J.J. Hox

Principes en toepassing van structurele modellen

Samenvatting: Dit artikel geeft een niet-technische inleiding in het gebruik van structurele modellen, ook wel kort aangeduid met LISREL-modellen. Structurele modellen stellen onderzoekers in staat ingewikkelde modellen te construeren, waaronder factoranalyse en multipele regressieanalyse, en een combinaties hiervan. Onderzoekers moeten voorafgaand aan de eigenlijke analyses een model specificeren voor hun gegevens. Vervolgens wordt dit model getoetst aan de empirische gegevens. Wanneer het model niet goed past, kunnen suggesties gegeven worden voor verbetering van het model middels zogenaamde modificatie-indexen.

200
Kind en Adolescent, 1999, 20 (3), p. 200 - 217

'Structurele modellen' is de technische term voor wat in de wandelgangen dikwijls wordt aangeduid als LISREL-modellen. In Engelstalige publicaties heten ze tegenwoordig kortweg SEM, de afkorting van Structural Equations Modeling. Wat structurele modellen inhouden is waarschijnlijk het beste toe te lichten met een aantal voorbeelden. Dit gebeurt aan de hand van een zogenaamd *paddiagram* (voor het eerst voorgesteld in Wright, 1921): een grafische weergave van de veronderstelde verbanden in het model. Een paddiagram is een figuur bestaande uit een verzameling vierkanten, cirkels en pijlen. De vierkanten representeren geobserveerde variabelen die empirisch gemeten kunnen worden. Daarnaast zijn er latente variabelen. Dit zijn hypothetische factoren waarvan wordt aangenomen dat ze 'onder' de geobserveerde variabelen liggen, en waarvoor de geobserveerde variabelen dienen als empirische indicatoren. Latente factoren worden in het paddiagram weergegeven met cirkels of ellipsen. Daarnaast bevat het paddiagram pijlen, die dienen om verbanden tussen de variabelen weer te geven. Enkelvoudige pijlen worden gebruikt voor gerichte verbanden, die causaal geïnterpreteerd worden. Pijlen met twee punten worden gebruikt om verbanden aan te geven waarvoor geen causale interpretatie geldt.

In het vervolg van dit artikel wordt allereerst ingegaan op verschillende soorten structurele modellen, namelijk factoranalyse, multipele regressieanalyse, en padmodellen. Daarna wordt ingegaan op de technieken die gebruikt worden om het model te schatten, en om te bepalen of het model goed past bij de empirische gegevens. Ten slotte wordt nog ingegaan op enkele speciale modellen, en op de meest gebruikelijke software.

Prof. dr. J.J. Hox is werkzaam bij de Capaciteitsgroep Methodenleer en Statistiek van de Universiteit Utrecht.

Contactadres: Capaciteitsgroep Methodenleer en Statistiek, Postbus 80140, 3508TC Utrecht, email: j.hox@fss.uu.nl, homepage: http://www.fsw.ruu.nl/ms/jh.

Met dank aan Timo Bechger en Catharina Hartman voor hun commentaar op een eerdere versie.

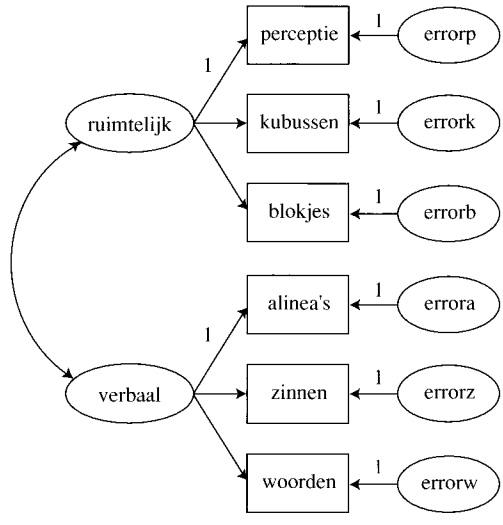
Voorbeelden van structurele modellen

Confirmatieve factoranalyse

Stel dat een onderzoeker beschikt over een intelligentietest die bestaat uit zes deeltests, waarvan er drie 'ruimtelijk inzicht' meten, en drie 'verbale intelligentie.' Vertaald naar een factoranalyse, houdt dat in dat er zes geobserveerde variabelen zijn (de zes deeltests), en twee latente variabelen of factoren (de twee achterliggende intelligentiefactoren). Figuur 1 bevat het paddiagram van dit factormodel. (Het is een voorbeeld uit het Amos-manual, Arbucle, 1997, p. 375).

De geobserveerde variabelen zijn de scores op de deeltests voor 'visuele perceptie', 'kubussen', 'blokjes' (drie nonverbale tests), 'alinea's herkennen', 'zinnen afmaken', en 'woordbetekenissen' (drie verbale tests). In het model van de factoranalyse wordt ervan uitgegaan, dat deze zes geobserveerde variabelen veroorzaakt worden door de twee onderliggende factoren ruimtelijk inzicht en verbale intelligentie. In de figuur wordt dit aangegeven door de pijlen die getrokken zijn van de factoren naar de geobserveerde variabelen. De dubbelzijdige pijl tussen de beide factoren geeft aan, dat we aannemen dat deze beide factoren gecorreleerd zijn, maar hier wordt geen oorzakelijke richting gespecificeerd. In het statistische model zijn de enkelzijdige pijlen *padcoëfficiënten*, die een verondersteld causaal verband aangeven. Het zijn eigenlijk regressiecoëfficiënten, die aangeven hoe goed een variabele voorspeld kan worden uit een andere. De dubbele pijlen zijn eenvoudig *correlaties*.

De factoren 'errorp' tot en met 'errorw' rechts in figuur 1 zijn residuele meetfouten. We nemen namelijk aan dat de geobserveerde variabelen nooit helemaal kunnen worden voorspeld uit de twee latente intelligentiefactoren. Daarom wordt voor elke geobserveerde variabele nog een residuele meetfout gespecificeerd. Deze meetfout wordt opgevat als het gevolg van de onbetrouwbaarheid van de geobserveerde variabele. Residuele meetfouten zijn niet ge-



Figuur 1. Confirmatieve factoranalyse op zes deeltests voor intelligentie: tweefactormodel

observeerd, we nemen hun bestaan aan omdat we de geobserveerde variabelen niet perfect kunnen voorspellen uit de factoren waar zij van afhangen. Omdat de residuele meetfouten ook latente factoren zijn, geeft het programma AMOS (dat gebruikt is om figuur 1 te maken) die ook consequent aan met een cirkel of ellips. Omdat de meetfoutfactoren echter inhoudelijk niet interessant zijn, worden ze in het paddiagram dikwijls weergegeven met alleen een pijl, zonder de bijbehorende cirkel (het programma LISREL doet het op deze manier).

