BEOORDELING 'VELDINSTRUMENT' ALS OBSERVATIE-INSTRUMENT

Huisman, 31 december 2014

Sardes is gevraagd het nieuwe observatie-instrument voor de toezichthouders, *'veldinstrument observatie kindercentrum'*, te valideren. De inhoud-validatie moet op voorhand als mislukt worden beschouwd omdat de theoretische concepten die zijn gebruikt (de pedagogische doelen uit de wet) niet voldoen aan wetenschappelijke eisen voor definiëring (conceptualisering) en operationalisering van begrippen, bijgevolge waarvan de informationeel betrouwbaarheid en cognitief geldigheid tekortschieten, zodat instrumentele betrouwbaarheid en validiteit onmogelijk wordt. Alleen al op deze grond moet het veldinstrument worden verworpen.
Sardes toetst ook nog de concurrent-validiteit, per abuis in het rapport 'concurrentie-validiteit' genoemd. Concurrent betekent 'gelijktijdig' en concurrent validatie stelt dus de vraag: *geeft het te valideren instrument (in dit geval: het veldinstrument) dezelfde resultaten als een onafhankelijk daarvan gelijktijdig gebruikt ander instrument?*
Een kritische analyse hiervan is goed beschouwd overbodig vanwege de hiervoor reeds gestelde principiële onhaalbaarheid van welke validiteit dan ook. Vanwege de vergaande tekortkomingen in de onderzoekaanpak van Sardes gaan we hier toch nader op in.

Concurrent-validatie eist het onafhankelijk van elkaar gelijktijdig gebruiken van 2 verschillende instrumenten die worden verondersteld (ongeveer) hetzelfde te meten.
En daarmee raken we meteen aan een **eerste** wetenschappelijk tekort van dit onderzoek: er waren weliswaar twee instrumenten maar die zijn niet onafhankelijk van elkaar gelijktijdig gebruikt! Hoe zit dat?
1
Er is gekozen voor paarsgewijze, dus dubbele inspecties op de groep, waarbij de ene inspecteur het veldinstrument invulde en de andere inspecteur het 'schaduw instrument'. Beide inspecteurs kennen elkaar want er deden 11 GGD'en mee, met 2 of meer inspecteurs, die met elkaar 94 dubbele inspecties hebben uitgevoerd. Dat wil zeggen dat meerdere keren dezelfde inspecteurs meerdere inspecties hebben gedaan, zij het in wisselende paren. Het aantal deelnemende inspecteurs is minimaal 22 (11 x 2) maar het maximum aantal weten we niet, want het rapport geeft hierover geen uitsluitsel (dus N is onbekend). Wij vermoeden dat dit geen toeval is. Maar stel er waren in totaal 31 inspecteurs, en daaronder ook eigen collega's, want dat was toegestaan, zo staat in het rapport, dan hebben die elk 6 observatielijsten ingevuld, afwisselend het ene of het andere instrument. En dat betekent dat zowel het veldinstrument als het schaduw instrument herhaaldelijk door dezelfde persoon is ingevuld. Van onafhankelijkheid is dan natuurlijk geen sprake meer.
2.
Die onafhankelijkheid is ook in een ander opzicht dubieus. Tijdens de inspectie was er een rolverdeling waarbij de inspecteur met het veldinstrument het voortouw nam en de andere inspecteur wat op de achtergrond bleef. Dat benadeelt het gelijktijdig waarnemen van feiten, maar suggereert ook een zekere rangschikking. En dat is evenmin 'onafhankelijk van elkaar'.
3.
Het gaat tenslotte om inspecteurs die elkaar kennen, uit dezelfde GGD komen en dus uit dezelfde culturele context waar zij dagelijks werken. Dat bevordert de kans op een gedeelde

bias (zie hoofdstuk 3 van deze analyse), waarbij juist niet meer onafhankelijk van elkaar wordt geoordeeld, zelfs ongeacht het instrument.

Deze drie tekorten zijn strikt genomen al voldoende om de resultaten te verwerpen. Maar er zijn meer gebreken, die noodzakelijk moeten worden genoemd.

