Variance Components in test generalizability research: which, when, why?

Elffers, H.; Tavecchio, L.W.C.

Variantiekomponenten en de ontwikkeling van meetinstrumenten: Welke, wanneer, waarom?

Henk Elffers
Geografisch Instituut
Rijksuniversiteit Utrecht

Louis W.C. Tavecchio
Pedagogisch Instituut
Rijksuniversiteit Leiden

Paper gepresenteerd op de ORD '79, 17 en 18 april te Nijmegen.

Inhoudsopgave:

Paper van Louis W.C. Tavecchio (paperlezer) en Henk Elffers, gepresenteerd tijdens de ORD '79, 17 en 18 april te Nijmegen. Adres paperlezer: Pedagogisch Instituut, Rijksuniversiteit Leiden, Schuttersveld 9, Leiden (tel. 071 - 148333, tst. 5699).

Variantiekomponenten en de ontwikkeling van meetinstrumenten: welke, wanneer, waarom?

1. Inleiding

De kwaliteit van tests en andere in sociaal-wetenschappelijk onderzoek gebruikte meetintrumenten wordt, voorzover het gaat om generaliseerbaarheid of betrouwbaarheid, tegenwoordig vaak geëvalueerd met behulp van variantiekomponenten in een lineair skore-model. Teneinde de grootte van de diverse komponenten vast te stellen dient een zgn. "generaliseerbaarheidsonderzoek" te worden uitgevoerd. Deze benadering is uitgewerkt door Cronbach c.s. (1972) in de monografie: "The dependability of behavioral measurements". Ondanks het baanbrekende werk van Cronbach en de zijnen rijzen er, ook bij "deskundigen", nogal eens moeilijkheden bij het opzetten van een generaliseerbaarheidsonderzoek (G-studie), bij de evaluatie van de resultaten, het naar waarde schatten van de relatieve grootte van de variantiekomponenten en het trekken van konklusies m.b.t. de kwaliteit van een meetinstrument in diverse omstandigheden. De problematiek kan worden samengevat in de vorm van twee vragen:

1. Wélke variantiekomponenten moeten in wélke situaties worden berekend, en waarom?
2. Welke (relatieve) grootte dienen variantiekomponenten te hebben teneinde het meetinstrument in de praktijk op een verantwoorde wijze toe te passen?

De problemen met betrekking tot de eerste vraag kunnen wellicht ontstaan door overkoneentratie op de techniek der variantiekomponentschatting zelf. Hoe belangrijk dit ook moge zijn, men dient zich te realiseren dat het slechts een middel vormt waarmee de kwaliteit van de verkregen skores kan worden vastgesteld. Te weinig aandacht voor de eigenschappen van de skores die door de variantiekomponenten worden gerepresenteerd kan leiden tot inkorrekt gebruik. Dit geldt in nog sterkere mate voor het gebruik van generaliseerbaarheidskoëfficienten als index voor kwaliteit, aangezien deze koëfficienten funkties zijn van variantiekomponenten.
Daarom is de benadering in dit paper toegespitst op de te verkrijgen skores en hun variabiliteit in diverse situaties. Op deze wijze worden de variantiekomponenten in het juiste licht gezien en kan een rechtstreeks antwoord op de eerste vraag worden geformuleerd. In vergelijking met hetgeen Cronbach c.s. hebben ontwikkeld is onze benadering niet nieuw. Er is wél sprake

van een andere invalshoek, een die o.i. vruchtbaar kan werken. Het antwoord op de tweede vraag wordt gegeven in het paper van Henk Elffers: "Hoe hoog behoort een generaliseerbaarheidskoëfficient te zijn?". Een geïntegreerde verhandeling over de hierboven opgeworpen problematiek vindt men in Elffers & Tavecchio (1979).

2. Het kader voor instrumentontwikkeling en de planning van instrumentgebruik

2.1. De reikwijdte van een instrument: bij het ontwikkelen van een instrument dient de ontwerper de aanstaande gebruikers te voorzien van informatie die noodzakelijk is voor de adekwate planning van onderzoek waarin het intrument toegepast gaat worden. Het is nuttig om de statistische aspekten van instrumentontwikkeling in het kader van een G-studie te onderscheiden van aspekten die te maken hebben met de toepassing ervan in het kader van diverse "D-studies" (D van "Decision"). Achtereenvolgens zullen nu een aantal begrippen worden geïntroduceerd op basis waarvan het mogelijk is op adekwate wijze G- en D-studie-opzetten te ontwerpen en uit te voeren. Deze begrippen spelen een belangrijke rol in het praktijkvoorbeeld van onderwijskundig onderzoek dat in 3 wordt geïntroduceerd.
In de eerste fase van instrumentontwikkeling specificeert een onderzoeker het interessegebied ("realm of interest"), d.w.z. hij gaat na in welke situaties het instrument een interessante skore ("score of interest") kan opleveren; in de situaties onderkent hij mogelijk relevante aspekten naast irrelevante aspekten.

Voorbeeld 1 : leerlingen verrichten verschillende taken, begeleid door leerkrachten. De interessante skore, zoals deze wordt geregistreerd door een aantal observatoren, is een maat voor de hoeveelheid initiatief; mogelijk relevante aspekten zijn het leerling aspekt, het observator aspekt, het taakaspekt, het leerkrachtaspekt. Alle andere verschillen in de meetsituatie worden beschouwd als irrelevante aspekten: bijv. uur van de dag, opeenvolging van taken, laatst bijgewoonde les, fitheid van de leerling, de klas, het klaslokaal.

De verschillende waarden die de aspekten kunnen aannemen noemen we niveaus. Zo worden de niveaus van het leerlingaspekt gevorm door de individuele leerlingen. Een aspekt kan mogelijk relevant zijn vanwege twee redenen:

1. direkte relevantie: de onderzoeker vindt het belangrijk om skores

te verkrijgen voor specifieke niveaus van het aspekt.

Vervolg voorbeeld 1: het instrument zou kunnen worden gebruikt om per leerling de prestatie op een bepaalde taak te evalueren: in dat geval zijn het leerling- en het taakaspekt beide direkt mogelijk relevant. Als, aan de andere kant, het instrument zou worden gebruikt om de effektiviteit van de leerkracht te bepalen is het leerkrachtaspekt direkt mogelijk relevant.