Stel dat onze onderzoeker beschikt over empirische gegevens, in de vorm van de scores van 73 meisjes op de zes tests. De veronderstelde factorstructuur kan dan onderzocht worden met exploratieve factoranalyse. In exploratieve factoranalyse, zoals uitgevoerd door het programma FACTOR in SPSS, worden de factorladingen van alle geobserveerde variabelen op alle factoren geschat. Achteraf probeert de onderzoeker de factoren zo goed mogelijk te interpreteren. Wanneer er twee factoren uit de analyse komen, één met hoge ladingen van de verbale tests, en één met hoge ladingen van de nonverbale tests, dan kunnen we concluderen dat de gegevens onze hypothese ondersteunen dat er sprake is van een verbale en een ruimtelijke factor. Deze interpretatie is echter volledig subjectief. Figuur 1 beschrijft daarentegen een aantal zeer expliciete hypothesen over de factorstructuur. De theorie achter figuur 1 stelt kennelijk dat de drie ruimtelijke tests alléén ladingen mogen hebben op de ruimtelijke factor, en de drie verbale tests alléén op de verbale factor. Figuur 1 impliceert daarmee een aantal strenge restricties, namelijk dat er géén ladingen zijn van de ruimtelijke tests op de verbale factor en van de verbale tests op de ruimtelijke factor. Omdat de onderzoekers voor ze aan de analyse beginnen expliciete hypothesen opstellen over de factorstructuur, kan er nu gebruik worden gemaakt van een *confirmatieve factoranalyse*. Het factormodel in figuur 1 is dus niet een resultaat, dat uit de analyses komt rollen, maar een model dat voorafgaand aan de analyses moet worden opgesteld. Zo'n model is doorgaans gebaseerd op een combinatie van theoretische overwegingen en eerdere onderzoeksresultaten. De houdbaarheid van het model als geheel, en de significantie van de verschillende paden in het paddiagram, kunnen dan statistisch getoetst worden.

Voor het model in figuur 1 beschikken we over gegevens, namelijk de scores van 73 meisjes op de zes deeltests. Net als bij exploratieve factoranalyse is het uitgangspunt van de analyse de matrix met correlaties tussen de variabelen. Wanneer we deze correlaties berekenen, verkrijgen we de correlatiematrix in tabel 1.

Uit de correlaties in tabel 1 blijkt al, dat de drie deeltests die bij de verbale factor horen, onderling hoog correleren. De correlaties die bij de

Tabel 1. Correlaties tussen de zes intelligentie-deeltests

	woorden	*zinnen*	*alinea's*	*blokjes*	*kubussen*	*perceptie*
woorden	1,00					
zinnen	0,70	1,00				
alinea's	0,74	0,72	1,00			
blokjes	0,37	0,34	0,33	1,00		
kubussen	0,18	0,18	0,21	0,49	1,00	
perceptie	0,23	0,37	0,34	0,49	0,48	1,00

ruimtelijke factor horen, correleren eveneens hoog, maar niet zo hoog als de verbale deeltests. In de tabel zijn deze correlaties vet weergegeven. De correlaties tussen de twee verschillende soorten deeltests zijn het laagste. Op het oog is er dus reeds een zekere bevestiging van de veronderstelde tweefactorstructuur.

De analyse van het structurele model houdt in dat schattingen berekend worden voor de parameters in het model: de factorladingen, de correlatie tussen de beide factoren, en de varianties van de residuele meetfouten. Voor elke geschatte parameter wordt ook de standaardfout berekend, waarmee per parameter afzonderlijk getoetst kan worden of die parameter significant afwijkt van nul. Daarnaast wordt de passing van het model als geheel op de gegevens met een chi-kwadraat toets statistisch getoetst. Deze chi-kwadraat toets is in feite een toets van de discrepanties tussen het veronderstelde model en de empirische gegevens. Op basis van het factormodel in figuur 1 moet het mogelijk zijn de correlaties in tabel 1 nauwkeurig terug te rekenen. Wanneer het resultaat van de chi-kwadraat toets significant is, geeft dit aan dat de discrepanties tussen de teruggerekende correlaties en de empirische correlaties in tabel 1 significant zijn, zodat het model verworpen moet worden. Als de toets niet significant is, dan zijn de discrepanties verwaarloosbaar klein, en wordt het model daarom (voorlopig) aanvaard. Er zijn ook speciale maten ontwikkeld om de passing van een model bij de gegevens te evalueren. De bekendste is een maat die door Jöreskog ontwikkeld is, en die eenvoudig Goodness-of-Fit Index (GFI) heet. De GFI is een maat die kan lopen van nul tot één; doorgaans wordt van een acceptabele passing gesproken wanneer de GFI minstens 0,90 is. Er zijn nog vele andere goodness-of-fit maten, die verderop meer gedetaileerd besproken zullen worden.

Voor het model in figuur 1 levert de statistische toets een chi-kwadraat op van 7,8, met acht vrijheidsgraden, en een p-waarde van 0,45. De passingsmaat GFI is gelijk aan 0,97, wat vrij hoog is. Het model wordt dus niet verworpen, en op basis van de GFI mogen we concluderen dat het tweefactormodel in figuur 1 een goed passend model is voor de gegevens. Dit is een belangrijke uitkomst, want het betekent dat het zinvol is om de geschatte factorladingen en dergelijke te interpreteren. Wanneer het model verworpen was, was het niet zinvol geweest de verdere resultaten te interpreteren. We zouden dan immers factorladingen enzovoort interpreteren van een model dat sterk afwijkt van de empirische gegevens. De interpretatie van de resultaten van een structureel model (SEM) begint dus altijd met het bepalen of het

Tabel 2. Factorladingen en standaardfouten (tussen haakjes) in het confirmatief factormodel

	Gestandaardiseerde lading		
	Ruimtelijk	*Verbaal*	*Z-waarde*
Perceptie	0,70 (0,12)		5,74
Kubussen	0,65 (0,12)		5,32
Blokjes	0,74 (0,12)		6,01
Alinea's		0,88 (0,10)	8,96
Zinnen		0,83 (0,10)	8,22
Woorden		0,84 (0,10)	8,42

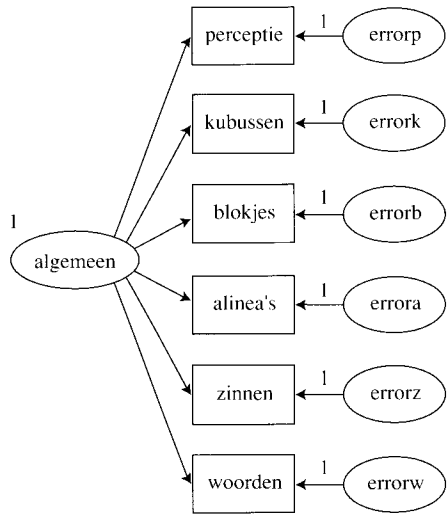
theoretische model goed bij de gegevens past. Als die passing slecht is, wat onder andere blijkt uit een significante chi-kwadraat, dan moet het model niet worden geïnterpreteerd, maar worden gewijzigd, totdat een model is gevonden dat wel goed bij de gegevens past. Op het aanpassen van een model wordt later meer gedetailleerd ingegaan.

De geschatte factorladingen van het confirmatieve factoranalyse model staan in tabel 2, met tussen haakjes de geschatte standaardfouten.

De standaardfouten (tussen haakjes weergegeven) kunnen worden gebruikt voor een toets van de significantie van de factorlading. De toetsingsgrootheid is Z = lading/standaardfout, waarvan de significantie bepaald wordt op basis van de standaard-normale verdeling. De kritische waarde van Z voor een tweezijdige toets met een alfa van 0,01 is 2,58. De Z-waarden voor de ladingen in tabel 2 zijn alle veel groter, dus alle ladingen zijn significant op een significantieniveau van één procent. De correlatie tussen de beide factoren is 0,49 (standaardfout 0,12, Z-waarde 4,12). De correlatie tussen de beide factoren is significant, maar niet extreem hoog.

