Psychologische Beiträge, Band 29, 1987, S. 60–77

Möglichkeiten der Kreuzvalidierung von Strukturgleichungsmodellen

K. OPWIS[1], A. GOLD[2] und W. SCHNEIDER[3]

Zusammenfassung, Summary, Résumé

Die Beurteilung der Angemessenheit theoretischer Überlegungen auf der Grundlage statistischer Hypothesentests ist für die empirische Forschung von zentraler Bedeutung. Im Mittelpunkt der Arbeit stehen die mit der Testung von multivariaten Strukturgleichungsmodellen (LISREL-Modelle) verbundenen Probleme. Zu Beginn wird das LISREL-Modell unter den Aspekten der simultanen Analyse von Kovarianz- und Mittelwertstrukturen und der Modelltestung diskutiert. Anschließend werden anhand eines Illustrationsbeispiels aus der pädagogisch-psychologischen Forschung verschiedene Vorgehensweisen der Kreuzvalidierung einander gegenübergestellt und an empirischen Daten demonstriert. Es zeigt sich, daß die Übernahme exakter numerischer Parameterwerte sowohl aufgrund inhaltlicher wie auch formaler Argumente unangemessen streng erscheint. Diese Einschätzung kann empirisch belegt werden, indem eine Approximation von Intervallrestriktionen vorgenommen wird.

Cross-validating structural equation models: Problems and perspectives

This study focuses on the problem of how to judge the adequacy of theoretical models on the basis of statistical hypothesis testing procedures. In particular, problems of testing multivariate structural equation models (LISREL-models) are discussed. In a first step, the possibilities of simultaneously analyzing covariance and mean structures and of testing LISREL models are discussed. Next, data from a study conducted in the field of educational psychology are used to illustrate different procedures of cross-validating LISREL models. As it turns out, the usual restriction to base the model on equivalent numerical parameter values appears to be inappropriately rigorous. Thus, the suggestion is that cross-validations of structural equation models should rely on approximations of interval restrictions instead.

Possibilités de validation croisée pour des modèles d'équation structurale

Le jugement de l'appropriation des modèles théoriques basé sur des tests d'hythèses statistiques est d'une importance capitale pour la recherche empirique. Les

1 Dr. Klaus Opwis, Psychologisches Institut der Universität Freiburg, Niemensstr. 10, D–7800 Freiburg.
2 Dipl. Psych. Andreas Gold, Institut für Pädagogische Psychologie der Universität Frankfurt/Main, Senckenberganlage 15, D–6000 Frankfurt/Main.
3 Dr. Wolfgang Schneider, Max-Planck-Institut für Psychologische Forschung, Leopoldstr. 24, D–8000 München 40.

problèmes liés à la vérification de modèles multivariants d'équation structurale sont au centre de cette étude.

On trouve pour commencer une discussion sur le modèle LISREL sous l'aspect de l'analyse simultanée de structures de covariance et de valeur moyenne et sous celui de la vérification du modèle. Suit ne comparaison de différents procédés de validation croisée, démontrée d'après des données empiriques et illustrée par des exemples pris dans la recherche psychopédagogique. Il en ressort que la reprise de paramètres numériques exacts parait inadéquate dans sa rigueur en raison d'arguments concernant tant le fond que la forme. Cette évaluation peut être prouvée de façon empirique en faisant une approximation des restrictions d'intervalles. (A.-E. Posse-Douhaire)

構造方程式モデルの交差妥当化の可能性

統計的仮説検定に基づく理論的考察の妥当性の評価は，経験的研究にとって非常に重要である．多変量構造方程式モデル（LISRELモデル）の検定と関連した問題が本研究の中心となる．はじめに，LISRELモデルが共分散構造と平均値構造の同時分析とモデル検定の観点から考察される．つぎに，教育心理学的研究からの例を用いて，交差妥当化の異なる処理法が互いに比較され，経験的データについて説明される．正確なパラメータ値を内容的および形式的論拠に基づかせることは，不適切といってもいいほど厳密すぎると思われる．この評価は，区間制約の近似が行われるので，経験的に証明されうる．
（山下利之 Dr.T.Yamashita）