2. indirekte relevantie: de onderzoeker is niet zo zeer geïnteresseerd in skores voor specifieke niveaus van het aspekt als zodanig, maar hij vermoedt dat het aspekt een aanzienlijke invloed zou kunnen uitoefenen op de interessante skore en hij wil die invloed zo goed mogelijk elimineren.

Vervolg voorbeeld 1: er is sprake van een sterk effekt van de observatoren op de interessante skore, terwijl dit effekt zelf niet van belang wordt geacht. Het observatoraspekt is dan indirekt relevant. Er wordt hier gesproken over mogelijk relevant omdat het van de konkrete toepassing van het instrument afhangt of een aspekt als relevant wordt beschouwd of daarentegen zal worden geëlimineerd of veronachtzaamd. Men kan zich voorstellen dat een andere gebruiker juist wel geïnteresseerd is in indirekt relevante aspekten, bijv. verschillen tussen observatoren.

De irrelevante aspekten worden irrelevant genoemd omdat het onmogelijk of niet de moeite waard is om ze te manipuleren. (Ze hebben natuurlijk vaak wel degelijk effekt op de interessante skore!). Dit vereist wel dat hun invloed niet gerelateerd is aan die van de mogelijk relevante aspekten. Dus, specifikatie van het interessegebied ("realm of interest") bestaat uit het na ampele overweging inventariseren van mogelijk relevante aspekten, waarbij alle overige aspekten irrelevant worden verklaard. Op de tweede plaats dient echter tevens een beslissing genomen te worden over toelaatbare kombinaties van niveaus van aspekten, d.w.z.: wélke niveaus van aspekten kunnen worden aangetroffen in kombinatie met wélke niveaus van andere aspekten?

Voorbeeld 2 : in een onderzoek naar doceerstijlen zijn "schoolvakken" en "leerkrachten" mogelijk relevante aspekten. Schoolvakken heeft als niveaus "wiskunde", "aardrijkskunde", "lichamelijke opvoeding" e.d. Leerkrachten heeft niveaus als "Jan de Groot", "Richard de Vries"

etc. Het is in zo'n kontekst niet van belang om zich voor te stellen dat Jan de Groot, leraar wiskunde, gekombineerd wordt met andere schoolvakken dan zijn eigen vak. Hij geeft gewoon nooit aardrijkskunde.

In het algemeen beslissen we voor ieder paar aspekten A en B of ze als gekruist of genest dienen te worden beschouwd. De aspekten A en B zijn gekruist als ieder niveau van A binnen het interessegebied gekombineerd met ieder niveau van B kan voorkomen. A is genest in B indien er voor ieder niveau van A precies één niveau van B is waarmee het in het interessegebied gekombineerd voorkomt. Idem voor het genest zijn van B in A. Andere mogelijkheden worden hier niet in beschouwing genomen.

Vervolg voorbeeld 2: het leerkrachtaspekt is genest in het schoolvakkenaspekt.

Er zij op gewezen dat aspekten slechts als gekruist worden beschouwd als we geïnteresseerd zijn in de invloed van beide afzonderlijk. Als aspekt B op een of andere wijze ondergeschikt is aan A, kan de keuze zijn om B te beschouwen als genest in A, terwijl het op zich wellicht mogelijk zou zijn ze als gekruist te beschouwen.

Vervolg voorbeeld 1: veronderstel dat de rol van leerkrachten bij de begeleiding van de leerlingen in het verrichten van diverse taken dermate prominent is dat we nauwelijks kunnen blijven spreken over dezelfde taak wanneer hij wordt gesuperviseerd door verschillende leerkrachten. In zo'n geval lijkt het van weinig belang om taakinvloed als onafhankelijk van leerkrachtinvloed te beschouwen. Dus worden taken opgevat als genest in leerkracht, hoewel men kan opwerpen dat iedere taak gekombineerd met iedere leerkracht voorkomt.
Er zij met nadruk gesteld dat de beslissing om aspekten als gekruist of genest te behandelen moet worden genomen op basis van wat de onderzoeker belangrijk vindt, binnen het kader van zijn theoretische opvattingen en het vakgebied waarbinnen hij zich beweegt. Overwegingen met betrekking tot de experimentele opzet zijn hier nog niet aan de orde. Irrelevante aspekten worden beschouwd als genest in alle andere aspekten, met als argument dat twee identieke realisaties van irrelevante aspekten niet voorkomen.

2.2. De planning van instrumentgebruik: alvorens over te gaan tot de tweede fase van instrumentontwikkeling, waarin informatie over de invloed van diverse aspekten wordt verzameld, dient eerst aan de orde te komen

wélke informatie nu eigenlijk gewenst is.
Laten we eens uitgaan van een onderzoeker die overweegt een meetinstrument in een bepaalde situatie toe te passen, in termen van Cronbach c.s. dus iemand die een D-studie wil gaan verrichten. Het blijkt nu dat op grond van de doelstellingen van zijn onderzoek een aantal, wellicht zelfs alle, mogelijk relevante aspekten werkelijk relevant worden, waarmee wordt bedoeld dat de onderzoeker geinteresseerd is in de skores die behoren bij bepaalde niveaus van deze aspekten. Merk op dat het niét noodzakelijk is dat de werkelijk relevante aspekten in een D-studie dezelfde aspekten zijn die de testontwerper als direkt mogelijk relevant beschouwde! In het kader van een D-studie moet de onderzoeker echter beslissen wat er dient te gebeuren met alle andere mogelijk relevante aspekten:

a) hij kan verkiezen ze konstant te houden, d.w.z. metingen op slechts één niveau van deze aspekten te verrichten, waardoor de algemeenheid van zijn resultaten wordt beperkt;
b) hij kan beslissen een willekeurig niveau te kiezen steeds wanneer hij wordt gedwongen er een te selekteren, hetgeen zal resulteren in een verlies aan nauwkeurigheid van de meting;
c) hij kan ze regelmatig variëren, en dan is de vraag hoe dat zou moeten.

