalentieliteratuur voornamelijk aandacht aan de cross-nationale vergelijkbaarheid van subjectieve en abstracte concepten. Deze focus is weinig verrassend, aangezien vooral dit type vergelijkingen grote uitdagingen stelt aan equivalentie.

Meet*in*equivalentie: bronnen en preventieve maatregelen

Bronnen van inequivalentie

Metingen kunnen onvergelijkbaar zijn om verschillende redenen. Van de Vijver (1998) ordent de bronnen van inequivalentie in drie grote categorieën. Dit schema vertrekt – net zoals Mellenbergh (1989) – van de notie van vertekening in de betekenis van ongewilde variatie die leidt tot over- of onderschatting van groepsverschillen. Deze vertekening is het tegendeel van meetequivalentie. Van de Vijver (1998) maakt een onderscheid tussen drie types vertekening of *bias* die equivalentie bedreigen, namelijk construct bias, methode bias en item bias. De eerste vorm, construct bias, situeert zich op conceptueel niveau en heeft betrekking op de theoretische geldigheid van concepten. Methode en item bias verwijzen daarentegen eerder naar meetgeldigheid (Meredith & Teresi, 2006).

Construct bias is de meest fundamentele vorm van vertekening en houdt in dat het theoretisch concept zelf een verschillende betekenis heeft over groepen. De scores die in de verschillende populaties gemeten worden, verwijzen niet naar hetzelfde concept. Dit leidt als het ware tot het vergelijken van appelen en citroenen (Horn & McArdle, 1992). Het feit dat bepaalde concepten een cultuur- of landenspecifieke betekenis hebben (zogenoemde 'emic'-concepten – Triandis, 1972) vormt een ware uitdaging in sociologisch onderzoek. Het concept 'welvaartsstaat', bijvoorbeeld, is betekenisloos in een land waar geen staatsinstellingen bestaan die de burgers tot op zekere hoogte economische zekerheid bieden.

Methode bias is het resultaat van verschillen in methodologische aspecten van de studie, zoals procedures van steekproeftrekking (Häder & Gabler, 2003; Heeringa & O'Muircheartaigh, 2010), non-responsmechanismen (Billiet *et al.*, 2007; Couper & De Leeuw, 2003) of gebruikte surveymodus. Ook cross-culturele verschillen in responsstijlen kunnen aanleiding geven tot methode bias. Empirisch onderzoek heeft aangetoond dat sociale wenselijkheidsvertekening (Johnson & van de Vijver, 2003) en volgzaamheidsbias – dit is de neiging om akkoord te gaan met stellingen los van de inhoud ervan (Smith, 2004) – sterker aanwezig zijn in culturele contexten die door een hoge mate van collectivisme worden gekenmerkt (bv. in Aziatische landen). Marin en collega's (Marin, Gamba & Marin, 1992) illustreren dat Amerikanen van Latijns-Amerikaanse origine vaker extreme antwoordcategorieën gebruiken dan blanken, maar ook dat processen van acculturatie ertoe leiden dat de verschillen in responsstijl na verloop van tijd eroderen. De variatie in geobserveerde scores die uit dergelijke methodologische verschillen voortkomt kan makkelijk verward worden met inhoudelijke verschillen en vormt daarom een bedreiging voor de geldigheid van cross-nationale vergelijkingen.

Item bias, ten derde, verwijst naar mogelijke anomalieën op het niveau van de surveyvraag zelf, zoals foutieve vertalingen of het gebruik van woorden die een landen- of cultuurspecifieke betekenis hebben. Een voorbeeld van een vertaalfout in één van de ESS-immigratie-items wordt gedocumenteerd door Billiet (2013, pp. 286-88). Op het eerste gezicht laten de data van 2002 vermoeden dat de Denen van alle Europeanen de meest tolerante houding hebben ten aanzien van immigranten die een misdrijf begaan hebben (Coenders, Lubbers & Scheepers, 2005, p. 98), een bevinding die erg onwaarschijnlijk is gezien het succes van extreem rechts in Denemarken. Testen voor meetequivalentie brachten echter aan het licht dat, omwille van een vertaalfout, het item met betrekking tot het terugsturen van migranten die een misdrijf begaan hebben niet vergelijkbaar is. De Deense vragenlijst gebruikt in plaats van misdrijf een woord dat ook naar veel lichtere vergrijpen verwijst, zoals bijvoorbeeld verkeersovertredingen (nl. *Lovovertrædelse*). Deze zwakkere formulering in Denemarken leidde tot een beduidende onderschatting van het aantal respondenten dat instemt met het terugsturen van migranten. Davidov en collega's (2012) illustreren hoe een item dat refereert aan milieubescherming – vaak gebruikt als indicator van universalistische of postmaterialistische waarden – uiteenlopende connotaties oproept in landen met een verschillende mate van economische ontwikkeling. In economisch minder ontwikkelde

contexten wordt milieubescherming eerder als een zaak van fysieke gezondheid geïnterpreteerd dan als een postmoderne bezorgdheid voor levenskwaliteit. Bijgevolg is dit item een vertekende indicator van universalisme in bepaalde contexten.

Inequivalentie voorkomen

Uit dit conceptueel schema vloeien een aantal strategieën voort die meetinequivalentie te helpen voorkomen tijdens het opstellen van vragenlijsten en van de datacollectie. Deze bespreking beperkt zich tot enkele voorbeelden; Johnson (1998) en van de Vijver (1998) bieden een uitgebreider overzicht van preventieve maatregelen.

Om construct bias te vermijden is het van cruciaal belang een diepgaand inzicht te verwerven in cross-culturele verschillen en gelijkenissen met betrekking tot de bestudeerde fenomenen. Kwalitatieve benaderingen kunnen hierbij erg nuttig zijn. Cognitieve interviewtechnieken (Fitzgerald *et al.*, 2011; Willis, 2005), bijvoorbeeld, laten respondenten luidop reflecteren over de wijze waarop zij bepaalde survey-items interpreteren. Een verwante techniek bestaat erin respondenten onvolledige zinnen te laten vervolledigen, om zo meer inzicht te krijgen in cultuurspecifieke interpretaties van bepaalde concepten. Ook 'lokale' experten kunnen een nuttige bron van informatie zijn en zij kunnen aangeven of bepaalde survey-items geschikt zijn om in hun cultuur de relevante dimensies van een concept te vatten (Johnson, 1998).