1. Einleitung

In den letzten Jahren hat der speziell mit den Namen JÖRESKOG und SÖRBOM verbundene LISREL-Ansatz zunehmend an Bedeutung für die sozialwissenschaftliche Modellbildung und Hypothesentestung gewonnen. Die Abkürzung LISREL („LInear Structural RELationship") bezeichnet einerseits ein statistisches Modell und andererseits das zugehörige Computerprogramm zur Durchführung der Parameterschätzungen und Modelltests. Die Vorzüge dieses Ansatzes gegenüber herkömmlichen statistischen Verfahren lassen sich auf drei Betrachtungsebenen aufzeigen:

1) Unter formal-statistischen Aspekten stellt er eine Generalisierung des allgemeinen linearen Modells (vgl. hierzu etwa TIMM, 1975) dar. Positiv hervorzuheben ist dabei im besonderen, daß neben einer einheitlichen Terminologie und Notation mächtige schätz- und testtheoretische Kalküle (Prinzip der Maximum-Likelihood-Schätzung; Methode des Likelihoodquotiententests) zur Verfügung gestellt werden. Sind notwendige Voraussetzungen, wie etwa eine multivariate Normalverteilung der in die Analyse eingehenden Variablen gegeben, ist die adäquate Behandlung einer Reihe bekannter statistischer Verfahren (uni- und multivariate Regressionsanalyse, rekursive und nicht-rekursive Pfadanalyse, konfirmatorische

Faktorenanalyse) als Spezialfälle gewährleistet (JÖRESKOG & SÖRBOM, 1984). Darüber hinaus können auch varianzanalytische Modelle als spezielle LISREL-Modelle formuliert und getestet werden (vgl. MÖBUS, 1986; OPWIS, 1986).

2) Vom inhaltlich-theoretischen Standpunkt aus betrachtet, bietet das LISREL-Modell vielfältige Möglichkeiten zur expliziten Formulierung, Abbildung und Prüfung von Hypothesen (a) über angenommene Effektmuster in den Beziehungen und Zusammenhängen und (b) über Mittelwertunterschiede bzw. -änderungen von Variablen. Beide Hypothesenarten können simultan und in sich ergänzender Weise, also integrativ behandelt werden (vgl. die Beiträge in MÖBUS & SCHNEIDER, 1986). Dabei kann insbesondere auch die für die Sozialwissenschaften fundamentale Meßfehlerproblematik in angemessener Weise berücksichtigt werden.

3) Unter pädagogisch-didaktischen Gesichtspunkten bietet die Auseinandersetzung mit LISREL den Vorteil, daß der Anwender das Verfahren selbst gestaltet: Man lernt Modelle zu konzipieren, Hypothesen zu formulieren und zu prüfen, sowie verstecke Annahmen zu explizieren.

Im Mittelpunkt dieser Arbeit stehen die mit der Testung von multivariaten Strukturgleichungsmodellen verbundenen Probleme (vgl. GREEN, 1977):

a) Können die beobachteten (Stichproben-)Daten durch das postulierte Modell angemessen reproduziert werden? (Anpassungsproblem)
b) Welche Auswirkungen haben Datenidosynkrasien auf die Parameterschätzungen? (Variabilitätsproblem)
c) Wie sensitiv reagieren die Prüfgrößen zur Beurteilung der Modellanpassung gegenüber Parameteränderungen? (Sensitivitätsproblem)

Aus der Diskussion dieser Fragen im Kontext von LISREL-Modellen wird abgeleitet, daß erst eine Kreuzvalidierung der geschätzten Modelle genauere Auskunft über ihre Angemessenheit geben kann, wobei unterschiedliche Vorgehensweisen möglich sind. Illustriert werden die Überlegungen anhand eines Beispiels aus dem Bereich der Schulleistungsforschung. Die mit Hilfe des LISREL VI Programms (JÖRESKOG & SÖRBOM, 1984) durchgeführten Analysen beruhen auf den Paneldaten zweier voneinander unabhängiger Schülerkohorten derselben Grundgesamtheit.