Bovendien zal hij zich afvragen of de irrelevante aspekten het in zijn geval niet volledig onmogelijk maken om zinvol te meten. Kortom, hij gaat na in welke mate hij de niet-werkelijk relevante aspekten zou moeten manipuleren zodat het mogelijk is om waargenomen verschillen tussen skores toe te schrijven aan verschillen tussen werkelijk relevante aspekten. Daartoe is het nodig te weten welk deel van de variabiliteit in de skores ontstaat vanwege het variëren van niet-werkelijk relevante en irrelevante aspekten, in vergelijking met de variabiliteit in de skores afkomstig van verschillen tussen werkelijk relevante aspekten. (We introduceren de term niet-relevante aspekten voor de vereniging van niet-werkelijk relevante en irrelevante aspekten). Het behoort tot de taak van een instrument ontwerper om de gebruikers te voorzien van gegevens omtrent deze verschillende soorten variabiliteit. Dus dient hij onderzoek te verrichten m.b.t. de ontwikkeling van het instrument (een G-studie in termen van Cronbach c.s.) om de hiertoe noodzakelijke gegevens te verkrijgen. Het schatten van variantiekomponenten voor de verschillende aspekten in een lineair model is een bruikbare methode om deze

informatie te verwerven, zoals we zullen zien.

2.3. De G-studie: bij het opzetten van een G-studie moet de instrument ontwerper het interessegebied ("realm of interest") enigszins inperken. Waar we voorheen spraken over aspekten met niveaus in het algemeen, dienen we nu op een meer specifieke wijze iets te zeggen over de uitgebreidheid van de verzamelingen niveaus van de aspekten, teneinde een adekwate opzet te realiseren. Een G-studie die informatie verschaft over alle mogelijke denkbare niveaus is eenvoudigweg niet uitvoerbaar, dus dienen we zorgvuldig te beschrijven op welke beperkte verzameling niveaus het onderzoek van toepassing is. De huidige statistische technieken laten slechts toe om uitspraken te doen over verzamelingen niveaus indien skores worden geanalyseerd afkomstig van een deelverzameling van niveaus die kunnen worden beschouwd als een aselekte steekproef uit de grotere verzameling. Dus specificeren we voor ieder aspekt de interessante niveaus ("levels of interest") als de grootste verzameling waarin we geïnteresseerd zijn en waaruit een aselekte deelverzameling kan worden getrokken die wordt opgenomen in de experimentele opzet van de G-studie. O.i. is het niet noodzakelijk dat ieder niveau afzonderlijk aselekt wordt getrokken; we zijn tevreden met het ontbreken van duidelijke indikaties dat de geanalyseerde niveaus aanzienlijk verschillen van de grotere verzameling.
Het kan voorkomen dat de niveaus van een aspekt dat in de G-studie is opgenomen alle interessante niveaus omvatten, hetzij omdat we niet geïnteresseerd zijn in andere niveaus, hetzij omdat we het onredelijk vinden om de geanalyseerde niveaus te beschouwen als een aselekte steekproef uit een grotere verzameling. Een dergelijk aspekt noemen we fixed, anders spreken we van random aspekten. De interessante populatie ("population of interest") wordt nu gevormd door de verzameling van alle skores die hypothetisch verkrijgbaar zijn bij alle toelaatbare kombinaties van interessante niveaus van mogelijk relevante aspekten en alle niveaus van de irrelevante aspekten. Merk op dat het verwijzen naar de toelaatbaarheid van kombinaties met zich meebrengt dat het gekruist en genest voorkomen van aspekten in de interessante populatie rechtstreeks wordt overgenomen uit het interessegebied ("realm of interest").
Een G-studie verschaft gegevens over de variabiliteit van skores in bovengenoemde interessante populatie . We benadrukken dat het vast-

stellen van interessante niveaus ("levels of interest") en de beslissing of een aspekt random of fixed is beide onderdeel vormen van een beslissingsproces waarin de praktische problemen van haalbare G-studie-opzetten moeten worden afgewogen tegen het verlies aan algemeenheid van uitspraken dat voortvloeit uit een sterke restriktie m.b.t. de interessante niveaus.

Vervolg voorbeeld 2: laten we aannemen dat het eenvoudig is om het instrument te gebruiken bij een aantal leerkrachten van school A, maar dat tevens bekend is dat deze school niet representatief is voor scholen van type B. We zien ons nu geplaatst voor een dilemma: ófwel de interessante niveaus van het aspekt "leerkracht" moeten worden beperkt tot "leerkrachten van school A", omdat we niet durven aan te nemen dat de onderzochte leerkrachten een quasi-aselekte steekproef vormen uit "leerkrachten afkomstig van scholen van type B", waarin we wérkelijk geïnteresseerd zijn, ófwel we zijn gedwongen om een meer gelijkmatig opgezet experiment uit te voeren bij leerkrachten afkomstig van verschillende scholen, hetgeen echter aanzienlijk meer tijd, geld en energie zou vergen.

Om de informatie die in het kader van een G-studie wordt verkregen toe te passen, dient, strikt genomen, iedere D-studie betrekking te hebben op een bepaalde deelverzameling uit de interessante populatie. Daarom is een zo groot mogelijke uitbreiding van de interessante niveaus wenselijk. In de praktijk van het onderzoek is men al gauw geneigd om deze eis af te zwakken teneinde het instrument aan te passen voor gebruik in andere situaties, "dichter bij huis". Het is belangrijk dat een instrument ontwerper nauwkeurig uiteenzet waarom hij de interessante niveaus op een bepaalde manier heeft ingeperkt. Slechts dán kunnen gebruikers van het instrument beoordelen of het door hen beoogde gebruik van het instrument geoorloofd is.

Vervolg voorbeeld 1: veronderstel dat in de G-studie de interessante niveaus van het observator- en taakaspekt worden gedefinieerd als "een gegeven verzameling observatoren $O_1,\ldots,O_l$", en "een gegeven verzameling taken $T_1,\ldots,T_k$", beide fixed aspekten. We zouden best willen generaliseren naar alle observatoren van een bepaald type, maar tijdens de observatortraining kan gebleken zijn dat bizondere kenmerken van observatoren een grote rol spelen, terwijl de selektie van de feitelijke observatoren ons niet toestaat hen te beschouwen als een quasi-aselekte steekproef uit alle observatoren. Taakaspekt

kan als fixed worden beschouwd omdat we niet geïnteresseerd zijn in andere taken, bijv. omdat de taken $T_1,....,T_k$ als groep onderwerp van onderzoek zijn. Anderzijds zouden we de interessante niveaus kunnen definiëren als "taken zoals $T_1,....,T_k$ binnen een bepaald gebied", daarmee $T_1,....,T_k$ beschouwend als een quasi-aselekte steekproef hieruit, waardoor "taak" een random aspekt wordt.

