Cluster bias: Testing measurement invariance in multilevel data

Jak, S.

Publication date
2013

SAMENVATTING (SUMMARY IN DUTCH)

In dit proefschrift worden methoden en procedures voorgesteld die gebruikt kunnen worden voor het toetsen van vraagonzuiverheid (measurement bias) in multilevel data.

Stel dat een onderzoeker geïnteresseerd is in de invloed van motivatie op rekenvaardigheid bij kinderen. Na wekenlang scholen te hebben gebeld, vindt ze 200 leraren en 700 leerlingen bereid aan haar onderzoek mee te werken. De leerlingen vullen een motivatie vragenlijst in met 10 items zoals "Ik denk dat leren rekenen goed voor me is" en "Ik vind rekenen leuk", die gescoord worden op een 7-puntsschaal van 1 (helemaal niet mee eens) tot 7 (helemaal mee eens). De kinderen maken ook een rekenvaardigheidstoets met 60 opgaven die goed of fout gemaakt kunnen worden.

Voordat de onderzoeker een hypothese kan toetsen over de relatie tussen motivatie en rekenvaardigheid, wil zij eerst weten: Zijn deze metingen valide? Resulteren verschillen in motivatie en rekenvaardigheid inderdaad in verschillen in de itemresponsen (Borsboom, Mellenbergh & van Heerden, 2004)? En meten de items dezelfde eigenschappen voor verschillende (groepen) respondenten (Mellenbergh, 1989; Meredith, 1993; Oort, 1992, 1993)? Als de rekenopgaven inderdaad hetzelfde meten voor bijvoorbeeld jongens en meisjes, dan zouden jongens en meisjes met gelijke rekenvaardigheid gemiddeld identieke test scores moeten behalen. Als dit het geval is, zijn de metingen meetinvariant ten opzichte van sekse. Als dit niet het geval is spreken we van vraagonzuiverheid. Om een voorbeeld te geven: een redactiesom zou makkelijker op te lossen kunnen zijn voor meisjes, doordat meisjes gemiddeld gezien beter kunnen lezen dan jongens (Wei et al., 2012). In dat geval zullen meisjes meer goede antwoorden op de som geven dan jongens, terwijl hun rekenvaardigheid gelijk is.

In algemenere zin is vraagonzuiverheid gedefinieerd als een schending van meetinvariantie (Mellenbergh, 1989). Stel dat item X (bijvoorbeeld de redactiesom) ontworpen is om de latente trek T (rekenvaardigheid) te meten. X is meetinvariant ten opzichte van een variabele V (bijvoorbeeld sekse) als de conditionele verdeling van X, gegeven T en V, gelijk is aan de conditionele verdeling van X, gegeven T. Met andere woorden, meetinvariantie geldt als alle invloed van de potentiele schender V op X via T loopt. Zie het figuur op pagina 19 van dit proefschrift voor een grafische weergave van vraagonzuiverheid. De twee meest gebruikte typen modellen voor het onderzoeken van meetinvariantie door middel van structural equation modeling (SEM) zijn multigroep modellen (Sörbom, 1974; Horn & McArdle, 1992; Little, 1997; Widaman & Reise, 1997) en restrictieve factoranalyse (RFA; Oort, 1992, 1998) of equivalente MIMIC (Muthén, 1989) modellen.

Een moeilijkheid is dat we geen directe maat hebben van de latente variabele waar we in geïnteresseerd zijn, zoals rekenvaardigheid of motivatie. We moeten werken met de

geobserveerde itemresponsen. De relatie tussen de geobserveerde itemresponsen en motivatie of rekenvaardigheid kan worden weergegeven in een meetmodel, zoals het lineaire factor model (Mellenbergh, 1994; Spearman, 1904, 1928). In het lineaire factor model wordt de variabele waarin we geïnteresseerd zijn weergegeven als een continue latente factor, die alle gedeelde variantie in de geobserveerde itemresponsen verklaart. Ieder item wordt ook beïnvloed door een unieke factor, die weer bestaat uit een structureel deel (dit deel zorgt voor item-specifieke variantie) en een random deel (de meetfout) (Bollen, 1989).

Het onderzoeken van vraagonzuiverheid dient altijd voorafgegaan te worden door het vinden van een correct meetmodel. **Hoofdstuk 1** van dit proefschrift dient als een introductie over vraagonzuiverheid. Door middel van twee voorbeelden uit een cognitieve vaardigheidstest lieten we zien dat vraagonzuiverheid en multidimensionaliteit nauw aan elkaar verbonden zijn. Een item dat onzuiver is, is multidimensioneel, aangezien het een dimensie meet die niet gemeten diende te worden. Als deze dimensie gerelateerd is aan potentiele schenders van meetinvariantie, (dit zijn vaak variabelen zoals sekse, etniciteit en leeftijd), dan zal dit item onzuiver blijken ten opzichte van deze variabele.