Der Arbeit liegt folgender Aufbau zugrunde: Zunächst wird eine erweiterte Notation des LISREL-Modells zur Analyse von Mittelwertstrukturen eingeführt und Fragen der Modelltestung erörtert. Anschließend wird als Illustrationsbeispiel aus der pädagogisch-psychologischen Forschung ein Schulleistungsmodell vorgestellt und in seinen wesentlichen Befunden diskutiert. Im letzten Teil werden Möglichkeiten der Kreuzvalidierung von Strukturgleichungsmodellen aufgezeigt und anhand eines zweiten Datensatzes demonstriert.

2. Die Analyse von Mittelwertstrukturen mit LISREL

Im einfachen LISREL-Modell (LONG, 1983; JÖRESKOG & SÖRBÖM, 1979, 1984) liegt das Augenmerk auf der Analyse der Kovarianzstruktur einer Gruppe von Personen. Allerdings ist eine Verallgemeinerung des einfachen Modells im Sinne einer zusätzlichen Einbeziehung von Mittelwertinformationen möglich, wenn die beobachteten Indikatorvariablen als Rohwerte in die Analyse eingehen, wovon die üblichen Darstellungen des allgemeinen linearen Modells ausgehen. Formal entsprechen Rohwerte den Abweichungswerten bzw. Momenten um Null. Die korrespondierende (Stichproben-)Momentenmatrix **M** hat bei zusätzlicher Einführung einer Dummyvariablen, die für alle Beobachtungen den konstanten Wert ‚1' besitzt, folgende Gestalt:

$$(1)\quad \mathbf{M} = \begin{bmatrix} \mathbf{S}_{yy'} + \bar{\mathbf{y}}\bar{\mathbf{y}}' & \mathbf{S}_{yx'} + \bar{\mathbf{y}}\bar{\mathbf{x}}' & \bar{\mathbf{y}} \\ \mathbf{S}_{xy} + \bar{\mathbf{x}}\bar{\mathbf{y}}' & \mathbf{S}_{xx'} + \bar{\mathbf{x}}\bar{\mathbf{x}}' & \bar{\mathbf{x}} \\ \bar{\mathbf{y}}' & \bar{\mathbf{x}}' & 1 \end{bmatrix}$$

mit: **y** = p-dimensionaler Vektor der beobachteten Indikatoren der latenten endogenen Variablen
x = q-dimensionaler Vektor der beobachteten Indikatoren der latenten exogenen Variablen
S = (Stichproben-)Kovarianzmatrix

Bei der Analyse von Momentmatrizen um Null mit Hilfe des LISREL-Programms geht man von der folgenden modifizierten Modelldarstellung aus (vgl. JÖRESKÖG & SÖRBOM, 1984):

$$(2)\quad \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \Gamma & \mathbf{B} \end{bmatrix} \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} + \begin{bmatrix} 1 \\ \kappa \\ \alpha \end{bmatrix} [1] + \begin{bmatrix} 0 \\ \xi - \kappa \\ \zeta \end{bmatrix}$$

$$(3)\quad \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \nu & \Lambda_x & 0 \\ \mu & 0 & \Lambda_y \end{bmatrix} \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} + \begin{bmatrix} \delta \\ \epsilon \end{bmatrix}$$

Dabei entsprechen B, Γ, Λ_x, Λ_y, ξ, η, ζ, x, y, δ und ϵ den im üblichen LISREL-Modell vorkommenden gleichnamigen Matrizen und Vektoren. Die zusätzlich auftretenden Parameter können als Regressionskonstanten bzw. „Lokalisationsparameter" interpretiert werden:

α = Vektor mit den Regressionskonstanten der Regression der η auf die ξ.

ν, μ = Vektor mit den Regressionskonstanten der Regression der x bzw. y auf die ξ bzw. η.