Merk op dat we de interessante niveaus van de irrelevante aspekten niet hebben gespecificeerd. Wellicht zou dit toch moeten gebeuren, omdat anders de irrelevante bijdragen in een D-studie aanzienlijk hoger zouden kunnen uitvallen dan aanvankelijk bleek uit de resultaten van de G-studie.

3. Een voorbeeld uit de praktijk van het onderwijskundig onderzoek

Elders wordt nader ingegaan op het lineaire model zoals dat geldt in de interessante populatie ("population of interest") en het model voor de G-studie (zie Elffers & Tavecchio (1979) pp. 6-9). De ontwikkelde inzichten en procedures, die overigens algemeen toepasbaar zijn, worden binnen het bestek van dit paper nu verder geïllustreerd aan de hand van een praktijkvoorbeeld uit de onderwijskunde, waarbij als meetintrument een observatieschema werd gebruikt.

3.1. PEIAS II

de tweede versie van het Physical Education Interaction Analysis System (PEIAS II, vgl. Tavecchio, 1977) wordt geïntroduceerd als voorbeeld van het gebruik van variantiekomponenten bij meetinstrumentontwikkeling. Het gaat hier om een gedeeltelijk geneste proefopzet, waarbij gebruik werd gemaakt van een fixed model zonder interaktie (zie Elffers & Tavecchio (1979) voor een voorbeeld van een random model met interaktie).

PEIAS II werd gekonstrueerd om het onderwijsgedrag van leerkrachten lichamelijke opvoeding te karakteriseren alsmede om ze onderling te vergelijken. Het is een zgn. "affektief" observatiesysteem, waarbij de nadruk valt op het sociaal-emotionele klimaat van het onderwijsleerproces in de les lichamelijke opvoeding. Het instrument bevat 16 kategorieën voor het koderen van nondirektief, neutraal en direktief onderwijsgedrag. PEIAS II was ontworpen met het doel de direktiviteit van leerkrachten lichamelijke opvoeding te meten en de leerkrachten onderling te vergelijken. Bij het opzetten van een G-studie wordt de onderzoeker echter gekonfronteerd met een groot aantal va-

riabelen die de kategorieskores van het observatiesysteem kunnen beïnvloeden (naast de "werkelijk" bestaande verschillen in direktiviteit tussen de leerkrachten). De evaluatie van PEIAS II als meetinstrument dient dan ook gebaseerd te zijn op het ontdekken en kwantificeren van (tenminste een aantal van) deze variabelen.
In termen van hetgeen in paragraaf 2 werd uiteengezet dient de onderzoeker een poging te doen definities te geven van het interessegebied ("realm of interest"), de interessante skore ("score of interest"), het interessante niveau ("level of interest") en de interessante populatie ("population of interest"):
- Specifikatie van het interessegebied: we gaan uit van de skore op een bepaalde PEIAS II-kategorie, bijv. kategorie 1: "Acceptance of feelings, collectively". Een groot aantal aspekten beïnvloedt de interessante skore, d.w.z. de fraktie tijd doorgebracht in een bepaalde kategorie, in dit geval een maat voor het aksepterend gedrag van de leerkracht; deze aspekten zijn: het leerkrachtaspekt, het leerjaaraspekt, het klasaspekt, het observatoraspekt, het lestypeaspekt, het schoolaspekt, het uur-van-de-dagaspekt, etc. De ontwerpers van PEIAS II beschouwden het leerkrachtaspekt, het leerjaaraspekt, het observatoraspekt, het klasaspekt en het lestypeaspekt als mogelijk relevante aspekten. Hiervan is het leerkrachtaspekt het meest prominente en op zich van groot belang omdat PEIAS II bedoeld is om de affektieve kwaliteit van leerkrachtgedrag te meten, m.a.w. het leerkrachtaspekt is direkt relevant.
De overige mogelijk relevante aspekten zijn belangrijk omdat ze aanzienlijke invloed kunnen uitoefenen op de interessante skore, hoewel de onderzoekers niet in deze aspekten als zodanig geïnteresseerd waren (indirekt relevant). Het klasaspekt en het lestypeaspekt werden geëlimineerd en zullen hier niet worden besproken.
- Vervolgens het vraagstuk van de gekruistheid en genestheid van aspekten: de onderzoekers besloten om het leerjaaraspekt te beschouwen als genest in het leerkrachtaspekt, niet omdat ze niet als gekruist konden worden opgevat (dit kan wel degelijk), maar omdat de onderzoekers argumenteerden dat variaties in onderwijsgedrag als gevolg van het lesgeven in verschillende leerjaren voor leerkracht A iets heel anders kan betekenen dan voor leerkracht B, althans voor bepaalde kategorieën. Dus werden leerjaarverschillen onderzocht als genest in leerkracht, met andere woorden: de onderzoekers beschouwden de gekombineerde invloed(en) van leerjaaraspekt en leerkrachtaspekt als

belangrijker en van meer betekenis dan het leerjaaraspekt afzonderlijk. De overige aspekten werden als gekruist beschouwd.

– Specifikatie van interessante niveaus en de interessante populatie: om een G-studie te kunnen opzetten dient de verzameling interessante niveaus van ieder aspekt te worden geschetst als de grootste verzameling waaruit een quasi-aselekte steekproef zou kunnen worden getrokken. In dit geval oordeelden de onderzoekers het onjuist om de diverse niveaus van de aspekten als aselekt getrokken te beschouwen. Ze werden als fixed opgevat. Met betrekking tot het leerkrachtaspekt bestonden de interessante niveaus uit leerkracht 1, 2, 3 en 4; zij allen maakten deel uit van de G-studie-opzet. Met betrekking tot het leerjaar-in-leerkrachtaspekt waren er 2 interessante niveaus, te weten het hoge leerjaarniveau en het lage leerjaarniveau, ook beide opgenomen in de experimentele opzet. Voor het observatoraspekt werden 3 interessante niveaus gespecificeerd, de observatoren 1, 2 en 3, allen participerend aan de G-studie. Uitspraken ten aanzien van D-studies met andere leerkrachten, observatoren, e.d. kunnen alleen worden gerechtvaardigd onder additionele assumpties!