Om methode bias te vermijden is het belangrijk om de gebruikte methodologie zoveel mogelijk constant te houden over de te vergelijken groepen heen. Maar het principe van de constante methodologie garandeert geenszins vergelijkbaarheid, aangezien individuen uit verschillende landen uiteenlopende reacties kunnen vertonen op één en dezelfde methodologische stimulus. Het is dan ook raadzaam tijdens de veldwerkfase zoveel mogelijk informatie te verzamelen (bijvoorbeeld over sampling designs, response ratio's, mode-effecten en responsstijlen), zodat hiermee rekening gehouden kan worden tijdens de analyse.

Een adequate vertaling van items is het voornaamste aandachtspunt om item bias te voorkomen. Verschillende vertaalprocedures werden ontwikkeld om zo goed mogelijk te verzekeren dat items dezelfde inhoud weergeven in verschillende talen. *Translation and back-translation*, bijvoorbeeld, houdt in dat één vertaler de vragenlijst eerst van taal A naar taal B omzet, en dat daarna een tweede vertaler de vertaling B terug naar taal A omzet. Vervolgens kunnen de twee versies in taal A met elkaar vergeleken worden. Eventuele verschillen worden verder onderzocht en kunnen tot een aanpassing van de vertaling leiden (Brislin, 1986; Johnson, 1998). Een recentere aanpak die door het ESS aanbevolen wordt, stelt voor om een groep van tweetalige experts bijeen te brengen die gezamenlijk besluiten wat de meest aangewezen vertaling is. Dit groepsmatige proces houdt een vijftal stappen in – namelijk vertaling (*Translation*), *Review*, gezamenlijke beslissing (*Adjudication*), *Pretest* en *Documentatie* van het vertaalproces – en wordt daarom ook wel met het acroniem TRAPD aangeduid (Harkness *et al.*, 2010b).

Meetequivalentie empirisch testen

Latente variabelenmodellen

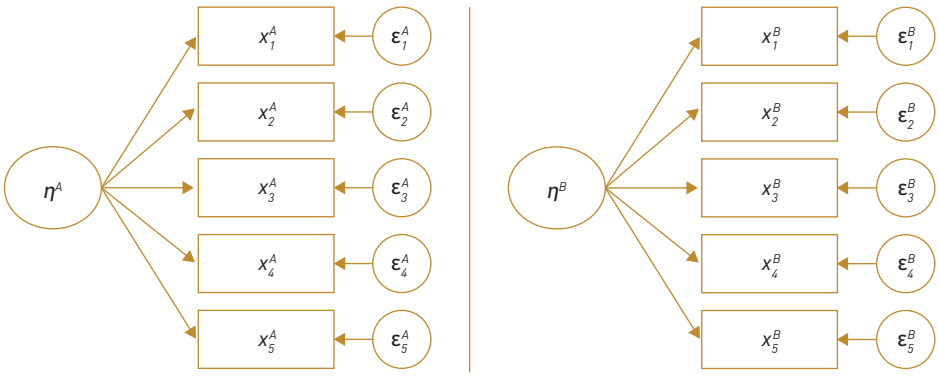
Zelfs de meest rigoureuze toepassing van de preventieve maatregelen kan geen waterdichte garantie bieden dat metingen vergelijkbaar zijn. Het is daarom een aangewezen praktijk om de assumptie van meetequivalentie empirisch te toetsen alvorens tot vergelijkingen over te gaan (Chen, 2008; Vandenberg & Lance, 2000). Tijdens de voorbije decennia is een brede waaier aan data-analytische technieken gebruikt om equivalentie te evalueren (zie Millsap & Meredith, 2007 voor een historisch overzicht). Voorbeelden zijn exploratieve factoranalyse (EFA – Meredith, 1964), multiple groep confirmatorische factoranalyse (MGCFA – Jöreskog, 1971; Steenkamp & Baumgartner, 1998), multidimensional scaling (MDS – Braun & Scott, 1998), item responsetheorie (IRT – Raju, Laffitte & Byrne, 2002) of latente klassenanalyse (LCA – Kankaras, Vermunt & Moors, 2011).

De technieken die momenteel de grootste populariteit genieten – MGCFA, IRT en LCA – hebben een gemeenschappelijk vertrekpunt, namelijk het zogenaamde veralgemeende latente variabelenmodel (*generalized latent variable approach*; Kankaras *et al.*, 2011). Deze aanpak is gebaseerd op de idee dat theoretische concepten vaak niet rechtstreeks geobserveerd kunnen worden (vandaar: latente variabelen). In plaats daarvan worden ze onrechtstreeks afgeleid uit meerdere manifeste indicatoren die de latente variabele reflecteren maar ook meetfouten bevatten. Neem bijvoorbeeld een algemene attitudedimensie zoals 'steun voor overheidsingrijpen', een begrip dat een centrale rol speelt in de literatuur over welvaartsstaatattitudes en solidariteit (Roller, 1995). Aangezien onderzoekers niet direct zintuiglijk kunnen waarnemen of iemand al dan niet voorstander is van een overheid die actief ingrijpt, wordt deze attitude doorgaans geoperationaliseerd door middel van meerdere concrete indicatoren, zoals survey-items (zie verder voor een illustratie aan de hand van ESS-data). Deze indicatoren bieden echter geen perfecte afspiegeling van de latente variabele, maar bevatten ook een unieke component (de meetfout). Figuur 2 geeft een grafische voorstelling van een latente variabelenmodel voor twee groepen (A en B), waarbij ε naar het latente concept verwijst, x_1 tot x_5 de indicatoren vormen en ε_1 tot de ε_5 respectieve meetfouten aanduiden.

De kernidee van deze latente variabelenaanpak is dat we kunnen testen of op equivalente wijze gemeten is in groep A en B door na te gaan in welke mate de relaties tussen latente variabele en multiple indicatoren over groepen heen gelijk zijn. Meer gelijke relaties (in termen van de parameters van het meetmodel) worden geïnterpreteerd als empirische steun voor de assumptie van meetequivalentie (Drasgow & Kanfer, 1985).

De verschillende benaderingen van meetequivalentie – MGCFA, IRT en LCA – maken gebruik van dit basisprincipe en vertonen dus heel wat gelijkenissen. Maar elke techniek voorziet een eigen statistische operationalisering van het latente variabelenmodel. Drie kenmerken zijn van belang om de verschillen tussen de statistische modellen te vatten, namelijk (1) het meetniveau van de geobserveerde variabelen, (2) het

meetniveau van de latente variabele, en (3) het karakter van de statistische functie die de latente variabele met de indicatoren verbindt (de zgn. linkfunctie) (Kankaras *et al.*, 2011). Tabel 1 vat de verschillen tussen de diverse aanpakken compact samen. Het MGCFA-model schat een continue latente variabele (factor) op basis van indicatoren die verondersteld worden eveneens continu te zijn. In de praktijk is het meetniveau van survey-items vaak ordinaal-categorisch (denk bv. aan een Likert-schaal). Voor dit soort data werd een aanpassing van het klassieke MGCFA-model ontwikkeld, die rekening houdt met de categorische aard van de indicatoren (Jöreskog, 1990; Millsap & Yun-Tein, 2004). In beide MGCFA-varianten zijn indicatoren en latente variabele lineair verbonden door middel van de identiteitsfunctie. Het IRT-model linkt (evenals categorische MGCFA) categorische indicatoren aan een continue factor. Het voornaamste verschil is echter dat IRT gebruik maakt van een logistische functie. Bij latente klassenanalyse, ten slotte, wordt een categorische latente variabele geschat (klassen in plaats van een continue factor dus). Ook hier is de statistische onderbouw logistisch van aard.