Für die Erwartungswerte der Modellvariablen gilt dann:

(3.1) $E(\xi) = \kappa$

(3.2) $E(\eta) = (I-B)^{-1}(\alpha + K_\kappa) := \tau$

(3.3) $E(x) = \nu + \Lambda_x \kappa$

(3.4) $E(y) = \mu + \Lambda_y \tau$

Allerdings sind die neu eingeführten Parametervektoren ohne weitere Restriktionen nicht identifiziert (vgl. SÖRBOM, 1982; OPWIS, 1986). Die Formulierung von Hypothesen über Mittelwertunterschiede ist für den Fall einer Mehrgruppenanalyse nur sinnvoll, wenn (a) bei allen Gruppen dieselben (bzw. vergleichbare) Indikatorvariablen beobachtet werden und (b) die Indikatorvariablen für alle Gruppen parallele Messungen der als zugrundeliegend angenommenen latenten Variablen darstellen.

Diese Voraussetzungen implizieren gruppeninvariante Meßmodellanteile (ν, μ, Λ_x, Λ_y sind für alle Gruppen als gleich zu spezifizieren), wodurch die Metrik der Erwartungswertvektoren bis auf den Nullpunkt eindeutig definiert ist. Die Festlegung des Nullpunktes kann durch eine Festlegung der Erwartungswerte einer beliebigen Gruppe vorgenommen werden, beispielsweise durch:

(4) $\kappa^{(1)} = \tau^{(1)} = 0$

Damit wird die erste Gruppe zur Referenzgruppe und die Erwartungswerte der übrigen Gruppen sind auf einer Differenzskala abgebildet, also identifiziert und als Unterschiede zur gewählten Referenzgruppe interpretierbar.

3. Die Testung von LISREL-Modellen

Unter der Annahme, daß die in die Analyse eingehenden Daten unabhängige Beobachtungen einer multivariaten Normalverteilung mit Mittelwertvektor μ und Kovarianzmatrix Σ sind, erhält man Parameterschätzungen – ihre Identifikation vorausgesetzt –, indem folgende ML-Schätzfunktion minimiert wird (vgl. MORRISON, 1967):

(5) $F = \ln|\Sigma| - \ln|S| + \text{Spur}(S\Sigma^{-1}) - (p+q)$

Auch für den Fall, daß die Matrizen Σ und $\mathbf{S}$ nicht Momente um die Mittelwerte (Kovarianzen) sondern Momente um Null enthalten, liefert die Minimierung von (5) weiterhin ML-Schätzungen (vgl. MÖBUS, 1986). Die statistische Testung von LISREL-Modellen beruht auf der Likelihoodquotienten-Technik. Die chi^2-verteilte Prüfgröße ist gegeben durch:

$$(6) \qquad \lambda = -2\ln (L_0/L_i) = N\, F_0(\hat{\Sigma}_0)$$

mit: L_i = Likelihood des ‚idiographischen' bzw. ‚saturierten' Modells („Alternativhypothese") mit dessen Hilfe eine beliebige positivdefinite Matrix erzeugt werden kann, also insbesondere eine solche, die der beobachteten Stichproben-Kovarianzmatrix exakt entspricht.

L_0 = Likelihood des ‚vollen' Modells („Nullhypothese"), das dem vom Anwender formulierten Modell entspricht.

N = Anzahl der Meßwertträger (Personen)

$F_0(\hat{\Sigma}_0)$ = Wert der ML-Schätzfunktion für das volle Modell, mit dessen Hilfe die Matrix $\hat{\Sigma}_0$ generiert wird.

Die Prüfgröße λ ist – bei hinreichend großem N – annähernd chi^2-verteilt, wobei die Anzahl der Freiheitsgrade durch die Differenz zwischen der Anzahl vorgegebener Datenelemente und zu schätzender Parameter festgelegt ist. Sie bildet die Grundlage der Entscheidung, ob die beobachteten Daten durch das postulierte Modell in statistisch angemessener Weise reproduziert werden können. Ein im Verhältnis zu den Freiheitsgraden hoher chi^2-Wert deutet darauf hin, daß dies nicht der Fall ist.