3.2. Het lineaire model voor de G-studie

Elk van de 4 leerkrachten werd bij twee gelegenheden in elk van de 2 leerjaren door elk van de 3 observatoren geobserveerd (er werd gebruik gemaakt van video-opnamen). Gebaseerd op de experimentele opzet en uitgaande van de skore op een gegeven PEIAS II-kategorie werd het volgende lineaire model aangenomen:

(1) $X(t,g:t,o,\underline{r}) = \mu + T(t) + G:T(g:t) + O(o) + \varepsilon(\underline{r})$

waarin:

$X(t,g:t,o,\underline{r})$ = de skore van de leerkracht op de kategorie;

μ = het algemeen gemiddelde;

$T(t)$ = het hoofdeffekt van leerkracht t (t = 1,2,3,4; als nevenvoorwaarde geldt $\sum_{t=1}^{4} T(t) = 0$; hieruit volgt dat T(t) het verschil is tussen het effekt van leerkracht t en het gemiddelde leerkrachteffekt, d.w.z. μ. Dezelfde redenering gaat op bij de overige nevenvoorwaarden;

$G:T(g:t)$ = het hoofdeffekt van het g-de leerjaar in leerkracht t (g = 1,2, genest in t; de nevenvoorwaarde is $\sum_{g=1}^{2} G:T(g:t) = 0$, voor iedere t;

$O(o)$ = het hoofdeffekt van observator o (o = 1,2,3, met als nevenvoorwaarde $\sum_{o=1}^{3} O(o) = 0$);

$\varepsilon(\underline{r})$ = bijdrage van het volgens aanname random irrelevante aspekt, $\varepsilon(\underline{r}) = \varepsilon(r:(t,g:t,o)), r = 1,2$.

Merk op dat t, g en o, die alle drie niveaus aanduiden van fixed aspekten, niét onderstreept zijn, terwijl r, een niveau van een random aspekt, onderstreept wordt. Verder is het van belang er op te wijzen dat in bovenstaand model leerkracht x observator interaktie en leerjaar-in-leerkracht x observator interaktie gelijk worden gesteld aan nul (vgl. Tavecchio, 1977, blz. 66). We zullen nu niet nader ingaan op de schatting van de variantiekomponenten in bovenstaand model, maar laten zien hoe ze gebruikt en geïnterpreteerd worden, een en ander aan de hand van de gegevens in onderstaande tabel.

Tabel 1: Geschatte variantiekomponenten voor PEIAS II kategorie 1, "Acceptance of feelings, collectively".

Aspekt	Komponent	Schatting
Leerkracht	$\hat{\sigma}^2(T)$	0.00097
Leerjaar-in-leerkracht	$\hat{\sigma}^2(G:T)$	0.00103
Observator	$\hat{\sigma}^2(O)$	0.00047
Irrelevant	$\hat{\sigma}^2(\underline{\varepsilon})$	0.00076
Totale variantie		0.00323

4. Evaluatie van geplande toepassingen van een instrument: het PEIAS II voorbeeld

We zullen nu nader ingaan op de vraag hóe variantiekomponenten kunnen worden gebruikt bij de evaluatie van geplande toepassingen van een instrument. Hiertoe volgen we een onderzoeker bij de evaluatie van zijn D-studies. In deze paragraaf bespreken we een aantal D-studies met het PEIAS II observatiesysteem uit 3, gebruikmakend van het lineaire model (1) met geneste aspekten, zonder interaktie. In 4.1 wordt de argumentatie nader uitgewerkt aan de hand van voorbeelden, terwijl in 4.2 getracht wordt algemene richtlijnen te formuleren.

4.1. Uiteenzetting van de evaluatieprocedure aan de hand van voorbeelden

We volgen een onderzoeker die van plan is een bepaald instrument toe te passen (D-studie). Bepaalde aspekten beschouwt hij als werkelijk relevant, d.w.z. hij is bereid om verschillen tussen verkregen skores toe te schrijven aan verschillen tussen deze werkelijk relevante aspekten en niét aan verschillen tussen de overige aspekten; daarbij vraagt hij zich af op welke wijze hij deze overige aspekten moet manipuleren om hun negatieve invloed zo veel mogelijk uit te schakelen. Bij wijze van voorbeeld nemen we een onderzoeker die van plan is het PEIAS II instrument toe te passen met als doelstelling de mate van direktiviteit van een aantal leerkrachten lichamelijke opvoeding te meten. We beperken ons in de diskussie verder tot kategorie 1, "Acceptance of feelings, collectively". De onderzoeker beschouwt het leerkrachtaspekt als werkelijk relevant: hij hoopt een eventueel verschil tussen waargenomen skores van leerkrachten toe te kunnen schrijven aan een verschil in direktiviteit tussen de betreffende leerkrachten. Verder vraagt hij zich af of het nodig zal zijn meer dan één observator te gebruiken. Een andere kwestie is of observaties moeten worden verricht in diverse leerjaar/klas kombinaties. Hoeveel lesparen moeten worden geobserveerd? M.a.w.: wat moet er gebeuren met de niet-relevante aspekten?