De confirmatieve factoranalyse bevestigt hiermee de hypothesen die ten grondslag liggen aan het model in figuur 1. De factorladingen zijn hoog en ze zijn alle significant. De correlatie tussen beide factoren is eveneens significant, maar de correlatie is niet zo hoog dat de beide factoren evengoed samengenomen kunnen worden. Ten slotte, de algemene chi-kwadraat toets verwerpt het model niet. Dat wil in feite zeggen dat het model goed bij de gegevens past.

De uitkomsten van een exploratieve factoranalyse op de correlaties in tabel 1 zullen niet tot andere interpretaties leiden dan de confirmatieve factoranalyse. De confirmatieve factoranalyse levert echter meer informatie. Allereerst weten we op basis van de chi-kwadraat toets dat twee factoren inderdaad voldoende zijn. Verder hebben we niet alleen een factormatrix, maar ook significanties van de ladingen. We weten in ons geval, dat alle ladingen significant zijn. Ten slotte, we kunnen alternatieve modellen zeer expliciet toetsen. Bijvoorbeeld, stel dat er een alternatief factormodel is, dat stelt dat er slechts



Figuur 2. Confirmatieve factoranalyse op zes deeltests voor intelligentie: eenfactormodel

één algemene intelligentiefactor is. Het paddiagram van dit model staat in figuur 2.

Over het model in figuur 2 hoeven we niet lang na te denken. De modeltoets leidt tot een chi-kwadraat van 40,8, met negen vrijheidsgraden en een p-waarde kleiner dan 0,001. De discrepantie tussen het model en de gegevens is ruim significant, en het model is daarmee verworpen. De passingsmaat GFI is 0,83, en dat is veel te laag. Ook de GFI geeft dus aan dat het model niet goed past. De conclusie is dat één enkele intelligentiefactor niet voldoende is om het patroon van correlaties in tabel 1 te verklaren. De theorie dat achter de zes intelligentietests een enkele intelligentiefactor schuilgaat, is daarmee verworpen.

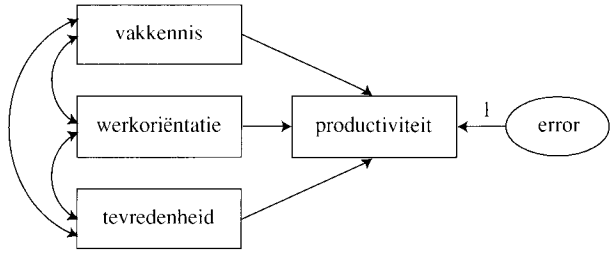
Multipele regressieanalyse

Wanneer we uitsluitend met geobserveerde variabelen werken, kunnen we met behulp van SEM een multipele regressieanalyse uitvoeren. Figuur 3 bevat het paddiagram van een multipele regressieanalyse met één afhankelijke variabele, en drie predictoren (eveneens een voorbeeld uit de Amos-handleiding, Arbuckle, 1997, p. 323). In deze multipele regressieanalyse wordt de productiviteit van arbeiders voorspeld uit de predictoren vakkennis, werkoriëntatie, en tevredenheid.

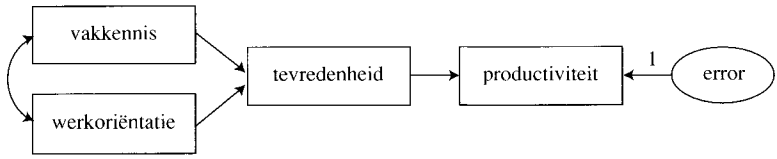
Figuur 3 maakt twee veronderstellingen expliciet, waar bij multipele regressieanalyse via bijvoorbeeld SPSS niet snel bij wordt stilgestaan. Allereerst is het niet voldoende om de paden van de predictoren naar de afhankelijke variabele weer te geven. Er moet ook worden aangegeven dat verondersteld wordt dat de predictoren onderling gecorreleerd zijn; dat wordt in het paddiagram aangegeven door de dubbelzijdige pijlen. Verder blijkt de residuele voorspellingsfout (error) een latente factor te zijn.

De parameters van het model in figuur 3 zijn de regressiecoëfficiënten en de variantie van de residuele voorspellingsfout. SEM programma's leveren daarnaast ook nog de gekwadrateerde multipele correlatie. Net als bij 'gewone' regressieanalyse is er sprake van ruwe regressiegewichten en van gestandaardiseerde gewichten (de beta-gewichten). Bij SEM analyse krijgen we de gestandaardiseerde gewichten als we de analyse uitvoeren op een correlatiematrix. De ongestandaardiseerde gewichten krijgen we als we de analyse uitvoeren op een covariantiematrix; de meeste programma's kunnen als extra uitvoer dan ook de gestandaardiseerde beta-gewichten leveren.

Het gebruik van SEM om multipele correlaties te berekenen heeft geen



Figuur 3. Multipele regressiemodel ter voorspelling van productiviteit in de notatie van een SEM paddiagram



Figuur 4. Padmodel voor tevredenheid en productiviteit in de notatie van een SEM paddiagram

speciale voordelen. Het onderscheid tussen SEM en gewone multipele regressie is daarin gelegen, dat we met SEM ook ingewikkeldere padmodellen kunnen opstellen en toetsen. Dat kunnen regressiemodellen zijn met meer dan één afhankelijke variabele, of met een interveniërende variabele tussen de predictoren en de afhankelijke variabele. Een voorbeeld daarvan is te vinden in figuur 4. Het model in figuur 4 is een padmodel met een interveniërende variabele. Dit is een model dat met gewone multipele regressie niet simultaan geschat kan worden. We kunnen hoogstens in twee aparte analyses eerst tevredenheid voorspellen uit vakkennis en werkoriëntatie, en vervolgens productiviteit voorspellen uit de tevredenheid. Met SEM kan het model als een geheel geschat worden.

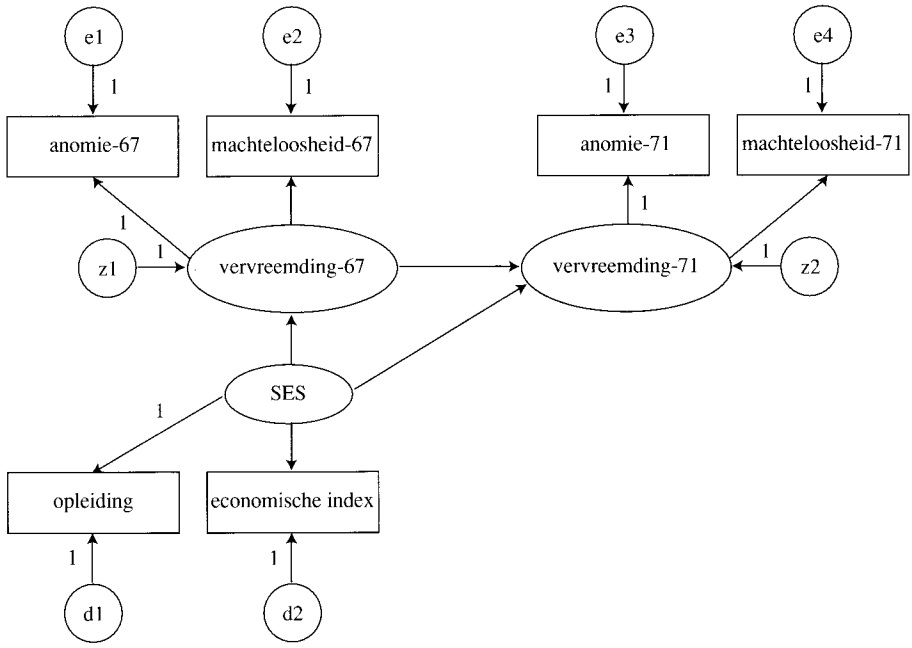
Wanneer we het model in figuur 4 analyseren met de beschikbare empirische gegevens (de gegevens zijn te vinden in het Amos manual), dan blijkt dat het model in figuur 4 leidt tot een chi-kwadraat van 1,7, met twee vrijheidsgraden en een overschrijdingskans van 0,43. Het model is daarmee niet verworpen. De gekwadrateerde multipele correlatie voor de afhankelijke variabele 'tevredenheid' is echter slechts 0,01. Als voorspellingsmodel is het dus toch weinig geslaagd. Het model past goed bij de gegevens, maar we moeten constateren dat uit het model ook blijkt dat tevredenheid slecht te voorspellen is uit vakkennis en werkoriëntatie.