Dieser zentrale Aspekt des LISREL-Ansatzes war in den letzten Jahren häufig Gegenstand kritischer Anmerkungen. Es handelt sich um eine globale Anpassungsprüfgröße, deren zugrundeliegende Logik einem Hypothesentest der Art unspezifizierte Alternativ- vs. spezifizierte Nullhypothese entspricht (vgl. BENTLER & BONETT, 1980). Die unspezifizierte Alternativhypothese entspricht dem idiographischen Modell und die Nullhypothese dem vom Forscher postulierten Modell. Ziel des Modelltests ist somit die Beibehaltung der Nullhypothese, die aber gegen die ‚strengstmögliche' Alternativhypothese einer perfekten Anpassung getestet wird. In diesem Zusammenhang haben verschiedene Autoren (etwa BENTLER & BONETT, 1980; CUDECK & BROWNE, 1983) angeregt, das vom Forscher postulierte Modell gegen ein Alternativmodell mit möglichst wenigen Parametern zu testen, wie es auch sonst üblich ist. Allerdings kann weder eine statistisch noch eine inhaltlich fundierte Begründung für die Verwendung eines der vorgeschlagenen speziellen Modelle („model of no correlation"; „equicorrelation model"; „one factor model") gegeben werden.

Eine statistisch bedingte Eigenschaft der Prüfgröße selbst, die zu Interpretationsproblemen führt, ist ihre direkt proportionale Abhängigkeit von der Stichprobengröße (vgl. MUNCK, 1979). Dies hat zur Konsequenz, daß für den schätztheoretisch erforderlichen Fall großer Stichproben praktisch jedes Modell als statistisch unbefriedigend beurteilt werden muß. Umgekehrt gilt für den Fall relativ kleiner Stichproben, daß oftmals auch inadäquate Modelle aufgrund insignifikanter chi^2-Werte beibehalten werden können. Weitgehend unbekannt ist wie die Prüfgröße auf eine Verletzung der ihr zugrundeliegenden Voraussetzungen reagiert (vgl. hierzu etwa eine von BOOMSMA (1982) durchgeführte Simulationsstudie).

Die vorherrschende ‚Bewältigungsstrategie' zur Umgehung dieser Probleme bei der Beurteilung der Anpassungsgüte eines Modells besteht in ihrer Relativierung, indem (a) zusätzliche Gesichtspunkte und/oder (b) hierarchisch ineinander geschachtelte Modelle herangezogen werden. So liefert die LISREL-VI Programmversion eine Reihe ergänzender Informationen, wie den „goodness-of-fit" Index (GFI), das „root mean square residual" (RMR) oder die normalisierten Residuen, auf die unten noch genauer eingegangen wird. Im unter (b) genannten Fall wird weniger der einzelne chi^2-Wert eines speziellen Modells interpretiert, als vielmehr ein Vergleich der chi^2-Wertdifferenzen hierarchisch ineinander geschachtelter Modelle bevorzugt. Diese Differenzen sind wiederum annähernd chi^2-verteilt. Allerdings wird diese Art des Modellvergleichs in der Regel an demselben Datensatz vorgenommen, wodurch sich das Signifikanzniveau in unkontrollierter Weise verändert. Man hat es bei dieser Vorgehensweise mit einer explorativen Datenanalyse bzw. einem ‚Modellfitting' und nicht mit einer statistischen Hypothesentestung zu tun.

Angesichts der angesprochenen Probleme bei der Testung von LISREL-Modellen, besteht der überzeugendste Weg zur Prüfung der Angemessenheit eines postulierten Modells in der Durchführung einer Kreuzvalidierung. Dies gilt selbstverständlich nicht nur für LISREL-Modelle, sondern generell für statistische Modelle. Im folgenden werden verschiedene im Rahmen des LISREL-Ansatzes realisierbare Formen der Kreuzvalidierung vorgestellt und praktisch demonstriert. Hierzu wird das im nächsten Abschnitt einzuführende Schulleistungsmodell auf eine zweite unabhängige Schülerkohorte derselben Grundgesamtheit übertragen.