Figuur 2. Latente variabelen model voor twee groepen (A en B).

Tabel 1. Classificatie van de MGCFA-, IRT- en LCA-benadering van meetequivalentie.

		Meetniveau latente variabelen	
		Continu	*Categorisch*
Meetniveau indicatoren	*Continu*	**MGCFA** (link: identiteitsfunctie)	
	Categorisch	**Categorische MGCFA** (link: identiteitsfunctie)	**LCA** (link: logistische functie)
		IRT (link: logistische functie)	

Meetequivalentie testen met MGCFA

In de empirische literatuur rond meetequivalentie is MGCFA zonder meer de meest prominent aanwezige benadering. Omdat dit overzichtsartikel niet toelaat elk van de technieken meer in detail uit te werken, beperken we ons tot deze aanpak.

De klassieke MGCFA-aanpak modelleert de geobserveerde scores op continue items x_i voor personen in populatie g als een lineaire regressiefunctie van een continue latente variabele η[2]:

$$x_i^g = \tau_i^g + \lambda_i^g \eta^g + \varepsilon_i^g \qquad (2)$$

In deze uitdrukking capteert λ_i^g (de factorlading) de sterkte van de relatie die in groep g tussen latente variabele η en indicator x_i bestaat. τ_i^g is het intercept van de functie en verwijst naar de verwachte score op item x_i voor personen uit groep g die score 0 hebben op de latent variabele. De foutentermen (ε_i^g) capteren de afwijking tussen observatie en voorspelling, en worden verondersteld een multivariate normaalverdeling te volgen.

De MGCFA-benadering toetst equivalentie door na te gaan in hoeverre het meetmodel en zijn respectieve parameters gelijk zijn over groepen g heen. Concreet gebeurt dit door restricties op de meetparameters aan te brengen, en meer en minder restrictieve modellen met elkaar te vergelijken (Vandenberg & Lance, 2000; Steenkamp & Baumgartner, 1998).

De literatuur onderscheidt verschillende niveaus van meetequivalentie, die hiërarchisch geordend kunnen worden. Hogere niveaus vereisen striktere gelijkheid van meetparameters en zijn bijgevolg moeilijker te verkrijgen, maar garanderen ook een grotere mate van vergelijkbaarheid over groepen heen. Configurele equivalentie (*configural equivalence*) is het eerste en laagste niveau van meetequivalentie. Configurele equivalentie vereist dat de structuur van de factor modellen gelijk is over groepen heen. Wanneer een item op een latente factor laadt in één groep, moet dit ook in de andere groepen het geval te zijn. Merk op dat enkel het patroon van substantiële en niet-substantiële ladingen gelijk moet zijn over groepen, niet de exacte sterkte van de factorlading (Steenkamp & Baumgartner, 1998). Configurele equivalentie is een eerder kwalitatieve vorm van meetequivalentie. Ze impliceert dat het gemeten concept in de verschillende groepen dezelfde betekenis draagt of, met andere woorden, dat construct bias afwezig is (van de Vijver, 1998). Configurele equivalentie betekent echter niet dat de transformatie van het kenmerk naar scores (de meetschaal) op identieke wijze verloopt, en garandeert dan ook geen kwantitatieve vergelijkbaarheid over groepen heen. Om de metafoor van de weegschalen te gebruiken: in twee groepen wordt weliswaar één en hetzelfde concept gemeten (namelijk gewicht) maar het is lang niet zeker of de weegschaal in beide groepen op dezelfde wijze geijkt is.

Kwantitatieve vergelijkingen vereisen hogere niveaus van meetequivalentie. Metrische equivalentie (*metric equivalence*) – ook wel zwakke equivalentie (*weak equivalence*; Meredith, 1993) genoemd – is een tweede en hoger niveau, en vereist dat factor-

ladingen (λ_i^g) gelijk zijn over groepen heen. In elke groep moeten de indicatoren even sterk samenhangen met het te meten concept. Een verschuiving van één eenheid in de latente variabele moet, ongeacht lidmaatschap van groepen, een identieke impact hebben op de geobserveerde indicatoren (Steenkamp & Baumgartner, 1998). Metrische equivalentie betekent dat er geen non-uniforme bias aanwezig is. Concreet garandeert metrische equivalentie dat de intervallen op de meetschaal van de latente variabele gelijk zijn: een toename van één eenheid op het gemeten concept heeft dezelfde betekenis in alle onderzochte groepen. Bijgevolg kunnen alle maten die gebaseerd zijn op verschilscores – dit zijn scores die gecorrigeerd zijn voor het groepsgemiddelde, zoals regressiecoëfficiënten of covarianties – op betekenisvolle wijze vergeleken worden (Steenkamp & Baumgartner 1998).

Maar ook metrische equivalentie is nog geen voldoende voorwaarde om ruwe scores of gemiddelden te kunnen vergelijken over groepen heen. Verschillen tussen groepen kunnen immers nog steeds op uniforme wijze over- of onderschat worden. Om ook uniforme bias uit te sluiten is scalaire equivalentie (*scalar equivalence*) – of sterke equivalentie (*strong equivalence*; Meredith, 1993) – vereist. Dit derde niveau van equivalentie gaat bijkomend na of de intercepten (τ_i^g) van het meetmodel invariant zijn. Deze intercepten stellen de verwachte waarde op de geobserveerde indicatoren voor de groep respondenten die waarde 0 heeft op de latente variabele. Gelijkheid van intercepten betekent dat respondenten uit verschillende groepen, conditioneel op de latente variabele, even hoog scoren op de indicatoren. Pas wanneer het meetmodel aan deze voorwaarde voldoet, zijn betekenisvolle groepsvergelijkingen van de scores op de latente variabele mogelijk (Steenkamp & Baumgartner, 1998).