Een padmodel met latente factoren

Structurele modellen kunnen zeer gecompliceerd zijn, met combinaties van zowel geobserveerde variabelen als van latente factoren. Een voorbeeld van een padmodel voor herhaalde metingen, dat zowel geobserveerde variabelen als latente factoren bevat, staat in figuur 5.

Het betreft hier een model voor de stabiliteit in de tijd van het sociologische begrip vervreemding. Het model in figuur 5 bestaat uit twee componenten. Allereerst hebben we het *meetmodel*, dat aangeeft dat er een latente factor 'vervreemding' is, die gemeten wordt door twee geobserveerde variabelen: 'anomie'en 'machteloosheid.' 'Vervreemding' is tweemaal gemeten, eenmaal in 1967 en eenmaal in 1971. De Sociaal-Economische Status (SES) is eveneens een latente factor, gemeten via de geobserveerde variabelen 'opleidingsniveau' en 'economische index.' De meetmodellen zijn eigenlijk confirmatieve factormodellen, waarin aangegeven staat hoe de latente factoren gemeten worden door de geobserveerde variabelen.

De andere component in figuur 5 is het *structurele model*, waarin gepoogd wordt de vervreemding in 1971 te voorspellen uit de SES en de vervreemding in 1967. Met name de padcoëfficiënt die loopt van vervreemding-67 naar de vervreemding-71 is inhoudelijk interessant, want die geeft aan hoe



Figuur 5. Paddiagram met latente factoren voor de stabiliteit van vervreemding

stabiel de vervreemding is over de tijd. Voor sociologen die geïnteresseerd zijn in de invloed van de SES op gevoelens van vervreemding zal ook het pad van SES naar vervreemding-67 inhoudelijk interessant zijn.

De empirische gegevens zijn oorspronkelijk van Wheaton Muthén, Alwin en Summers (1987), en te vinden bij de AMOS voorbeelden (Arbucle, 1997, p. 229). Het model in figuur 5 blijkt slecht bij de gegevens te passen: de chi-kwadraat is 71,5, bij zes vrijheidsgraden, en de overschrijdingskans kleiner dan 0,001. Het model moet dus verworpen worden, en interpretatie van de parameters (de factorladingen en padcoëfficiënten) is niet erg zinvol. We moeten in dit geval allereerst proberen het model te verbeteren tot de passing wel acceptabel is. Hoe we daarbij te werk kunnen gaan, wordt in een volgende paragraaf uiteengezet.

Enige technische aspecten van de parameterschatting en passing van het model

Bij het schatten van de parameters van structurele modellen wordt er doorgaans van uitgegaan, dat de geobserveerde variabelen een *normale verdeling* hebben. De berekeningsmethode is gebaseerd op een statistische techniek die bekend staat als Maximum Likelihood (ML) schatting. De berekeningen worden uitgevoerd met gespecialiseerde programma's, waarvan de bekendste zijn (in alfabetische volgorde) AMOS, EQS, en LISREL. (Meer gedetailleerde programma-informatie staat aan het eind van dit artikel.). De Maximum Likelihood methode veronderstelt niet al te kleine steekproeven. Methodologisch onderzoek heeft uitgewezen dat wanneer de gegevens inderdaad een normale verdeling hebben, een steekproef van minstens 200 respondenten doorgaans

voldoende is (Hoogland & Boomsma, 1997). Als de gegevens *niet* normaal verdeeld zijn, is er een alternatieve schattingsmethode beschikbaar. Deze staat bekend als de Asymptotically Distribution Free (ADF) schatter; in LISREL heet deze schatter WLS (Weighted Least Squares). Een nadeel van ADF (WLS) schatting is, dat zeer grote steekproeven nodig zijn, in sommige gevallen meer dan 5000 respondenten (vergelijk Chou & Bentler, 1995; Hoogland & Boomsma, 1997). Bij gegevens waarvan de afwijking van normaliteit niet al te groot is, levert de Maximum Likelihood methode desondanks goede schattingen op, maar dan zijn wel grotere steekproeven nodig, doorgaans minstens 400 respondenten (Hoogland & Boomsma, 1997). Daarnaast is het mogelijk bij niet-normaal verdeelde gegevens de chi-kwadraat te corrigeren. Deze correctie, de Satorra-Bentler correctie, blijkt volgens simulatieonderzoek de meest aantrekkelijke optie om het verstorende effect van niet-normaal verdeelde variabelen op de chi-kwadraat en te verdisconteren (Chou & Bentler, 1995; Hoogland & Boomsma, 1997). Recente versies van de computerprogramma's hebben de Satorra-Bentler correctie ingebouwd, aangevuld met correcties voor de standaardfouten. Een nadeel van deze correctie is dat de analyse op ruwe data moet plaatsvinden; het invoeren van alleen een correlatiematrix is niet voldoende.

Een ander probleem is de analyse van *categoriale gegevens*. De verschillende schattingsmethoden veronderstellen alle dat de variabelen continu zijn. Veel variabelen zijn echter categoriaal. Met categoriale gegevens worden gegevens bedoeld zoals bijvoorbeeld de antwoorden op een twee- of vijfpuntsschaal, waarbij bovendien de antwoordschaal eerder als ordinaal dan als interval meetniveau opgevat wordt. Wanneer er sprake is van een groot aantal categorieën, zoals bij een somscore op een test, is dat niet echt een probleem. We hebben dan zoveel categorieën, dat de variabele gerust als een continue variabele opgevat mag worden. Anders is dat bij de analyse van afzonderlijke vragen, die bijvoorbeeld met een twee- of vijfpuntsschaal gemeten zijn. De gebruikelijke aanpak is, om zulke categoriale gegevens te beschouwen als een ordinale meting van een variabele die een eigenlijk een continue en normale verdeling heeft. Bijvoorbeeld, bij een vijfpunts attitudevraag nemen we aan dat de gemeten attitude in werkelijkheid een normale verdeling heeft, die echter door de vijfpuntsvraag op een grove manier gemeten wordt. Gegeven deze aanname, is het mogelijk om op basis van de correlatie tussen twee categoriale variabelen, te schatten wat de correlatie tussen de achterliggende normaal verdeelde variabelen is. Bij dichotome variabelen heet deze correlatie de tetrachorische correlatie, bij meer dan twee antwoordcategorieën spreken we van een polychorische correlatie. Bij categoriale gegevens wordt de analyse vervolgens op de polychorische correlaties uitgevoerd. Een probleem daarbij is, dat hiervoor de ADF (WLS) methode gebruikt moet worden, die grote steekproeven vereist. Met name bij twee- en driepuntsvariabelen worden de polychorische correlaties erg onnauwkeurig geschat, en zijn te kleine steekproeven dus een probleem. Wanneer het aantal antwoordcategorieën minstens vijf is, en de frequentieverdelingen weinig afwijken van een normale verdeling, levert ook een analyse op gewone correlaties goede resultaten op (Chou & Bentler, 1995).