Op grond van financiële en organisatorische overwegingen en de behoefte om het schoolgebeuren zo min mogelijk te beïnvloeden zou het goed uitkomen als hij zou kunnen volstaan met de observatie van één willekeurig paar lessen, in één willekeurig leerjaar door één willekeurige observator. Maar is dat werkelijk mogelijk? Laten we de skores die op die manier worden verkregen eens nader bestuderen. Laten we eens kijken naar de leerkrachten t_1 en t_2 en hun aldus verkregen skores, d.w.z.: we kijken naar onafhankelijk en aselekt gekozen leerjaren $g_1{:}t_1$ voor leerkracht t_1 en $g_2{:}t_2$ voor leerkracht t_2. Merk op dat we "willekeurig leerjaar" in de werkelijke situatie vertalen in "leerjaar g:t aselekt gekozen uit de interessante niveaus" binnen ons model. Op dezelfde wijze kiezen we voor ieder paar lessen dat moet worden geobserveerd op aselekte wijze een observator, zeg $\underline{o}_1$ en $\underline{o}_2$, uit de verzameling observatoren. We nemen aan dat de meting plaatsvindt op onafhankelijke en aselekte niveaus van irrelevante aspekten $\underline{r}_1$ en $\underline{r}_2$ ("overige omstandigheden"). De aldus verkregen skores zijn $x(t_1, g_1{:}t_1, \underline{o}_1, \underline{r}_1)$ en $x(t_2, g_2{:}t_2, \underline{o}_2, \underline{r}_2)$, die we verder voor het gemak aanduiden met respektievelijk $\underline{y}(t_1)$ en $\underline{y}(t_2)$.

De door de onderzoeker geprefereerde opzet is adekwaat indien het gerechtvaardigd is om uit $\underline{y}(t_1) > \underline{y}(t_2)$ af te leiden dat $T(t_1) > T(t_2)$, en vice versa, d.w.z. verschillen tussen waargenomen skores worden "veroorzaakt" door verschillen tussen leerkrachtkenmerken. Om na te gaan of een dergelijke konklusie juist is kijken we naar het verschil tussen de waargenomen skores:

$$(4.1.)\quad \underline{\Delta}(t_1,t_2) \stackrel{def}{=} \underline{y}(t_1) - \underline{y}(t_2) = \mu + T(t_1) + G{:}T(\underline{g}_1{:}t_1) + O(\underline{o}_1) + \underline{\varepsilon}_1 - \mu - T(t_2) - G{:}T(\underline{g}_2{:}t_2) - O(\underline{o}_2) - \underline{\varepsilon}_2$$

$$= \left\{T(t_1) - T(t_2)\right\} + \left\{G{:}T(\underline{g}_1{:}t_1) - G{:}T(\underline{g}_2{:}t_2)\right\} + \left\{O(\underline{o}_1) - O(\underline{o}_2)\right\} + \left\{\underline{\varepsilon}_1 - \underline{\varepsilon}_2\right\}$$

met als vraagstelling of de bijdragen van alle termen aan de rechterkant te verwaarlozen zijn in vergelijking met het T-verschil, want daar gaat het om. Dus proberen we te achterhalen welke bijdragen naar verwachting afkomstig zijn van een aselekte keuze van niveaus van observator-, leerjaar- en irrelevante aspekten. We stellen daarom voor om $E(\underline{\Delta}(t_1,t_2))^2$ te evalueren, dit is de <u>verwachting</u> (bij aselekte trekkingen uit de verzamelingen interessante niveaus) van het gekwadrateerde waargenomen verschil, waarbij de kwadratering slechts dient om te voorkomen dat positieve en negatieve verschillen tegen elkaar wegvallen. Deze grootheid ziet er nader uitgewerkt betrekkelijk doorzichtig uit, vanwege de onafhankelijkheid van keuzen en de nul-verwachting van individuele termen*):

$$(4.2.)\quad E(\underline{\Delta}(t_1,t_2))^2 = \left\{T(t_1) - T(t_2)\right\}^2 + E\left\{G{:}T(\underline{g}_1{:}t_2) - G{:}T(\underline{g}_2{:}t_2)\right\}^2 + E\left\{O(\underline{o}_1) - O(\underline{o}_2)\right\}^2 + E(\underline{\varepsilon}_1 - \underline{\varepsilon}_2)^2$$

$$= \left\{T(t_1) - T(t_2)\right\}^2 + 2\sigma^2(G{:}T) + 2\sigma^2(O) + 2\sigma^2(\underline{\varepsilon})$$

Tot nu toe hebben we ons beziggehouden met twee bepaalde leerkrachten, t_1 en t_2. We zijn echter niet geïnteresseerd in deze twee leerkrachten, we willen de verschillen tussen <u>welke twee dan ook</u> evalueren. Daarom bepalen we deze grootheid voor twee onafhankelijke, aselekt gekozen leerkrachten, $\underline{t}_1$ en $\underline{t}_2$, d.w.z. we kijken naar:

---

*) Met betrekking tot de variantie van leerjaar bijdragen gaan we hierbij uit van een assumptie die niet noodzakelijk is maar de afleidingen vereenvoudigt, nl.: voor alle t: $E((G{:}T)(\underline{g}{:}t))^2 = \sigma^2(G{:}T)$.

$$(4.3.)\quad E(\underline{\Delta}(\underline{t}_1,\underline{t}_2))^2 = E\left\{\underline{Y}(\underline{t}_1) - \underline{Y}(\underline{t}_2)\right\}^2 =$$

$$= E\left\{T(\underline{t}_1) - T(\underline{t}_2)\right\}^2 + 2\sigma^2(G:T) + 2\sigma^2(O) + 2\sigma^2(\underline{\varepsilon}) =$$

$$= 2\sigma^2(T) + 2\sigma^2(G:T) + 2\sigma^2(O) + 2\sigma^2(\underline{\varepsilon})$$

onder gebruikmaking van de onafhankelijkheid van leerkrachtkeuze en de nevenvoorwaarde $E(T(\underline{t}_1)) = E(T(\underline{t}_2)) = 0$

Uit 4.3. konkluderen we:
de te verwachten verschillen tussen skores in deze D-studie opzet worden voornamelijk veroorzaakt door het relevante aspekt T indien $2\sigma^2(T)$ "veel groter" is dan de overige komponenten $2\sigma^2(G:T) + 2\sigma^2(O) + 2\sigma^2(\underline{\varepsilon})$.
In hoofdstuk 6 van Elffers & Tavecchio (1979) wordt het begrip "veel groter" nader uitgewerkt. Daar vindt men de aanbeveling om de kwaliteit van een D-studie opzet als net voldoende te interpreteren indien veel groter wordt ingevuld als "tenminste 2 x zo groot"; als redelijk indien veel groter wordt ingevuld als "tenminste 4 x zo groot"; als goed bij "tenminste 9 x zo groot". De kwaliteit is slecht als zelfs de kwalifikatie net voldoende niet haalbaar blijkt. Als we het teken $\gg$ introduceren voor "is veel groter dan", zijn we in dit geval redelijk tevreden indien:

$$(4.4.)\quad \sigma^2(T) \gg \sigma^2(G:T) + \sigma^2(O) + \sigma^2(\underline{\varepsilon})$$