Scalaire equivalentie is niet het hoogste equivalentieniveau. Het is bijvoorbeeld ook mogelijk om te testen of de variantie van foutentermen ε_i^g gelijk is over groepen, wat impliceert dat het meetinstrument evenveel toevalsfout (ruis) bevat in verschillende groepen. Ook een test van gelijkheid van de variantie van de latente variabelen – een voorwaarde om bijvoorbeeld gestandaardiseerde coëfficiënten te vergelijken – behoort tot de mogelijkheden (Vandeberg & Lance, 2000; Steenkamp & Baumgartner, 1998). Omdat deze equivalentieniveaus minder praktische toepassingen hebben, gaat deze bijdrage er niet dieper op in.

Ten slotte dient opgemerkt te worden dat verschillende auteurs (Byrne, Shavelson & Muthen, 1989; Steenkamp & Baumgartner 1998) geargumenteerd hebben dat het niet noodzakelijk is dat de parameters (ladingen en intercepten) voor alle indicatoren invariant zijn. Valide vergelijkingen zouden ook gerechtvaardigd zijn indien slechts een deel van de indicatoren op equivalente wijze functioneert. Dan wordt van partiële equivalentie gesproken. Concreet is vergelijkbaarheid gegarandeerd zodra ten minste twee indicatoren per concept gelijke meetparameters hebben.

Illustratie: steun voor overheidsingrijpen in Vlaanderen, Wallonië en Nederland

Om de diverse meetniveaus te illustreren, stellen we hier een toepassing voor van het MGCFA-model om meetequivalentie te testen. Deze illustratie schuift eveneens een praktische strategie naar voor om de gelijkheid van meetparameters empirisch te toetsen.

Concreet gaan we de vergelijkbaarheid na van de schaal die in ronde 4 van het ESS werd opgenomen om het concept 'steun voor overheidsingrijpen' (Roller, 1995) te meten. Deze analyse focust op vijf items die verwijzen naar concrete domeinen van sociaal beleid, namelijk pensioenen, werkloosheidsuitkeringen, gezondheidszorg, kinderopvang en familiaal verlof. Respondenten dienen op een schaal van 0 tot 10 aan te geven in hoeverre de overheid verantwoordelijkheid moet hebben over deze domeinen (zie Tabel 2).[3]

Tabel 2. Vraagverwoording van de ESS-items m.b.t. steun voor overheidsingrijpen.

Er bestaan verschillende opvattingen over wat wel en niet onder de verantwoordelijkheid van de overheid zou moeten vallen. Geef voor elk van de taken die ik voorlees op een schaal van 0 tot 10 aan in welke mate u denkt dat de overheid ervoor de verantwoordelijkheid zou moeten hebben?		
d16	...ervoor te zorgen dat de gezondheidszorg toereikend is?	0 (Zou helemaal niet de verantwoordelijkheid van de overheid mogen zijn) - 10 (Zou volledig de verantwoordelijkheid van de overheid moeten zijn)
d17	...ervoor te zorgen dat ouderen een redelijke levensstandaard hebben?	
d18	...om te voorzien in een redelijke levensstandaard voor werklozen?	
d19	...om te voorzien in voldoende kinderopvang voor werkende ouders?	
d20	...om te voorzien in betaald verlof voor werkenden die tijdelijk voor zieke familieleden moeten zorgen?	

Deze analyse gaat na of dit instrument vergelijkbare metingen oplevert in drie groepen, namelijk respondenten uit Nederland (N = 1778), Vlaanderen (N = 1060) en Wallonië (N = 575).[4] Om diverse redenen bestaat het risico dat respondenten er per regio uiteenlopende interpretaties van de items op na houden. Naast de taalkloof (Nederlandstalig vs. Franstalig) zijn er immers ook belangrijke institutionele (Nederlandse vs. Belgische welvaartsstaat), economische (bv. werkloosheidscijfers), culturele en politieke verschillen.

Tabel 3. Fit indices voor de geteste MGCFA-modellen.

Model		chi²	df	RMSEA	CFI	TLI
M1	Configurele equivalentie	33.8	9	0.049	0.99	0.98
M2	Metrische equivalentie	55.5	17	0.043	0.99	0.98
M3	Scalaire equivalentie	433.5	25	0.120	0.88	0.86
M3b	Partieel scalaire equivalentie	141.0	22	0.069	0.97	0.95

Om vergelijkbaarheid van het instrument na te gaan, schatten we een reeks CFA-modellen met één latente variabele en vijf indicatoren in drie verschillende groepen. Hierbij passen we een bottom-up strategie (Vandenberg & Lance, 2000) toe: er wordt gestart met het minst restrictieve model (configurele equivalentie) en gaandeweg worden bijkomende pa-

rameters gelijkgesteld over groepen heen (eerst factorladingen – metrische equivalentie; vervolgens intercepten – scalaire equivalentie). De evaluatie of de geobserveerde data de toepaste restricties wel degelijk toelaten gebeurt door verschillende fit indices (chi^2, RMSEA, CFI en TLI – Hu & Bentler, 1999) over modellen heen te vergelijken (Chen, 2007). Wanneer het model niet voldoende bij de data past, inspecteren we de zogenaamde modificatie-indices om na te gaan voor welke meetparameters gelijkheid over groepen niet houdbaar is. Deze modificatie-indices drukken uit met hoeveel eenheden de chi^2-waarde zou dalen indien welbepaalde restricties weggelaten worden. Een meer uitgebreide uiteenzetting over deze procedure kan teruggevonden worden bij Meuleman en Billiet (2012).

Zoals gezegd, test het configurele equivalentie-model (M1 in Tabel 3) geen strikte gelijkheid van parameters, maar wordt nagegaan of een identieke, goed passende factorstructuur gevonden kan worden in de drie groepen. Op het initiële model (vijf items die laden op één enkele factor) dienden enkele aanpassingen te worden gemaakt. Met name bleek het nodig in elk van de groepen twee covarianties tussen de foutentermen toe te laten. Het betreft de covariantie tussen d16 (gezondheidszorg) en d17 (pensioenen) enerzijds, en tussen d19 (kinderopvang) en d20 (familiaal verlof) anderzijds. Deze foutencovarianties zijn theoretisch te verantwoorden. In het eerste geval gaat het om twee beleidsdomeinen gericht op groepen (ouderen en zieken) die een hoge mate van hulpwaardigheid hebben in de ogen van de bevolking omdat de betreffende risico's (ouderdom en ziekte) iedereen treffen in het leven. In het tweede geval betreft het twee beleidsdomeinen die rechtstreeks met zorg voor familieleden te maken hebben. Na deze aanpassingen vertoont het configurele model een vrij goede fit. De RMSEA-waarde ligt beduidend onder de drempelwaarde en zowel CFI als TLI liggen voldoende dicht bij 1. Aangezien een identieke factorstructuur voor de drie regio's past, kunnen we besluiten dat het concept 'steun voor overheidsingrijpen' dezelfde betekenis heeft in Nederland, Vlaanderen en Wallonië.