Voor een deel betreft het hier problemen die te voorzien zijn, en bij de opzet van het onderzoek zoveel mogelijk voorkomen moeten worden. Bij onderzoek waarbij het gebruik van SEM technieken voorzien wordt, is het verstandig van meet af aan een steekproef van ongeveer 400 cases te plannen. Bij

de variabelen moeten variabelen met weinig antwoordcategoriën vermeden worden, bijvoorbeeld door met somscores te werken, of het aantal antwoordcategoriën in de vragenlijst te vergroten. Ten slotte loont het de moeite om met een recente versie van de software te werken; allerlei recente correctietechnieken voor niet-normale gegevens zijn in oudere versies van de software doorgaans nog niet beschikbaar.

De passing van het model: goodness-of-fit indexen

Een belangrijk kenmerk van structurele modellen is, dat er een statistische toets beschikbaar is om na te gaan of het model de gegevens goed beschrijft. Als de statistische toets significant is, zijn er niet-verwaarloosbare discrepanties gevonden tussen de geobserveerde correlaties en de correlaties die door het model voorspeld worden. Bij een significante uitkomst van de toets, dient het model dus verworpen te worden, en moeten de onderzoekers een ander model opstellen. Een probleem hierbij is echter, dat de uitkomst van een statistische toets onder andere afhangt van de grootte van de steekproef. Bij een kleine steekproef zijn tamelijk grote discrepanties nog niet significant, en worden dus tamelijk slechte modellen niet snel verworpen. Bij een grote steekproef zullen ook kleine discrepanties statistisch significant zijn, en worden ook vrij goede modellen verworpen.

Om dit probleem te omzeilen, zijn *goodness-of-fit* indexen voorgesteld, die aangeven hoe goed het model de gegevens beschrijft, zonder dat daarbij de steekproefgrootte meespeelt. Sommige goodness-of-fit indexen wegen daarbij ook de complexiteit van het model mee: bij gelijke passing op de gegevens krijgt een ingewikkeld model dan een lagere passings-index dan een eenvoudig model. Bij dergelijke passings-indexen zijn eenvoudige modellen in het voordeel vergeleken met ingewikkelder modellen. Moderne SEM-software produceert een enorm aantal goodness-of-fit indexen. De meest bekende zijn de door Jöreskog en Sörbom (1989) voorgestelde GFI (Goodness of Fit) en AGFI (Adjusted GFI). De GFI is een passings-index, en de AGFI is een aanpassing hiervan die de complexiteit van het model meeweegt. Andere bekende indexen zijn de door Bentler en Bonett (1980) voorgestelde NFI (Normed Fit Index) en NNFI (Non-Normed Fit Index). De NNFI is eerder voorgesteld door Tucker en Lewis (1973), en daarom wordt hij ook wel de TLI (Tucker-Lewis Index) genoemd. De NNFI (TLI) compenseert net als de AGFI ook voor de complexiteit van het model. In simulatieonderzoek is gebleken dat veel van de voorgestelde fit-indexen toch beïnvloed worden door de steekproefgrootte, of weer andere nadelen hebben. De bekende en veel gebruikte GFI en AGFI maten zijn bijvoorbeeld sterk afhankelijk van de grootte van de steekproef (hoewel dat uiteraard niet de bedoeling was), en daarom eigenlijk af te raden. De passingsmaat NNFI/TLI is eenvoudig te berekenen, en wordt slechts weinig beïnvloed door de steekproefgrootte of niet-normale verdelingen. Doorgaans wordt gesteld dat een passings-index van 0,95 of groter aangeeft dat het model goed past, en van kleiner dan 0,90 dat het model slecht past, en dus niet geïnterpreteerd kan worden. Een passingsmaat tussen 0,90 en 0,95 geeft aan dat het model aardig past, maar dat enige verbeteringen wenselijk zijn. Dit soort vuistregels zijn tamelijk grof, en kunnen bij verschillende indexen enigszins anders liggen.

Een tamelijk nieuwe aanpak is om er van uit te gaan dat *alle* modellen slechts benaderingen zijn, en dat een perfecte passing dus niet mogelijk is. De

vraag is hoe goed de benadering is. Daarvoor is een statistische index ontwikkeld, de RMSEA (Steiger, 1990), voor de Root Mean Square Error of Approximation. Als de benadering goed is, dan is de RMSEA klein. De gebruikelijke norm daarvoor is dat de RMSEA kleiner moet zijn dan 0,05. Het voordeel van de RMSEA is, dat het mogelijk is daarvoor een betrouwbaarheidsinterval te berekenen, waardoor getoetst kan worden of de RMSEA significant groter is dan de grens van 0,05. Deze toets heet de 'test for close fit'.

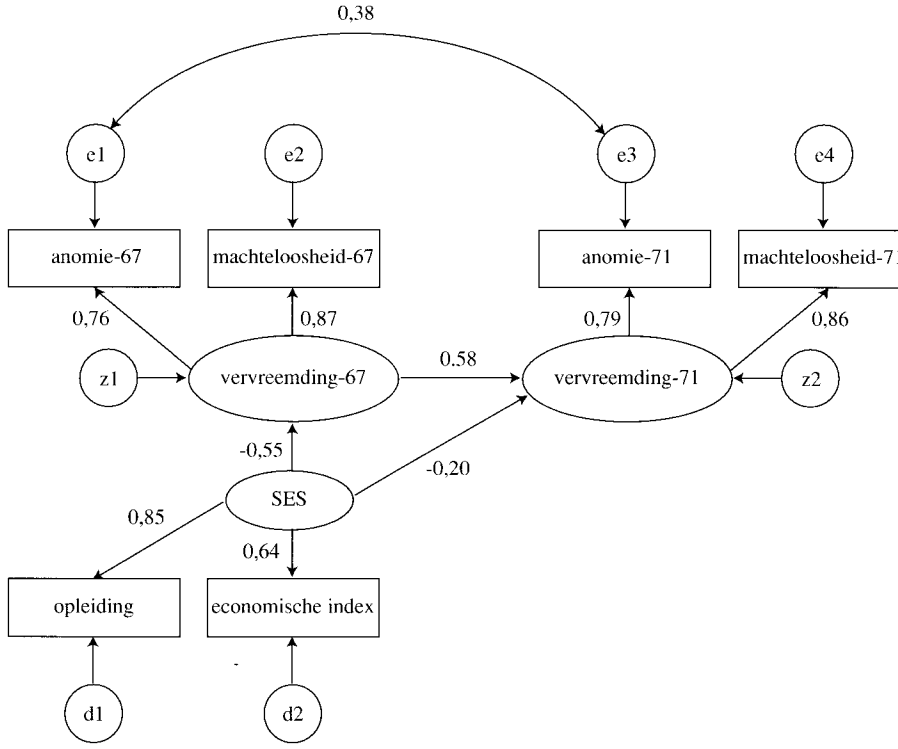
Een simulatieonderzoek dat de verschillende criteria vergelijkt (Hu & Bentler, 1999) leidt tot het advies standaard twee verschillende goodness-of-fit maten te vermelden. Een goede combinatie is de RMSEA met de TLI (met een gewenste waarde van minstens 0,95).