4. Ein Modell schulischer Leistungen als Anwendungsbeispiel aus der pädagogisch-psychologischen Forschung

Die den Berechnungen zugrundeliegenden Daten wurden im Zusammenhang mit der modellschulbegleitenden Forschung an einer kooperati-

ven Gesamtschule in Süddeutschland erhoben. Für insgesamt 218 Schüler einer ersten Kohorte lagen aus der Mitte der fünften Klassenstufe Leistungsbeurteilungen in Form von Zeugnisnoten für vier Schulfächer (Deutsch, Englisch, Mathematik, Naturlehre) vor. 18 Monate später wurden die Schulleistungen ein zweites Mal erfaßt. In der Grundschule, etwa neun Monate vor der ersten Schulleistungsmessung, hatten die Schüler u.a. die folgenden Fähigkeitstests zu bearbeiten:

- die Untertests 1+2 („Allgemeinbegabung"), 6 („Ratefähigkeit"), 3 und 4 („Schlußfolgerndes Denken") des Prüfsystems für Schul- und Bildungsberatung (PSB) von HORN (1969),
- den Untertest RS („Rechtscheiben") aus dem Allgemeinen Schulleistungstest für 4. Klassen (AST4) von FIPPINGER (1967),
- den CFT2 („Nichtverbale Grundintelligenz") nach Cattell in der deutschen Bearbeitung von CATTELL & WEISS (1972).

Anhand des Leistungsniveaus (A-, B-, C-Kurs), das die Schüler nach fünf weiteren Schuljahren, also zu Ende der neunten Klasse, aufwiesen, ließ sich die Gesamtstichprobe retrospektiv in drei Teilstichproben aufteilen. Die Leistungsanforderungen sind im A-Kurs besonders hoch, im C-Kurs am geringsten. Im Analysezeitraum während des 5. und 6. Schuljahres wurden die Schüler gemeinsam unterrichtet. Eine fächerübergreifende Differenzierung in A-, B-, und C-Kurs fand erst mit Beginn des 7. Schuljahres statt. Dieselbe Datenstruktur liegt für eine zweite, im darauffolgenden Schuljahr untersuchte Kohorte mit 215 Schülern vor. Das zur Beschreibung der Daten[4] angenommene Strukturgleichungsmodell ist in Abbildung 1 wiedergegeben.

Es ist so aufgebaut, daß exogene Fähigkeitskonstrukte („Verbale" und „Nicht-Verbale" Fähigkeiten) auf jeweils analoge Schulleistungskonstrukte („Sprachliche" und „Nicht-sprachliche" Schulleistung) wirken. Die als Indikatoren der Schulleistungskonstrukte wiederholt beobachteten Schulnoten beinhalten neben zufälligen auch fachspezifisch systematische Fehleranteile, die über die Zeit miteinander kovariieren (vgl. JÖRESKOG, 1979). Aufgrund der empirischen Evidenz zur Bedeutung kognitiver Fähigkeiten für den Schulerfolg (vgl. etwa KÜHN, 1983; PARKERSON, LOMAX, SCHILLER & WALBERG, 1984; SCHNEIDER & TREIBER, 1984) wird im Modell unterstellt, daß die Schulleistungen allein von korrespondierenden Fähigkeitskonstrukten abhängen. In Anwendung der Struktur eines Simplex-Modells (JÖRESKOG, 1970; WERTS,