In het vervolg spreken we over ongelijkheden als deze in termen van kwaliteitsongelijkheden, want een D-studie opzet is van bevredigende kwaliteit als deze ongelijkheid opgaat. Soms zullen we ook spreken over de kwaliteitsratio Q, waarmee we de ratio tussen de linkerkant en de rechterkant van de kwaliteitsongelijkheid aanduiden. De kwaliteitsongelijkheid gaat alleen op indien $Q \gg 1$.
We noemen de zojuist besproken D-studie opzet voorbeeld 4.1. en verkrijgen de volgende getallen:

Voorbeeld 4.1.: Voor PEIAS II kategorie 1 verkrijgen we voor deze D-studie opzet het volgende resultaat (zie Tabel 1, blz. 11): $\sigma^2(T) = 0.00097$, $\sigma^2(G:T) = 0.00103$, $\sigma^2(O) = 0.00047$ en $\sigma^2(\underline{\varepsilon}) = 0.00076$. Het blijkt dat de kwaliteitsongelijkheid niét opgaat: $0.00097 \not\gg 0.00226 (= 0.00103 + 0.00047 + 0.00076)$, of wanneer men de kwaliteitsratio Q gebruikt: $Q = 0.00097/0.00226 = 0.43 \not\gg 1$.
Als de kwaliteitsongelijkheid niet opgaat, kunnen we proberen de experimentele opzet aan te passen teneinde de invloed van de niet-relevante

aspekten te verminderen. Er zijn twee mogelijkheden:

1. het verrichten van metingen bij meer, of zelfs alle, niveaus van een aspekt en dan het gemiddelde van de aldus verkregen skores nemen; dit resulteert in een kleinere bijdrage van het aspekt, maar het kost nogal wat extra tijd aan metingen.
2. het konstant houden van een aspekt, d.w.z. het verrichten van metingen bij een enkel fixed niveau van dat aspekt; dit resulteert in het verdwijnen van het aspekt uit de kwaliteitsongelijkheid, maar de meting is op deze wijze wel moeilijker te realiseren (bijv.: steeds opnieuw dezelfde observator(en) inzetten is moeilijk te organiseren).

Laten we ter illustratie eens kijken naar het gemiddelde van de skores voor een bepaalde leerkracht in beide leerjaren, zoals geobserveerd door een zelfde vaste observator, zeg o_1, daartoe kijken we naar:

$$(4.5.)\quad \underline{Y}_2(t_1) = \tfrac{1}{2}\left\{x(t_1,g_1{:}t_1,o_1,\underline{r}_1) + x(t_1,g_2{:}t_1,o_1,\underline{r}_2)\right\}$$

$$\underline{Y}_2(t_2) = \tfrac{1}{2}\left\{x(t_2,g_1{:}t_2,o_1,\underline{r}_3) + x(t_2,g_2{:}t_2,o_1,\underline{r}_4)\right\}$$

Merk op dat:

$$(4.6.)\quad \underline{Y}_2(t_1) = \mu + T(t_1) + \tfrac{1}{2}\left\{G{:}T(g_1{:}t_1) + G{:}T(g_2{:}t_1)\right\} + O(o_1) + \tfrac{1}{2}(\underline{\varepsilon}_1+\underline{\varepsilon}_2)$$

$$= \mu + T(t_1) + O(o_1) + \tfrac{1}{2}(\underline{\varepsilon}_1+\underline{\varepsilon}_2)$$

waarbij de leerjaartermen verdwijnen vanwege de nevenvoorwaarde die behoort bij het leerjaareffekt. We stellen nu opnieuw de vraag of het verschil $\underline{Y}_2(t_1) - \underline{Y}_2(t_2)$ mag worden toegeschreven aan het verschil in leerkrachteffekt $T(t_1) - T(t_2)$; daartoe evalueren we:

$$(4.7.)\quad \underline{\Delta}_2(t_1,t_2) = T(t_1) - T(t_2) + \tfrac{1}{2}(\underline{\varepsilon}_1+\underline{\varepsilon}_2+\underline{\varepsilon}_3+\underline{\varepsilon}_4)$$

waarin, zoals we zien, het observatoreffekt $O(o_1)$ is weggevallen aangezien dit in beide skores aanwezig is vanwege het feit dat we gebruik maken van een vaste observator. Evenals tevoren uitgaande van een aselekt paar leraren $\underline{t}_1$, $\underline{t}_2$ en onder gebruikmaking van de gekwadrateerde verwachtingen, komen we nu uit bij:

$$(4.8.)\quad E\left\{\underline{\Delta}(\underline{t}_1,\underline{t}_2)\right\}^2 = 2\sigma^2(T) + \sigma^2(\underline{\varepsilon})$$

konkluderend dat het verantwoord is om skoreverschillen in $\underline{Y}_2$ toe te schrijven aan leerkrachtverschillen indien:

$$(4.9.)\quad \sigma^2(T) >> \tfrac{1}{2}\sigma^2(\underline{\varepsilon})$$

Deze kwaliteitsongelijkheid gaat vanzelfsprekend eerder op dan (4.4.), maar dit valt te verwachten op grond van de meer uitgebreide metingen. Als we deze D-studie opzet voorbeeld 4.2. noemen verkrijgen we de volgende getallen:

Voorbeeld 4.2.: (PEIAS II, kategorie 1): $\sigma^2(T) = 0.00097$, $\sigma^2(\underline{\varepsilon}) = 0.00076$, dus de kwaliteitsongelijkheid vereist dat $0.00097 >> 0.00038$, $Q = 2.55$, hetgeen resulteert in de uitspraak dat deze D-studie opzet net voldoende is.