Het probleem van de grote steekproef zou een rol kunnen spelen bij de verwerping van het model voor de stabiliteit van vervreemding in figuur 5. De steekproefgrootte is hier N = 932, en dat is vrij groot. Voor het (statistisch verworpen) model in de figuur zijn de bovengenoemde fit-indexen als volgt: GFI = 0,98, AGFI = 0,91, NFI = 0,97, NNFI/TLI = 0,92, RMSEA = 0,11. De RMSEA is significant groter dan 0,05. Hoewel het model niet dramatisch slecht past, wijzen met name de AGFI en NNFI/TLI en de RMSEA erop, dat verbetering wenselijk is.

Aanpassing van het model en modificatie-indexen

Als de passing van een model niet goed is, kan geprobeerd worden het model aan te passen door niet-significante parameters te verwijderen, en extra parameters toe te voegen. Het verwijderen van niet-significante factorladingen en padcoëfficiënten kan eenvoudig gebeuren op basis van de standaardfout. Net niet-significante parameters kunnen beter in het model gehandhaafd worden. Een simpele vuistregel is dat parameters die kleiner zijn dan de bijbehorende standaardfout veilig uit het model geschrapt kunnen worden. Moeilijker is het om te bepalen welke factorladingen en padcoëfficiënten eventueel moeten worden toegevoegd. Moderne SEM-software biedt hiervoor een hulpmiddel. Voor elke parameter die op nul is gefixeerd (niet in het model is opgenomen) wordt een zogenaamde *modificatie-index* berekend. De modificatie-index is een schatting van de mate waarin de chi-kwadraat van de modeltoets verlaagd wordt als de betreffende parameter vrij geschat wordt, met andere woorden, wanneer de betreffende parameter wordt toegevoegd aan het model. Een mogelijke strategie om het model te verbeteren lijkt dan ook, om stapsgewijs steeds die parameter in het model op te nemen, die de grootste modificatie-index heeft. Dit is echter een zuiver statistische redenering. Beter is het om de parameters met de grote modificatie-indexen zorgvuldig te inspecteren, en alleen die parameters toe te voegen die theoretisch zinvol zijn. Blindelings de grootste modificatie-index volgen is geen goede strategie!

Een voorbeeld hiervan kan gegeven worden aan de hand van het model in figuur 5. De model-toets van dat model geeft als uitkomst een chi-kwadraat van 71,5 bij zes vrijheidsgraden. Er zijn een aantal grote modificatie-indexen, de grootste is 40,9, en deze suggereert een covariantie tussen de meetfouten e1 en e3, de meetfouten van de anomie-maat in 1967 en 1971. Dat is op zich een plausibele aanpassing van het model; het veronderstelt dat de test-hertest correlatie voor de anomie-maat groter is dan voor de machteloosheid-maat. Wanneer we deze modificatie uitvoeren, krijgen we een chi-kwadraat van 6,4



Figuur 6. Aangepast padmodel met latente factoren voor de stabiliteit van vervreemding, met gestandaardiseerde schattingen van de parameters

bij vijf vrijheidsgraden, met een overschrijdingskans van 0,27. Het gewijzigde model is statistisch plausibel. Niet alleen is de statistische toetsing niet meer significant, ook de passings-indexen zijn nu alle 'goed': GFI = 1,00, AGFI = 0,99, NFI = 1,00, NNFI/TLI = 1,00, RMSEA = 0,02.

Figuur 6 bevat de extra correlatie, en geeft in de figuur de gestandaardiseerde schattingen van de parameters.

De stabiliteit van de vervreemding over de tijd blijkt aanzienlijk te zijn: de gestandaardiseerde padcoëfficiënt is 0,58. Het opleidingsniveau is een betere indicator van de SES dan de economische index. De beide vervreemdingsmaten hebben een tamelijk vergelijkbare lading op de vervreemding.

Bij aanpassingen van het model op basis van modificatie-indexen kan natuurlijk op kans worden gekapitaliseerd. Elke eigenaardigheid van de steekproef wordt uitgebuit, en als er veel modificaties worden doorgevoerd is het gevaar niet denkbeeldig dat er zwaar geleund wordt op toevallige eigenaardigheden van de voorliggende steekproef. Het algemene advies is dan ook: niet teveel modificaties, en alleen modificaties die ook theoretisch interpreteerbaar zijn. Het probleem hierbij is dat (achteraf) bijna altijd wel een theoretische interpretatie kan worden gegeven. Uit simulatieonderzoek is gebleken dat het aanpassen via modificatie-indexen dikwijls niet tot het goede model leidt (Spirtes, Scheines & Glymour, 1991), en dat gemodificeerde modellen bovendien vaak slecht repliceerbaar zijn (MacCallum, 1989; MacCallum, Roznowski & Necowitz, 1992). Het is dus altijd nodig om zulke model-aanpassingen te

toetsen door replicatie of kruisvalidatie (Cudeck & Henley, 1991; MacCallum e.a., 1992), dat wil zeggen door een nieuwe analyse op nieuwe gegevens.

Complexe structurele modellen

In de loop van de tijd zijn de mogelijkheden van SEM analyses steeds verder uitgebreid. Twee uitbreidingen worden hier besproken: het opnemen van gemiddelden in het model, en multigroep-analyse. Het gaat hier om modellen die nogal complex zijn, en een goede beheersing van zowel de methodologie als de software veronderstellen. Op de details van de techniek ga ik daarom niet in, daarvoor verwijs ik naar de literatuur.

Het opnemen van gemiddelden in een SEM model is overigens betrekkelijk eenvoudig. We kunnen nu niet meer alleen de correlaties invoeren; we moeten nu ruwe gegevens gebruiken, of naast de correlatiematrix ook de gemiddelden en de standaardafwijkingen van de geobserveerde variabelen invoeren. Het structureel model wordt uitgebreid met factorgemiddelden voor de latente factoren, en om het verband te leggen tussen de latente factoren en de geobserveerde variabelen hebben we nu niet alleen regressiegewichten nodig (de factorladingen), maar ook voor elke geobserveerde variabele ook een regressie-intercept. Dit is alleen zinvol wanneer complexe datasets worden geanalyseerd, met herhaalde metingen of meerdere groepen. Bij modellen voor herhaalde metingen wordt bijvoorbeeld vaak een model geconstrueerd dat het verloop van de factor-gemiddelden over de tijd beschrijft.

Bij gegevens die bestaan uit verschillende groepen kunnen we deze simultaan analyseren. Deze zogenaamde *multigroep-analyse* kan gebruikt worden om te onderzoeken of hetzelfde model past in twee of meer verschillende groepen. Dit doen we door voor beide groepen hetzelfde model te specificeren, en vervolgens *gelijkheidsrestricties* op te leggen. De software krijgt daarmee de opdracht, om voor overeenkomstige parameters voor beide groepen dezelfde waarde te schatten. De chi-kwadraat van een model met gelijkheidsrestricties kan vergeleken worden met de chi-kwadraat van hetzelfde model zonder die gelijkheidsrestricties. Het model met de gelijkheidsrestricties heeft een grotere chi-kwadraat, met meer vrijheidsgraden. Het aantal vrijheidsgraden meer is gelijk aan het aantal opgelegde gelijkheidsrestricties. Wanneer het verschil tussen de beide chi-kwadraten significant is, dan is de passing van het model met de gelijkheidsrestricties significant slechter. De gelijkheidsrestricties zijn dan verworpen, het is niet aannemelijk dat in beide groepen precies hetzelfde model geldt.