Vergelijkbaarheid van scores vereist echter hogere meetniveaus. Een tweede model stelt factorladingen gelijk over de drie groepen heen (M2 – metrische equivalentie). Deze bijkomende restricties leiden tot een relatief kleine toename in de chi^2-waarde (27.7 voor 8 bijkomende vrijheidsgraden). CFI en TLI blijven gelijk en de RMSEA vertoont zelfs een lichte daling. Kortom, alle fit indices bevestigen dat de factorladingen wel degelijk gelijk zijn voor de drie groepen en leveren zo empirische steun voor metrische equivalentie. Bijgevolg is het toegelaten bijvoorbeeld regressiecoëfficiënten tussen steun voor overheidsingrijpen en een tweede variabele over groepen te vergelijken (uiteraard op voorwaarde dat deze tweede variabele eveneens op voldoende vergelijkbare wijze gemeten is).

Om scalaire equivalentie te testen stellen we bijkomend de intercepten aan elkaar gelijk (M3). In vergelijking met het metrische equivalentiemodel gaat de fit er echter drastisch op achteruit. De chi^2-waarde neemt toe met maar liefst 378 eenheden (opnieuw voor 8 vrijheidsgraden), en zowel RMSEA (0.120), CFI (0.88) en TLI (0.86) wijzen er duidelijk op dat het model onvoldoende bij de data aansluit. De conclusie luidt dan ook dat niet alle intercepten gelijk zijn voor de Nederlandse, Vlaamse en Waalse data. Inspectie van de modificatie-indices leert echter dat de slechte fit grotendeels teruggevoerd kan worden op de restricties met betrekking tot drie specifieke intercepten (namelijk d19 en d20 voor Nederland; d18 voor Wallonië). Wanneer we deze drie in-

tercepten vrij schatten in plaats van ze gelijk te stellen over landen (M3b), verdwijnt het leeuwendeel van de misfit. De chi^2-waarde van dit model bedraagt 141.0, wat een daling van bijna 300 eenheden impliceert. RMSEA, TLI en CFI suggereren dat model 3b een aanvaardbare beschrijving geeft van de geobserveerde data. Bovendien zijn de mogelijkheden uitgeput om het model substantieel te verbeteren door bijkomende intercepten vrij te laten. We kunnen dan ook besluiten dat de resterende restricties op de intercepten niet tegengesproken worden door de data. Omdat twee items (d16 en d17) gelijke ladingen en intercepten hebben voor de drie groepen, kunnen we besluiten dat de meting aan het criterium van partiële scalaire equivalentie voldoet. Bijgevolg kan de gemiddelde steun voor overheidsingrijpen over de drie regio's op betekenisvolle wijze vergeleken worden. De parameterschattingen voor model 3b (weergegeven in Tabel 4) tonen aan dat het gemiddelde op de latente variabele in Vlaanderen (0.02) en Wallonië (0.12) weliswaar iets hoger ligt dan in Nederland (0 – referentiegroep), maar dat de verschillen te klein zijn om tot significantie te besluiten. Inwoners uit de drie regio's zijn de idee van overheidsingrijpen dus in dezelfde mate genegen.

Tabel 4. Parameterschattingen voor model 3b (gestandaardiseerde parameters tussen haakjes).

	Nederland			Vlaanderen			Wallonië		
	Par.	**Std. Par.**	**Sign.**	**Par.**	**Std. Par.**	**Sign.**	**Par.**	**Std. Par.**	**Sign.**
Factor ladingen									
d16 (gezondheidszorg)	1.00	(0.55)		1.00	(0.57)		1.00	(0.53)	
d17 (ouderen)	1.27	(0.65)	***	1.27	(0.69)	***	1.27	(0.72)	***
d18 (werklozen)	1.22	(0.55)	***	1.22	(0.51)	***	1.22	(0.47)	***
d19 (kinderopvang)	1.15	(0.39)	***	1.15	(0.52)	***	1.15	(0.51)	***
d20 (familiaal verlof)	1.22	(0.47)	***	1.22	(0.58)	***	1.22	(0.51)	***
Intercepten									
d16 (gezondheidszorg)	8.15	(6.09)	***	8.15	(6.06)	***	8.15	(5.26)	***
d17 (ouderen)	7.80	(5.43)	***	7.80	(5.47)	***	7.80	(5.37)	***
d18 (werklozen)	6.28	(3.85)	***	6.28	(3.39)	***	5.64	(2.66)	***
d19 (kinderopvang)	6.29	(2.93)	***	7.26	(4.32)	***	7.26	(3.91)	***
d20 (familiaal verlof)	6.71	(3.53)	***	7.34	(4.53)	***	7.34	(3.74)	***
Foutentermen									
d16 (gezondheidszorg)	1.26	(0.70)	***	1.22	(0.67)	***	1.72	(1.72)	***
d17 (ouderen)	1.20	(0.58)	***	1.08	(0.53)	***	1.01	(1.01)	***
d18 (werklozen)	1.87	(0.70)	***	2.55	(0.74)	***	3.50	(3.50)	***
d19 (kinderopvang)	3.92	(0.85)	***	2.05	(0.73)	***	2.55	(2.55)	***
d20 (familiaal verlof)	2.82	(0.78)	***	1.75	(0.67)	***	2.85	(2.85)	***
Latente variabele									
Gemiddelde	0.00	(0.00)		0.02	(0.02)		0.10	(0.12)	
Variantie	0.53	(1.00)	***	0.59	(1.00)	***	0.68	(1.00)	***