Een andere vraag die de onderzoekster uit het voorbeeld kan stellen is: Worden rekenvaardigheid en motivatie zuiver gemeten in verschillende schoolklassen? Aangezien ze data verzameld heeft van kinderen die gegroepeerd zijn in klassen, hebben de data een multilevel structuur. We kunnen in dit voorbeeld twee niveaus ("levels") onderscheiden: het kindniveau noemen we Niveau 1 en het klasniveau noemen we Niveau 2. Met multilevel SEM kunnen we modellen specificeren op verschillende niveaus van de multilevel data. Kinderen die in dezelfde klas zitten delen kenmerken op klasniveau, zoals de leraar, de samenstelling van de klas en de grootte van de klas. Verschillen in deze kenmerken kunnen leiden tot verschillen in de gemiddelde testscores van kinderen uit verschillende klassen, die niet verklaard worden door de factoren rekenvaardigheid of motivatie. In **Hoofdstuk 2** van dit proefschrift stellen we een toets voor die gebruikt kan worden om te toetsen of metingen onzuiver zijn ten opzichte van schoolklas. Deze toets is algemeen geschikt om onzuiverheid ten opzichte van de clusterende variabele in multilevel data te onderzoeken (bijvoorbeeld bij data van mensen in landen, patiënten in ziekenhuizen, kinderen in families, etc.), vandaar de naam "toets voor clusteronzuiverheid". Clusteronzuiverheid kan onderzocht worden door te toetsen of de factor ladingen op Niveau 1 gelijk zijn aan de factor ladingen op Niveau 2, en of de residuele varianties op Niveau 2 nul zijn. De toets wordt geïllustreerd met data uit onderwijskundig onderzoek. Daarnaast laten we in een simulatie onderzoek zien dat met continue data afkomstig van vijf items, en een groot genoeg aantal clusters, de likelihood ratio test genoeg statistische power heeft om clusteronzuiverheid te ontdekken. Met 50 clusters van 25 observaties per cluster is de power voldoende als de onzuiverheid 3% of meer van de totale variantie van de indicator veroorzaakt. Met slechts 20 clusters met ieder 25 observaties is de power om

clusteronzuiverheid te ontdekken voldoende als de onzuiverheid zorgt voor meer dan 5% van de totale variantie. De proporties vals negatieven waren hoger dan het gekozen significantieniveau in de condities met 100 clusters, maar lager in de condities met 20 clusters.

In **Hoofdstuk 3** breiden we de test voor clusteronzuiverheid uit naar ordinale itemresponsen, met behulp van het ordinale twee-niveau factor model (Grilli & Rampichini, 2007). Op basis van een simulatie onderzoek concluderen we dat clusteronzuiverheid in ordinale data getoetst kan worden met de likelihood ratio test en met de Wald test. Beide tests hebben voldoende power om aanzienlijke hoeveelheden onzuiverheid te detecteren, en hebben acceptabele proporties vals negatieve resultaten. De geschaalde likelihood ratio test, zoals geïmplementeerd in het programma Mplus, wordt niet aangeraden voor het toetsen van clusteronzuiverheid, aangezien in alle condities substantiële aantallen ontoelaatbare resultaten werden gevonden. De voorgestelde toetsen worden geïllustreerd met data over leerkracht-leerling relaties.

De oorzaak van clusteronzuiverheid is per definitie een variabele op clusterniveau. In het voorbeeld van kinderen in klassen, kan de oorzaak van clusteronzuiverheid liggen in onzuiverheid ten opzichte van de didactische kwaliteiten van de leraar. Om de toets voor clusteronzuiverheid toe te passen, hoeft de werkelijke schender van meetinvariantie (als die er is) niet gemeten te zijn. De toets voor clusteronzuiverheid kan daarom gebruikt worden als een algemene toets voor onzuiverheid ten opzichte van alle mogelijke variabelen op clusterniveau. In **Hoofdstuk 4** vergelijken we de power en de proporties vals positieven van de toets voor clusteronzuiverheid en de RFA-toets. Zoals verwacht heeft de RFA-toets meer power dan de toets voor clusteronzuiverheid. De toets voor clusteronzuiverheid heeft in het algemeen een kleinere hoeveelheid vals positieve resultaten. We concluderen dat het niet vinden van clusteronzuiverheid niet uitsluit dat er significante onzuiverheid ten opzichte van een Niveau 2-variabele gevonden wordt met de RFA-toets.

In het laatste hoofdstuk van dit proefschrift, **Hoofdstuk 5**, stellen we een stapsgewijze aanpak voor om vraagonzuiverheid te onderzoeken ten opzichte van een schender op Niveau 1, een schender op Niveau 2, en de cluster variabele. In deze procedure is de eerste stap het toetsen of multilevel-analyse werkelijk nodig is, Stap 2 is het vinden van een geschikt meetmodel op Niveau 1, Stap 3 is het toetsen van vraagonzuiverheid op Niveau 1, Stap 4 is het toetsen op clusteronzuiverheid, en in Stap 5 wordt de mogelijkheid getoetst dat de clusteronzuiverheid verklaard wordt door Niveau 2-variabelen. De vijf stappen zijn zo ontworpen dat het uiteindelijke model alle bias en inhoudelijke resultaten op beide niveaus laat zien.

Dit proefschrift biedt onderzoekers een methode om op een uitvoerbare en systematische manier vraagonzuiverheid te toetsen in multilevel data.