Een voorbeeld van mogelijke gelijkheidsrestricties vinden we in het model voor vervreemding in figuur 6. De factorladingen voor vervreemding in 1967 en in 1971 lijken sterk op elkaar. Misschien is het mogelijk te stellen dat ze in beide jaren voor overeenkomstige variabelen gelijk zijn. We kunnen nu aan de software de opdracht geven voor deze factorladingen gelijke schattingen te maken. Dit is een gelijkheidsrestrictie tussen twee parameters binnen dezelfde groep. Dan hebben we niet langer vier factorladingen, maar nog slechts twee: de lading van anomie en de lading voor machteloosheid, die nu in beide jaren hetzelfde moeten zijn. Dat blijkt te kunnen. Het resulterende model heeft een chi-kwadraat van 7,16 bij zes vrijheidsgraden en een overschrijdingskans van 0,30. De passingsmaten GFI en NNFI/TLI zijn 1,00. Het model past goed, en heeft als voordeel boven het eerdere model (dat ook goed

paste) dat de vervreemding in beide jaren nu op *precies dezelfde wijze* geoperationaliseerd wordt.

Wanneer we een multigroep-model uitbreiden met gemiddelden, dan kunnen we ook gelijkheidsrestricties opleggen aan de factorgemiddelden. Als we dat doen, dan voeren we een test uit op de gelijkheid van gemiddelden, analoog aan een t-toets of een eenweg variantieanalyse. We voeren die test echter uit op de latente, ongeobserveerde factoren. Met een gewone t-test kan dat niet. De procedure is wat ingewikkeld, maar ze wordt gedetailleerd uitgelegd door Byrne, Shavelson en Muthén (1989), Sörbom (1982) en Werts (1979).

Bij wijze van voorbeeld een analyse op het factor-model in figuur 1. Figuur 1 beschrijft een model waarin zes intelligentietests worden verklaard door twee achterliggende factoren: een ruimtelijke en een verbale factor. De beschikbare gegevens zijn de covarianties en gemiddelden van 73 meisjes en 72 jongens. Tabel 1 bevat de resultaten voor het tweefactormodel bij alleen de meisjes. Het is echter ook mogelijk een multigroep-model te specificeren voor de meisjes en de jongens tezamen. Om de factoren te vergelijken, moeten we gelijkheidsrestricties opleggen aan de factorladingen. Daarmee kunnen we toetsen of we in beide groepen precies dezelfde factoren meten (vergelijk Byrne e.a., 1989). Wanneer dit model verworpen wordt, moeten we aannemen dat bij de meisjes andere achterliggende factoren aanwezig zijn dan bij de jongens. Dan is met name dè *begripsvaliditeit* van de intelligentiefactoren in het geding. De factoren zijn dan immers niet hetzelfde voor meisjes en voor jongens, en een vergelijking van de factorgemiddelden is dan uiteraard niet zinvol. Wanneer het model met de gelijkheidsrestricties goed past, kunnen we vervolgens de factorgemiddelden vergelijken. Een vergelijking van de modellen met en zonder gelijkheidsrestricties op de factorgemiddelden levert een toets op de gelijkheid van de factorgemiddelden voor de meisjes en de jongens.

We kunnen nog veel meer restricties opleggen. Zo kan het interessant zijn om te onderzoeken, of de varianties en de correlatie van de twee factoren in beide groepen gelijk zijn, of dat de varianties van de meetfouten gelijk zijn. Tabel 3 geeft een aantal maten voor de passing van een aantal van de besproken modellen.

De verschillende passingsindexen geven aan dat *alle* modellen een acceptabele passing hebben op de gegevens. Wanneer we de opeenvolgende modellen formeel vergelijken, door het verschil tussen de chi-kwadraten op significantie te toetsen, dan blijkt de laatste restrictie, die op de factorgemiddelden, het model significant te verslechteren. Het verschil tussen de beide chi-kwadraten is daar 8,2, met als verschil van vrijheidsgraden twee. Een chi-kwadraatwaarde van 8,2 bij twee vrijheidsgraden levert een overschrijdingskans op

Tabel 3. Passing van verschillende multigroepmodellen voor meisjes en jongens

Gelijkheidsrestricties op:	chi^2	*df*	*p*	*TLI*	*RMSEA*
a) intercepts + ladingen	22,6	24	0,54	1,00	0,00
b) plus meetfouten	27,1	30	0,62	1,00	0,00
c) plus factorvariantie en correlatie	30,2	33	0,61	1,00	0,00
d) plus factorgemiddelden	38,4	35	0,32	1,00	0,03

van 0,02. Bij een alfa van 0,05 is daarmee de nulhypothese verworpen, dat de factorgemiddelden van de meisjes en de jongens gelijk zijn.

Het beste model om te interpreteren is model (c). Model (a) past ook, maar model (c) is door al die gelijkheidsrestricties eenvoudiger, en daarom beter. De mogelijkheid om al deze restricites op te leggen is voor onderzoekers goed nieuws. In ons voorbeeld betekent de gelijkheid van de factorladingen en intercepts dat we een goede begripsvaliditeit hebben; beide factoren zijn bij de jongens op dezelfde manier gedefiniëerd als bij de meisjes. Daarnaast zijn ook nog de meetfouten gelijk; de tests meten dus in beide groepen met dezelfde nauwkeurigheid. De gelijkheid van de correlatie tussen de beide factoren is opnieuw een indicatie van een goede begripsvaliditeit. Het enige verschil voor jongens en meisjes is de latente factoren verbale vaardigheid en ruimtelijk inzicht.

We kunnen het verschil tussen de meisjes en de jongens nog wat verder uitpluizen door naar de geschatte factorgemiddelden en de bijbehorende standaardfouten te kijken. In model (c) zijn de factorgemiddelden voor de jongens gefixeerd op nul, en voor de meisjes geschat. De geschatte factor-gemiddelden voor de meisjes zijn: −1,26 (standaardfout 0,86) voor de ruimtelijke factor en 0,95 (standaardfout 0,52) voor de verbale factor. De Z-ratio voor het ruimtelijk gemiddelde is −1,26, en voor het verbale gemiddelde 1,82. Deze zijn beide kleiner dan de kritische waarde van 1,96. We kunnen concluderen dat jongens en meisjes niet verschillen wat betreft de ruimtelijke factor, maar dat meisjes misschien iets hoger scoren op de verbale factor. Een laatste verfijning is dat we een gelijkheidsrestrictie aanbrengen voor de gemiddelden van de ruimtelijke factor. Dat leidt tot een goed passend model (chi-kwadraat 31,8, df = 34, p = 0,58, TLI = 1,00, RMSEA = 0,00). In dit model wordt het gemiddelde op de verbale factor voor de meisjes geschat op 1,23 (standaardfout 0,47). Dit verschil is duidelijk significant, de p-waarde is 0,004.

Software en literatuur

Voor het analyseren van structurele modellen is allerlei software beschikbaar. Bekende programma's zijn LISREL en EQS, en het meer recente programma AMOS. Er zijn een aantal vergelijkende recensies over deze programma's: van Waller (1993), Hox (1995), en Kline (1998). De conclusie is dat elk van deze programma's goed is voor standaard analyses. Er zijn kleine verschillen, maar die zijn alleen van belang bij specialistische toepassingen. In oudere versies van LISREL worden de verschillende soorten parameters (regressiegewichten, covarianties, varianties van meetfouten) ondergebracht in verschillende matrixen, die met Griekse letters worden aangeduid. Het model wordt dan gespecificeerd door de matrixen te definiëren en specifieke elementen daarin te fixeren of vrij te laten schatten. Dat is tamelijk ingewikkeld. Vanaf de achtste versie van LISREL is ook een eenvoudige commando-taal beschikbaar om modellen te specificeren, SIMPLIS genaamd. In SIMPLIS wordt het model gespecificeerd door het te omschrijven in een soort pseudo-Engels. De programma's EQS en AMOS kennen een commandotaal die het model specificeert in de vorm van vergelijkingen. In recente versies van EQS en LISREL kan het model ook gespecificeerd worden door het paddiagram te *tekenen*. AMOS kent die mogelijkheid al langer. Dat is veel eenvoudiger, en het voordeel is dat het

paddiagram via standaard Windows knip- en plak-technieken in de tekst van het verslag kan worden ingevoegd.