*** p < .0001; ** p < .01; * p < .05

Een blik op de parameterschattingen verschaft bijkomend inzicht in hoe de werking van de indicatoren verschilt tussen Nederland, Vlaanderen en Wallonië. Zo is het Nederlandse intercept voor het item over kinderopvang een volledig punt lager dan in Vlaanderen of Wallonië (6.29 i.p.v. 7.26). Concreet betekent dit dat Nederlanders met gemiddelde 0 op latente variabele 'steun voor overheidsingrijpen' gemiddeld 6.29 scoren op dit item. Vlaamse of Waalse respondenten die eveneens waarde 0 hebben op de latente factor scoren daarentegen gemiddeld 7.26. Onder controle voor de mate van steun voor overheidsingrijpen in het algemeen zijn Belgen dus sterkere voorstanders van publieke kinderopvang dan Nederlanders. Groepsverschillen op dit item kunnen dus niet volledig in termen van het gemeten concept geduid worden, maar wijzen op een regiospecifieke werking van dit item. Het item over kinderopvang is met andere woorden een vertekende indicator van steun voor overheidsingrijpen. Wellicht kan de specificiteit van het Nederlandse zorgregime, waar heel wat vaders en moeder deeltijds werken en gebruik van formele kinderopvang erg laag is (Kremer, 2007; Meuleman & Chung, 2012) verklaren waarom dit item anders functioneert. Een soortgelijk maar minder uitgesproken patroon vinden we terug bij het item over familiale verloven (d20). In Wallonië blijkt het intercept voor het item over werkloosheid dan weer lager dan in Vlaanderen en in België. In Wallonië is steun voor werkloosheidsuitkeringen, onder controle voor de algemene steun voor overheidsingrijpen, dus relatief laag. Ook dit is geen toeval: Wallonië kent een beduidend hoger percentage werklozen, wat de publieke steun voor deze doelgroep kan aantasten (van Oorschot & Meuleman, 2012).

Wat te doen wanneer metingen niet equivalent zijn?

De vorige sectie bood een methodologisch kader om na te gaan in welke mate metingen internationaal vergelijkbaar zijn. In de praktijk wijzen deze testen vaak uit dat instrumenten niet over het beoogde niveau van equivalentie beschikken. In dat geval rijst de vraag hoe best om te gaan met deze inequivalentie. Op het eerste gezicht ligt het voor de hand te besluiten dat de metingen eenvoudigweg niet vergelijkbaar zijn en dat pogingen om te vergelijken beter gestaakt kunnen worden. Deze drastische aanpak kan echter tot het verlies van relevante informatie leiden. In de literatuur zijn dan ook een aantal alternatieve benaderingen terug te vinden, die de aanwezige informatie maximaal benutten (Poortinga, 1989).

Een eerste strategie bestaat erin om de reikwijdte van de vergelijking in te perken en de vergelijking te focussen op landen en concepten die wel vergelijkbaar zijn. Wanneer studies een groot aantal nationale of culturele groepen onderzoeken, bestaat het risico dat de groepen te heterogeen zijn en dat enkel subgroepen van landen onderling vergelijkbaar zijn. Davidov en collega's (2008), bijvoorbeeld, concluderen dat de ESS-meting van de Schwartz-waarden niet scalair equivalent is voor alle landen, maar vinden kleinere groepen van gelijkaardige landen waarvoor metingen vergelijkbaar zijn. Welkenhuysen-Gybels en collega's (2007) introduceren een clustertechniek om

subsets van vergelijkbare landen te identificeren. Het is eveneens mogelijk dat slechts enkele van de gemeten concepten met equivalentieproblemen kampen. In dat geval is het uiteraard mogelijk de concepten die wel equivalent gemeten zijn in de vergelijking te betrekken.

Een tweede benadering is gebaseerd op de idee dat afwijkingen in de meetmodellen slechts problematisch zijn indien ze de inhoudelijke conclusies beïnvloeden. Niet inequivalentie op zich maar wel de mate van vertekening die deze inequivalentie met zich meebrengt is relevant. Oberski (2014) ontwikkelde vanuit deze optiek een methode om in te schatten hoe gevoelig regressiecoëfficiënten en latente gemiddelden (de zogenaamde *parameters of interest*) zijn voor verschillen in meetmodellen. Ook de hierboven vermelde idee van partiële equivalentie past binnen dit denkkader: Indien twee invariante indicatoren voldoende garantie bieden om onvertekende vergelijkingen te maken, dan kan genegeerd worden dat de parameters van de andere indicatoren verschillen.

Een derde – wellicht de meest ambitieuze – aanpak beschouwt verschillen in de meetmodellen niet als problematisch, maar probeert ze daarentegen tot bruikbare bron van informatie over cross-culturele verschillen in te zetten. Verschillen in meetparameters leggen de specifieke natuur van nationale en culturele contexten bloot, en dienen dus inhoudelijk geïnterpreteerd te worden. In de illustratie hierboven, bijvoorbeeld, linkten we verschillen in intercepten aan uiteenlopende zorgculturen en werkloosheidscijfers. Davidov en collega's (2012) zetten nog een stap verder in deze richting. Zij gebruiken een multilevel structureel vergelijkingsmodel om te verklaren waarom bepaalde items verschillend functioneren over landen heen.

Conclusie

Internationale vergelijkingen vormen een belangrijk ingrediënt van het sociologisch onderzoek nu internationale datasets in toenemende mate beschikbaar zijn. Om dergelijke vergelijkingen tot een goed einde te brengen, is het van cruciaal belang dat de gebruikte concepten op een equivalentie wijze gemeten zijn. Inequivalentie kan erg diverse wortels hebben. Theoretische concepten kunnen verschillende betekenis dragen in diverse culturele contexten (construct bias), onderzoeksmethoden kunnen per land verschillend uitpakken (methode bias) of items kunnen bij de respondenten uiteenlopende connotaties oproepen (item bias). Zelfs een strikte implementatie van allerhande preventieve maatregelen tijdens de veldwerkfase zal de vergelijkbaarheid van metingen nooit volledig kunnen garanderen. Meetequivalentie is dan ook een assumptie die niet zonder meer voor waar aangenomen mag worden, maar daarentegen empirische toetsing vereist.

Vandaag hebben comparatieve sociologen keuze uit diverse analysetechnieken om meetequivalentie te evalueren. MGCFA is op relatief korte tijd uitgegroeid tot de populairste techniek. Dit analytisch kader onderscheidt diverse meetniveaus – configurele, metrische en scalaire equivalentie – die elk eigen implicaties hebben voor de vergelijk-

baarheid van metingen. Vooralsnog besteedt slechts een minderheid van comparatieve studies aandacht aan de equivalentie van de metingen. Maar een gestage toename in het aantal toegepaste papers laat vermoeden dat deze technieken meer en meer toegankelijk worden, ook voor de empirische onderzoeker zonder specialisatie in statistische analyse. Dit overzichtsartikel hoopt een steentje bij te dragen aan deze trend.

Noten

1 Deze bijdrage is deels gebaseerd op het review-artikel Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P. & Billiet, J. (2014). Measurement Equivalence in Cross-national Research. *Annual Review of Sociology, 40*.

2 Voor de eenvoud beperken we ons hier tot de situatie met één enkele latente variabele. Dit model kan echter makkelijk uitgebreid worden naar een situatie met meer factoren (Jöreskog, 1971).