Een **tweede** gebrek is de representativiteit. De opzet is select in plaats van a-select (random). Van de aangeschreven 26 GGD'en deden er uiteindelijk maar 11 mee. Dat kan wijzen op overeenkomstige kenmerken van die 11 waardoor zij wèl mee deden. En omdat niet wordt prijs gegeven om welke GGD'en het gaat is het ook zeer de vraag of die wel voldoende verspreid waren over het land. Maar ook de selectie van de opvangsoorten die zijn geïnspecteerd is zeer waarschijnlijk niet a-select. Het is aannemelijk dat inspecteurs in Groningen voor opvang kozen in die omgeving en niet naar Limburg zijn afgereisd omdat daar geen GGD mee deed. En dat betekent dat ze zijn gaan inspecteren in voor hen reeds bekende opvang: 38 kdv's, 24 peuterspeelzaal en 32 bso's. En de inspecteurs zijn wel de enige bron die is gebruikt voor deze validering, met als reden: a) zij zijn al getraind in hun eigen instrument b) voor meer valideerders was geen tijd en budget c) inspecteurs worden de gebruikers d) dus dit geeft meteen meer betrokkenheid en draagvlak. Tja, voor een wetenschappelijke validatie allemaal geen relevante afwegingen. Integendeel.

Ten **derde** zien we een cluster van grote methodisch-technische gebreken in de voorbereidende en controlerende analyse t.b.v het 'opschonen' van data, zodat deze geschikt werden voor beschrijvende en toetsende analyses, aldus het rapport (blz 33).

- Zo zijn lijsten met meerdere scores in plaats van één per item, toch mee genomen. Eigenhandig maakten de onderzoekers dan een keuze uit die scores, waarbij doorgaans de meest extreme score dan werd gekozen, dus de hoogste of de laagste. Ongewis blijft wat dan de afstand was tussen die scores: groot? klein? En ook hoe vaak dit zich voordoet in het bestand. Maar hoe het ook zij, het is een ontoelaatbare manipulatie van de data en de legitimering die het rapport daarvoor geeft overtuigt allerminst: "blijkbaar had zo'n inspecteur die extremen wel gezien" .
- Verder zijn lijsten die niet met een score maar met 'ja' of 'gezien' zijn ingevuld, eveneens mee genomen, en dan door de onderzoekers eigenhandig vertaald naar de 'bijpassende score mogelijkheid'. Het is onnavolgbaar hoe zij 'ja' of 'gezien' hebben vertaald naar '*onvoldoende, minimaal, extra' (veldinstrument)* ofwel naar *'laag, middelmatig, hoog' (schaduw instrument).*
- En als er over dit type scoringsgedrag twijfel bestond, werd de betreffende inspecteur gebeld en om uitleg gevraagd! Echt ongelooflijk, het is bijna amateuristisch: een overleg over een oordeel in een valideringsonderzoek van een instrument.
- De uiteenlopende scoringswijze van beide instrumenten is ook nog 'vergelijkbaar' gemaakt. De totaal andere scoringsmogelijkheden, naar inhoud en systematiek, laat zich echter niet goed voegen. Doe je dat toch dan dwingt dit tot het kiezen van het laagste data-niveau, dus in plaats van ordinaal bij het schaduwinstrument wordt alles dan nominaal (van het veldinstrument). En naar inhoud lukt dat niet zonder de ruwe scores geweld aan te doen.

En dat vormt een **vierde** ernstige fout in het rapport: de onderzoekers hebben toch een ordinaal niveau aangehouden! En dus statistische maten toegepast die wel passen bij

ordinaal maar niet geschikt zijn voor nominaal niveau, zoals Spearman's r, de rangcorrelatie coëfficiënt waarmee correlaties worden berekend. Daar komt bij dat ook de N (= aantal respondenten) in dit onderzoek zeer dubieus is. De N moet minimaal 30 zijn om Spearman's r te mogen berekenen, en er zijn weliswaar 2x 94 lijsten ingevuld (dus 188), maar niet door 188 inspecteurs! De conclusie is onontkoombaar: alle gevonden correlaties zijn goed voor de prullenbak. Dat is niet echt heel erg, want het merendeel van gepresenteerde correlaties is zwak, al noemen de onderzoekers ze 'hoog'! Daar komen we straks op terug.