In een derde voorbeeld laten we zien dat op overeenkomstige wijze verschillende keuzen t.a.v. welke aspekten relevant zijn kunnen worden behandeld. Bijvoorbeeld, als we zowel leerkracht- als leerjaar-in-leerkrachtaspekten als werkelijk relevant beschouwen en de skores zijn toegekend door een aselekt gekozen observator, dan hebben we te maken met de skores:

(4.10) $\underline{Y}_3(t_1,g_1:t_1) = x(t_1,g_1:t_1,\underline{o}_1,\underline{r}_1)$

$\underline{Y}_3(t_2,g_2:t_2) = x(t_2,g_2:t_2,\underline{o}_2,\underline{r}_2)$

waarbij we ons afvragen of de verschillen $\underline{\Delta}_3((t_1,g_1:t_1),(t_2,g_2:t_2))$ kunnen worden toegeschreven aan verschillen tussen leerkracht/leerjaar kombinaties. Omdat we niet speciaal geïnteresseerd zijn in $(t_1,g_1:t_1)$, $(t_2,g_2:t_2)$, evalueren we $\underline{\Delta}_3$ voor onafhankelijk en aselekt gekozen $\underline{t}_1$, $\underline{t}_2$, $g_1:\underline{t}_1$, $g_2:\underline{t}_2$, door te kijken naar:

(4.11) $E\{\underline{\Delta}_3((\underline{t}_1,g_1:\underline{t}_1),(\underline{t}_2,g_2:\underline{t}_2))\}^2 = 2\sigma^2(T) + 2\sigma^2(G:T) + 2\sigma^2(O) + 2\sigma^2(\underline{\varepsilon})$

konkluderend dat we tevreden kunnen zijn indien de kwaliteitsongelijkheid

(4.12) $\sigma^2(T) + \sigma^2(G:T) >> \sigma^2(O) + \sigma^2(\underline{\varepsilon})$

opgaat, aangezien nu zowel het leerkracht- als het leerjaaraspekt relevant zijn. Als we deze D-studie opzet voorbeeld 4.3. noemen krijgen we:

Voorbeeld 4.3.: (PEIAS II, kategorie 1): de kwaliteitsongelijkheid luidt: $0.00097 + 0.00103 >> 0.00047 + 0.00076$?; dit is niet zo bevredigend. De kwaliteitsratio Q is 1.63. De konklusie moet luiden dat deze opzet niet geschikt is voor zijn doel.

4.2. Het evaluatieschema

Het is duidelijk dat we zouden kunnen doorgaan met het presenteren van een groot aantal variaties in beoogde skores vergezeld van de bijbehorende evaluaties, net zo goed als het mogelijk is om een groot aantal voor-

beelden te geven van geplande toepassingen van een instrument. De algemene lijn die men moet volgen bij de evaluatie van een D-studie is de volgende:

(1) Specificeer welke aspekten werkelijk relevant zijn;

(2) Definiëer de daarbij behorende skores en geef aan over hoeveel niveaus van niet-relevante aspekten zal worden gemiddeld en welke aspekten konstant zullen worden gehouden;

(3) Evalueer het verwachte gekwadrateerde verschil van de aldus gedefiniëerde skores, uitgaande van onafhankelijke, aselekte trekkingen van niveaus van relevante aspekten;

(4) Vergelijk de bijdrage(n) behorend bij de relevante aspekten met die afkomstig van de niet-relevante aspekten met behulp van de kwaliteitsongelijkheid of de kwaliteitsratio.

Indien deze vergelijking niet gunstig uitvalt, kan getracht worden de D-studie opzet aan te passen. Op de eerste plaats identificeren we de verantwoordelijke niet-relevante aspekten, d.w.z. degene die erg grote bijdragen leveren aan het verwachte gekwadrateerde verschil. Vervolgens verkleinen we deze bijdragen door de opzet op een van de volgende manieren bij te stellen:

(a) door het opnemen van meer aselekt gekozen niveaus van de verantwoordelijke aspekten; de beoogde skore is dan het gemiddelde van de skores verkregen op deze niveaus; deze procedure reduceert de bijdrage van een dergelijk aspekt tot $\frac{1}{k}\sigma^2$ (aspekt), wanneer k niveaus worden opgenomen.

(a') als een speciaal geval van (a) kunnen we, als het aspekt fixed is, het aantal niveaus maximaal uitbreiden, d.w.z. metingen verrichten op alle interessante niveaus, hetgeen resulteert in het volledig verdwijnen van de variantiekomponent voor dat aspekt, vanwege de nevenvoorwaarde dat de som van alle niveaubijdragen van een fixed aspekt gelijk is aan nul.

(b) door, in plaats van een aselekt gekozen verzameling, steeds dezelfde konstante verzameling niveaus van een niet-relevant aspekt te nemen bij iedere kombinatie van niveaus van relevante aspekten. Dit resulteert in het verdwijnen van de niveaubijdragen van deze identieke niveaus in de $\underline{\Delta}$-skore en dientengevolge tevens in het verdwijnen van de variantiekomponent uit de kwaliteitsongelijkheid.

Merk op dat (a) en (a') meer metingen vereisen, terwijl (b) soms moeilijk te realiseren is, bijv. het is moeilijk om steeds hetzelfde team

observatoren bij iedere meting in te zetten. In Elffers & Tavecchio (1979, pp. 21-22) vindt men een nadere uitwerking van dit evaluatie-schema, waarin ook rekening wordt gehouden met de aanwezigheid van interaktietermen in een model en tevens aandacht wordt besteed aan de irrelevante aspekten.

5. Literatuur:

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N.: The dependability of behavioral measurements. New York: Wiley, 1972.

Elffers, H. & Tavecchio, L.W.C.: Variance components in test generalizability research: Which, when, why? Vereniging voor Onderwijsresearch, VOR-publikatie nr. 9, 1979.

Tavecchio, L.W.C.: Quantification of teaching behavior in physical education: A methodological study. Akademisch proefschrift, Universiteit van Amsterdam, 1977. (Ook verkrijgbaar bij University Microfilms International, order no: 77-70,039).

De WEP-REEKS omvat (pré)publikaties (EX), interne notities (IN), skripties en werkstukken (SW), rekenprogramma's (RP), onderwijsprépublikaties (OW) en boekbesprekingen (BB). Ze kunnen worden aangevraagd bij de auteurs of het sekretariaat van de WEP (Schuttersveld 9, kamer 500, 2316 XG Leiden), behoudens de interne publikaties, die in het algemeen niet verder worden verspreid.