3 Hoewel deze items strikt genomen ordinaal en categorisch zijn, behandelt deze analyse ze als metrische variabelen. Simulatiestudies hebben aangetoond dat categorische variabelen als metrisch behandelen geen noemenswaardige vertekening oplevert op voorwaarde dat de steekproeven voldoende groot zijn en de schalen minimaal 5 punten bevatten (DiStefano, 2002).

4 Brussel wordt niet opgenomen in deze vergelijking omdat de steekproefomvang te klein is (N = 124).

Bibliografie

Berry, J. W., Poortinga, Y. H., Segall, M. H. & Dasen, P. R. (Eds.) (1992). *Cross-cultural Psychology. Research and Applications*. Cambridge: University Press.

Billiet, J. (1993). *Ondanks beperkt zicht. Studies over waarden, ontzuiling en politieke verandering in Vlaanderen*. Brussel: VUBPress.

Billiet J. (2013). Quantitative Methods with Survey Data in Comparative Research. In P. Kennett (Ed.), *A Handbook of Comparative Social Policy* (2nd edition) (pp. 264-300). Cheltenham: Edward Elgar.

Billiet, J., Koch, A. & Philippens, M. (2007). Understanding and Improving Response Rates. In R. Jowell, C. Roberts, R. Fitzgerald & G. Eva (Eds.), *Measuring Attitudes Cross-nationally: Lessons from the European Social Survey* (pp. 113-38). London: Sage.

Braun, M. & Scott, J. (1998). Multidimensional Scaling and Equivalence: Is Having a Job the Same as Working? In J. A. Harkness (Ed.), *Zuma-Nachrichten Spezial Volume 3. Cross-Cultural Survey Equivalence* (pp. 129-44). Mannheim: Zuma.

Brislin, R. W. (1986). The Wording and Translation of Research Instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field Methods in Cross-Cultural Research* (pp. 137-64). Beverly Hills, CA: Sage.

Byrne, B. M., Shavelson, R. J. & Muthen, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures – The Issue of Partial Measurement Invariance. *Psychological Bulletin, 105*(3), 456-66.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling, 14*(3), 464-504.

Chen, F. F. (2008). What Happens if we Compare Chopsticks with Forks? The Impact of Making Inappropriate Comparisons in Cross-cultural Research. *Journal of Personality and Social Psychology, 95*(5), 1005-18.

Coenders, M., Lubbers, M. & Scheepers, P. (2005). Majorities' Attitudes towards Minorities in European Societies: Results from the European Social Survey 2002-2003. In M. Coenders, M. Lubbers & P. Scheepers (Eds.), *Majority Population's Attitudes Towards Migrants and Minorities*. Vienna: European Monitoring Centre on Racism and Xenophobia.

Couper, M. P. & De Leeuw, E. D. (2003). Nonresponse in Cross-cultural and Cross-national Surveys. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural Survey Methods* (pp. 157-78). New York: John Wiley.

Cronshaw, S. F., Hamilton, L. K., Onyura, B. R. & Winston, A. S. (2006). Case for Non-Biased Intelligence Testing Against Black Africans Has Not Been Made: A Comment on Rushton, Skuy, and Bons (2004). *International Journal of Selection and Assessment, 14*(3), 278-87.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P. & Meuleman, B. (2012). Using a Multilevel Structural Equation Modeling Approach to Explain Cross-Cultural Measurement Noninvariance. *Journal of Cross-Cultural Psychology, 43*(4), 558-75.

Davidov, E., Schmidt, P. & Schwartz, S. H. (2008). Bringing Values back in. The Adequacy of the European Social Survey to Measure Values in 20 Countries. *Public Opinion Quarterly, 72*(3), 420-45.

DiStefano, C. (2002). The Impact of Categorization With Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(3), 327-46.

Drasgow, F. & Kanfer, R. (1985). Equivalence of Psychological Measurement in Heterogeneous Populations. *Journal of Applied Psychology, 70*(4), 662-80.

Durkheim, E. (1982 [Orig. 1895]). *The Rules of the Sociological Method*. New York: The Free Press.

Fitzgerald, R., Widdop, S., Gray, M. & Collins, D. (2011). Identifying Sources of Error in Cross-national Questionnaires: Application of an Error Source Typology to Cognitive Interview Data. *Journal of Official Statistics, 27*, 569-99.

Ganzeboom, H. B. G., de Graaf, P. & Treiman, D. J. (1992). A Standard International Socio-Economic Index of Occupational Status. *Social Science Research 21*(1), 1-56.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Häder, S. & Gabler, S. (2003). Sampling and Estimation. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural Survey Methods* (pp. 117-36). New York: John Wiley.

Harkness, J .A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., Pennell, B.-E. & Smith, T. W. (2010a). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, NJ: John Wiley & Sons.

Harkness, J. A. & Schoua-Glusberg, A. (1998). Questionnaires in Translation. In J. A. Harkness (Ed.), *Zuma-Nachrichten Spezial Volume 3. Cross-Cultural Survey Equivalence* (pp. 87-126). Mannheim: Zuma.

Harkness, J. A., van de Vijver, F. J. R. & Mohler, P. P. (Eds.) (2003). *Cross-cultural Survey Methods*. New York: John Wiley.

Harkness, J.A., Villar, A. & Edwards, B. (2010b). Translation, Adaptation, and Design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp . 117-40). Hoboken, NJ: John Wiley & Sons.

Heeringa, S. G. & O'Muircheartaigh, C. (2010). Sampling Designs for Cross-cultural and Cross-national Survey Programs. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 251-68). Hoboken, NJ: John Wiley & Sons.

Horn, J. L. & McArdle, J. J. (1992). A Practical and Theoretical Guide to Measurement Invariance in Aging Research. *Experimental Aging Research, 18*(3), 117-44.

Hu, L. & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Johnson, T. (1998). Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research. In J. A. Harkness (Ed.), *Zuma-Nachrichten Spezial Volume 3. Cross-Cultural Survey Equivalence* (pp. 1-40). Mannheim: Zuma.

Johnson, T. P. & van de Vijver, F. (2003). Social Desirability in Cross-cultural Research. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural Survey Methods* (pp. 195-206). New York: John Wiley.

Jöreskog, K. G. (1971). Simultaneous Factor Analysis in Several Populations. *Psychometrika, 36*(4), 409-26.

Jöreskog, K. G. (1990). New Developments in LISREL: Analysis of Ordinal Variables Using Polychoric Correlations and Weighted Least Squares. *Quality and Quantity, 24*(4), 387-404.

Jowell, R. (1998). How Comparative is Comparative Research? *American Behavioral Scientist, 42*(2), 168-77.

Jowell, R., Roberts, C., Fitzgerald, R. & Eva, G. (2007). *Measuring Attitudes Cross-nationally. Lessons from the European Social Survey*. London: Sage.