Ten behoeve van de statistische bewerkingen zijn eveneens vooraf keuzes gemaakt, zoals:

- In het schaduw instrument zijn alle 'nieten' en 'wellen' met elkaar in verband gebracht onder de voorwaarde van tenminste 1 'niet' of 1 'wel' per item, als aanwijzing dat er tijdens de observatie wel op is gelet. Ook hier weer ontoelaatbaar gegoochel. Een heel item wordt mee genomen ook al is er maar 1 x op gescoord, met als excuus: dat item is dus heus wel geobserveerd, jammer dat maar 1 score staat. De 'somscore' van zo'n item is dan die ene score! Ook hier wordt niet vermeld hoe vaak dit voorkomt.
- Een vergelijkbare keuze is gemaakt voor het berekenen van de gemiddelden per item en per indicator. Wat blijkt: bij meer indicatoren is niet van alle items, en van meerdere items niet alle onderliggende praktijkvoorbeelden een score gegeven. Er worden dus veel scores gemist! En daarom besloten de onderzoekers om een minimum aantal scores vast te stellen waarmee ze dan toch een gemiddelde (M) gingen uitrekenen, nl: als van minimaal de helft van de voorbeelden/ items/indicatoren een score was ingevuld. Kortom: met 50% data wordt 100% M gemaakt.
- Naast de gammele M-somscores per indicator zijn tevens evenzeer gammele M-totaalscores (van beide instrumenten tezamen) uitgerekend (hoe?) en daarover is Spearman's rangorde correlatie coëfficiënt berekend.

Nog één keer Spearman (Rs) als correlatiecoëfficiënt.
De wetenschappelijke state of the art stelt de volgende eisen aan de waardering van de gevonden correlaties:
een gevonden significantie is pas relevant, mogelijk betekenisvol, bij een Rs > .30.
dat heeft er mee te maken dat elke Rs kleiner dan .30 als 'zeer zwak' gewaardeerd moet worden, er blijft immers 91% 'onverklaard' van de gevonden variaties. Rs tussen .30 - .50 is zwak, er blijft in het beste geval (.50) nog steeds 75% onverklaard. Rs tussen .50 - .70 is matig, en verklaart maximaal 49% (51% blijft onverklaard). Dit is waar in de sociale wetenschappen regelmatig genoegen mee wordt genomen, terwijl het nauwelijks iets zegt: het kan immers vriezen óf dooien. Een Rs tussen .70 - .85 mogen we sterk noemen, want nu blijft nog maar 28% onverklaard en dat is een prima basis voor conclusies.
Maar niet in dit onderzoek! Het uitgangspunt is immers dat beide observatie-instrumenten globaal hetzelfde meten. En dat daarmee het veldinstrument is gevalideerd. De volgende hypothesen gelden dan:
hypothese 1: instrument V is hetzelfde als instrument S, het verband = 1
hypothese 2: instrument V is niet hetzelfde als instrument S, het verband = 0
Alles lager dan +1 of elke negatieve correlatie (bv -.50) is dus een ondermijning van hypothese 1. Vooraf kan een ondergrens worden gesteld voor het aanvaarden of verwerpen van de hypothese. Stel we nemen genoegen met 75% overeenstemming, dan is de ondergrens voor hypothese 1, Rs $\geq$ .87!

De veel te lage correlaties in het onderzoek kunnen derhalve niet worden glad gestreken met de opmerking: 'echt hoge correlaties en pos significanties mogen we niet verwachten, gezien het verschil in achtergrond en insteek van beide instrumenten.' (blz 34). Wat hier feitelijk staat is: we weten eigenlijk dat die twee instrumenten goed beschouwd onvergelijkbaar zijn! En dat lijkt ons een juiste vaststelling, ook gelet op de lage correlaties. Die bevestigen eerder hypothese 2 dan hypothese 1! Het is dan weer nonsens om te stellen: 'vaak is het niet goed mogelijk om gegevens over validiteit op een methodische en statistisch verantwoorde manier te verkrijgen.' (blz 35 in de box) Het mislukt valideren in dit onderzoek is niet te wijten aan methodische noch statistische onmogelijkheden. Immers, de beoogde concurrent validatie is prima te doen met twee instrumenten die wel sterk op elkaar lijken!
En de lezer wordt afgescheept, of beter gezegd voor de gek gehouden met de bewering dat 'het veldinstrument observatie kindercentrum uitdrukkelijk niet is bedoeld om kwaliteit te meten, maar om te bepalen of de kwaliteit al dan niet voldoet.' (blz 18) En **hoe** gaan we dan vaststellen of de kwaliteit voldoet? Precies: door te meten! Niet door maar een beetje in het wilde weg te oordelen.

Conclusie: noch de inhoud-validatie noch de concurrent validatie van het veldinstrument observatie kindercentrum, is geslaagd. Op inhoud is het niet-valide gebleken, en de concurrent validiteit wijst robuust op het aanvaarden van hypothese 2: het veldinstrument meet heel andere dingen dan het schaduw instrument! Een nadere analyse van de resultaten kan derhalve achterwege blijven.