Al is het specificeren van een structureel model eenvoudiger geworden, het blijft essentieel dat de gebruikers van de software begrijpen wat ze aan het doen zijn. Er zijn inmiddels een aantal handboeken die zich richten op beginnende SEM gebruikers. Goede introducties zijn de boeken van Loehlin (1998), Mueller (1996), en Schumacker en Lomax (1996). Een aantal speciale onderwerpen, zoals een bespreking van de verschillende goodness-of-fit maten, en analyse van niet-normaal verdeelde gegevens, zijn te vinden in een bundel geredigeerd door Hoyle (1995). Een verzameling artikelen over modellen voor multigroep gegevens en herhaalde metingen zijn te vinden in een bundel geredigeerd door Little, Schnabel en Baumert (1999). Een klassiek maar geavanceerd handboek is Bollen (1989). Een geavanceerde maar toch redelijk toegankelijke behandeling van verschillende speciale onderwerpen is te vinden in de bundel geredigeerd door Marcoulides en Schumacker (1996). Barbara Byrne heeft een serie boeken geschreven met een vast stramien voor de titel: Structural Equation Modeling with <PROGRAMMANAAM>. Ik geef daarvan de verwijzing expliciet niet; het is een doorlopend project, en geïnteresseerde lezers dienen eenvoudig het meest recente exemplaar op te sporen. Tot nu toe zijn verschenen versies voor LISREL (de matrixversie), LISREL+SIMPLIS, EQS en AMOS.

Er is één tijdschrift dat zich geheel richt op SEM toepassingen: 'Structural Equation Modeling, an interdisciplinary journal.' Methodologische tijdschriften die veel aandacht besteden aan SEM zijn 'Multivariate Behavior Research,' 'Psychological Methods' en 'Sociological Methods & Research.'

SEM op het Internet

Op Internet zijn een aantal interessante SEM activiteiten aanwezig. Een goed beginpunt voor starters is een FAQ (Frequently Asked Questions) document op de homepage van Ed Rigdon: http://www.gsu.edu/mkteer. Een ander startpunt is de homepage van de Europese (oorspronkelijk Duitse) SEM werkgroep op http://uni-muenster.de/SoWi/struktur (NB: de hoofdletters zijn verplicht), en de SEM page van Joel West op http://students.gsm.uci/joelwest/SEM. Op deze homepages staan weer verwijzingen naar andere locaties, informatie over boeken en artikelen, speciale modellen, besprekingen van software, enzovoorts.

Softwaremakers hebben tegenwoordig doorgaans homepages over hun producten. AMOS is te vinden op http://www.smallwaters.com, EQS op http://www.mvsoft.com, en LISREL http://www.ssi.com. Van alle drie is op de betreffende homepage ook een demonstratieversie van de software op te halen: alle voorbeelden uit dit artikel kunnen met deze demoversies geanalyseerd worden (de data worden meegeleverd met de AMOS demo). Een speciaal geval is het programma Mx, dat gratis is (maar ook moeilijk), en opgehaald kan worden op de homepage van Mike Neale: http://griffin.vcu.edu/mx. Vanuit Nederland kunnen de programma's het eenvoudigst besteld worden via ProGamma: http://www.gamma.rug.nl.

Er is ook een aan SEM gewijde discussielijst: SEMNET. Op deze discussielijst zijn een groot aantal mensen geabonneerd, zowel beginners als erkende grootheden. De inhoud van de discussies varieert van eenvoudige vragen over

software problemen tot lange verhalen over de filosofische implicaties van causale modellen. Hoe men zich op SEMNET abonneert is te vinden op Ed Rigdons homepage (http://www.gsu.edu/mkteer).

Literatuur

Arbucle, J. (1997). Amos user's guide. Chicago: Smallwaters.

Bentler, P.M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bollen, K.A. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Byrne, B.M., Shavelson, R.J. & Muthén, B. (1989). Testing the equivalence of factor covariance and mean structure: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456-466.

Chou, C.-P. & Bentler, P.M. (1995). Estimates and tests in structural equation modeling. In R.H. Hoyle (Ed.), *Structural equation modeling: concepts, issues and applications* (pp. 37-55). Newbury Park, CA: Sage.

Cudeck, R. & Henley, S.J. (1991). Model selection in covariance structure analysis and the 'problem' of sample size: a clarification. *Psychological Bulletin, 109,* 512-519.

Hoogland, J. & Boomsma, A. (1997). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26,* 329-367

Hoyle, R.H. (Ed.) (1995). *Structural equation modeling: concepts, issues and applications.* Newbury Park, CA: Sage.

Hox, J.J. (1995). Amos, Eqs and Lisrel for Windows: A comparative review. *Structural Equation Modeling, 2,* 79-91.

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Jöreskog, K.G. & Sörbom, D. (1989). *Lisrel 7: A guide to the program and applications.* Chicago: SPSS.

Kline, R, (1998). Software programs for structural equation modeling: Amos, Eqs and Lisrel. *Journal of Psychoeducational Assessment, 16,* 302-323.

Little, T., Schnabel, K.U. & Baumert, J. (Eds.) (1999). *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples.* Hillsdale, NJ: Erlbaum.

Loehlin, J.C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

MacCallum, R.C. (1989). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 533-541.

MacCallum, R.C., Roznowskei, M. & Necowitz, L.B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 114,* 490-504.

Marcoulides, G.A. & Schumacker, R.A. (Eds.) (1996). Advanced structural equation modeling: Issues and techniques. Hillsdale, NJ: Erlbaum.

Mueller, R. (1996). *Basic principles of structural equation modeling.* New York: Springer.

Schumacker, R.E. & Lomax, R.G. (1996). *A beginners guide to structural equation modeling.* Hillsdale, NJ: Erlbaum.

Sörbom, D. (1982). Structural equation models with structured means. In K.G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation.* Amsterdam: Elsevier.

Spirtes, P., Scheines, R. & Glymour, C. (1991). Simulation studies of the reliability of computer-

aided model specification using the Tetrad II, Eqs, and Lisrel. *Sociological Methods & Research*, *19*, 3-66.

Steiger, J.H. (1990). Structural modeling evaluation and modification: An interval estimation approach. *Multivariate Behavior Research*, *25*, 173-180.

Tucker, C. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10.

Waller, N.G. (1993). Software review. Seven CFA programs: EQS, EZPATH, LICS, LISCOMP, LISREL7, SIMPLIS, and CALIS. *Applied Psychological Measurement*, *17*, 73-100.

Werts, C.E. (1979). Confirmatory factor analysis applications: missing data problems and comparison of path models between populations. *Multivariate Behavior Research*, *14*, 199-213.

Wheaton, B., Muthén, B.O., Alwin, D.F. & Summers, G.F. (1987). Assessing reliability and stability in panel surveys. In D.R. Heise (Ed.), *Sociological Methodology 1977* (pp. 84-136). San Francisco: Jossey-Bass.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557-585.