Kankaras, M. & Moors, G. (2009). Measurement Equivalence in Solidarity Attitudes in Europe. Insights from a Multiple Group Latent Class Factor Approach. *International Sociology, 24*(4), 557-79.

Kankaras, M., Vermunt, J. K. & Moors, G. (2011). Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches. *Sociological Methods & Research,* 40(2), 279-310.

Kremer, M. (2007). *How Welfare States Care. Culture, Gender and Parenting in Europe*. Amsterdam: Amsterdam University Press.

Marin, G., Gamba, R. J. & Marin B. V. (1992). Extreme Response Style and Acquiescence among Hispanics. The Role of Acculturation and Education. *Journal of Cross-cultural Psychology, 23*(4), 498-509.

Mellenbergh, G. J. (1989). Item Bias and Item Response Theory. *International Journal of Educational Research, 13(2)*, 127-43.

Meredith, W. (1964). Rotation to Achieve Factorial Invariance. *Psychometrika 29*(2), 177-85.

Meredith, W. (1993). Measurement Invariance, Analysis and Factorial Invariance. *Psychometrika, 58*(4), 525-43.

Meredith, W. & Teresi, J. A. (2006). An Essay on Measurement and Factorial Invariance. *Medical Care, 44*(1), 69-77.

Meuleman, B. & Billiet, J. (2012). Measuring Attitudes toward Immigration in Europe: The Cross-cultural Validity of the ESS Immigration Scales. *ASK.Research&Methods, 21*, 5-29.

Meuleman, B. & Chung, H. (2012). Who Should Care for the Children? Support for Government Intervention in Childcare. In H. Ervasti, J. Goul Andersen, T. Fridberg & K. Ringdal (Eds.), *The Future of the Welfare State. Social Policy Attitudes and Social Capital in Europe* (pp. 107-33). Cheltenham, UK: Edward Elgar.

Millsap, R. E. & Meredith, W. (2007). Factorial Invariance: Historical Perspectives and New Problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor Analysis at 100. Historical Developments and Future Directions* (pp. 131-52). Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E. & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research*, *39*(3), 479-515.

Oberski, D. L. (2014). Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models. *Political Analysis, 22*(1), 45-60.

Poortinga, Y. H. (1989). Equivalence of Cross-cultural Data: An Overview of Basic Issues. *International Journal of Psychology, 24*(6), 737-56.

Poznyak, D., Meuleman, B., Abts, K., Bishop, G. F. (2013). Trust in American Government: Longitudinal Measurement Equivalence in the ANES, 1964-2008. *Social Indicators Research*. DOI: 10.1007/s11205-013-0441-5.

Raju, N. S., Laffitte, L. J. & Byrne, B. M. (2002). Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory. *Journal of Applied Psychology, 87*(3), 517-29.

Roller, E. (1995). The Welfare State: The Equality Dimension. In O. Borre & E. Scarbrough (Eds.), *The Scope of Government*. New York/Oxford: Oxford University Press.

Schneider, S. (2009). *Confusing Credentials: The Cross-Nationally Comparable Measurement of Educational Attainment*. Oxford: University of Oxford.

Smith, P.B. (2004). Acquiescent Response Bias as an Aspect of Cultural Communication Style. *Journal of Cross-cultural Psychology, 35*(1), 50-61.

Steenkamp, J.- B. E. M. & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-national Consumer Research. *Journal of Consumer Research, 25*(1), 78-90.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S. & Schwartz, S. H. (2009). Testing Measurement Invariance Using Multigroup CFA: Differences between Educational Groups in Human Values Measurements. *Quality & Quantity*, *43*(4), 599-616.

Tourangeau, R., Rips L. J. & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Vandenberg, R. J. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4-69.

Van de Velde, S., Bracke, P., Levecque, K. & Meuleman, B. (2010). Gender Differences in Depression in 25 European Countries after Eliminating Measurement Bias in the CES-D 8. *Social Science Research, 39*(3), 396-404.

Van de Vijver, F. (1998). Towards a Theory of Bias and Equivalence. In J. A. Harkness (Ed.), *Zuma-Nachrichten Spezial Volume 3. Cross-Cultural Survey Equivalence* (pp. 41-65). Mannheim: Zuma.

Van de Vijver, F. J. R. & Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. London: Sage

Van Oorschot, W. & Meuleman, B. (2012). Welfare Performance and Welfare Support. In S. Svallfors (Ed.), *Contested Welfare States: Welfare Opinions in Europe and beyond* (pp. 25-57). Palo Alto: Stanford University Press.

Warner, U. & Hoffmeyer-Zlotnik, J. H. P. (2005). Discussion of the Income Measure in the European Social Survey: A Proposal of Revised Survey Questions About the "Total Net Household Income". In J.A. Harkness (Ed.), *Zuma-Nachrichten Spezial Volume 12. Conducting Cross-National and Cross-Cultural Surveys* (pp. 53-66). Mannheim: Zuma.

Welkenhuysen-Gybels, J. (2003). *The Detection of Differential Item Functioning in Likert Score Items*. Leuven: KU Leuven.

Welkenhuysen-Gybels, J., van de Vijver, F. & Cambré, B. (2007). A Comparison of Methods for the Evaluation of Construct Equivalence in a Multi-group Setting. In G. Loosveldt, M. Swyngedouw & B. Cambre (Eds.), *Measuring Meaningful Data in Social Research* (pp. 357-72). Leuven: Acco.

Willis, G. (2005). *Cognitive Interviewing. A Tool for Improving Questionnaire Design*. London: Sage.

Abstract

It is indisputable that comparative research contributes to sociological knowledge by providing insight in the differences that exist across national and cultural contexts. However, valid cross-national comparisons require theoretical constructs to be measured equivalently across countries. This is especially the case for abstract concepts, such as values, attitudes, or opinions. As such, measurement equivalence cannot be readily assumed; it is a hypothesis that needs to be tested empirically. This article reviews the social science literature on the cross-national comparability of measurements. We start with some conceptual clarifications regarding the central notion in this field, namely 'measurement equivalence'. Possible sources of inequivalence as well as preventive measures are discussed. A subsequent section deals with statistical models to test empirically whether the conditions for measurement equivalence are fulfilled. Most attention is paid to the most popular technique for testing equivalence, namely multiple group confirmatory group analysis (MGCFA). By means of illustration, we test whether the ESS-scale measuring support for the welfare state is comparable across respondents from the Netherlands, Flanders and Wallonia. Finally, we suggest what may be done when equivalence is not supported by the data.

Keywords

Cross-national research; operationalisation; comparability; measurement error