4 Die entsprechend der Gleichung (1) aufgebauten Momentenmatrizen finden sich im Anhang.

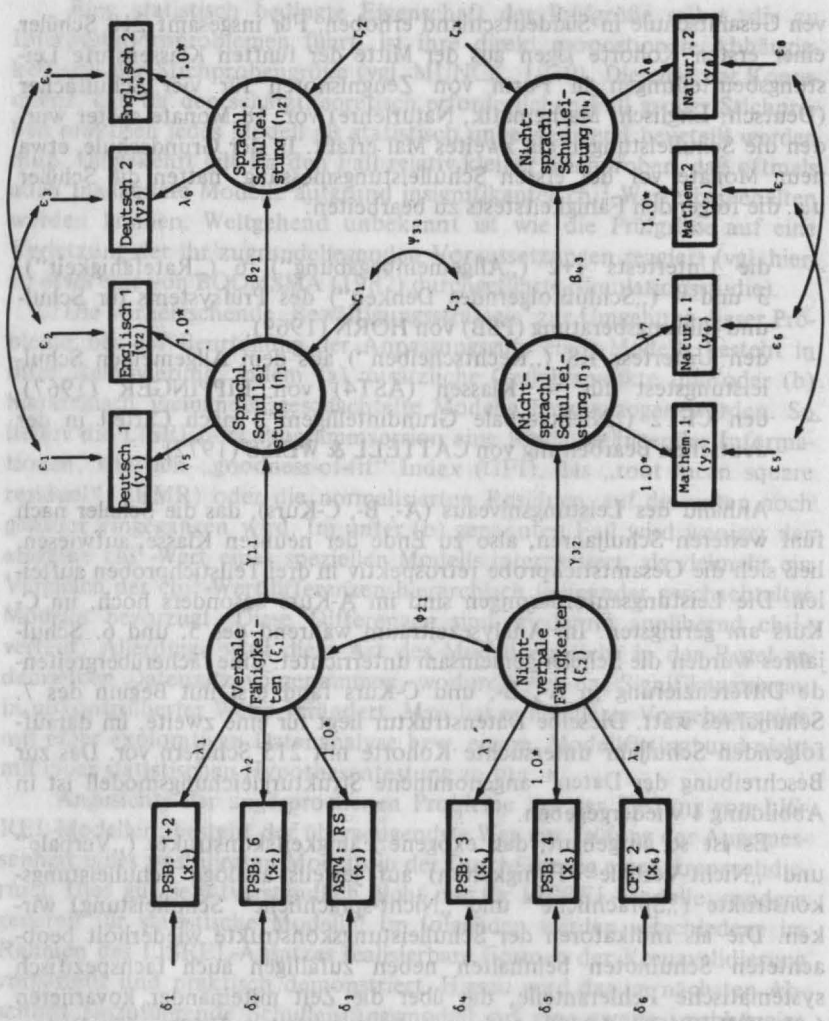

Abb. 1: Angenommenes Bedingungsmodell schulischer Leistungen.

LINN & JÖRESKOG, 1977) wird weiter angenommen, daß bedeutsame direkte Effekte auf die Schulleistungen zum zweiten Meßzeitpunkt nur für die in der fünften Klasse beobachteten Schulleistungen vorhanden sind. Die Kovariation der Gleichungsfehler zum ersten Meßzeitpunkt beinhaltet die gemeinsamen Varianzanteile der Schulleistungskonstrukte, die nicht durch die Fähigkeitskomponenten aufgeklärt werden.

Bei der durchgeführten simultanen Mehrgruppenanalyse wurde in den drei Teilstichproben (A-, B-, C-Kurs) ein identisches Meßmodell zugrundegelegt. Es resultierte eine globale Modellanpassung von $chi^2_{(df=233)} = 371.3$ ($p < .01$). Wie oben ausgeführt, stellt die im Sinne des Likelihood-quotienten signifikante Modellabweichung jedoch nur ein Kriterium zur Beurteilung der Anpassungsgüte von Modellen dar. Als zusätzliche Maße werden von JÖRESKOG & SÖRBOM (1984) der robustere „goodness-of-fit index" (GFI) und das „root mean square residual" (RMR) empfohlen. GFI gibt an, welcher Anteil der beobachteten Varianzen und Kovarianzen durch das Modell erklärt wird. RMR entspricht der Größe der durchschnittlichen Residualvarianzen und -kovarianzen und liefert beim Ver-

Tabelle 1:

Sequenz der untersuchten Modellvarianten

Modell-variante	Modellsemantik	Modell-spezifikation	Kriterien der Modellanpassung[1,2] χ^2-Prüfgröße	GFI	RMR
M 1	Test des Ausgangsmodells und seiner Annahmen (vgl. Text)	vgl. Abb. 1	$\chi^2_{(233)} = 371.3$	0.82 0.86 0.78	8.18 7.85 16.32
M 2	Test der „Zeitvarianz"-Hypothese der wiederholt verwendeten Meßinstrumente	$\mu_1 = \mu_3, \mu_2 = \mu_4,$ $\mu_5 = \mu_7, \mu_6 = \mu_8,$ $\lambda_5 = \lambda_6, \lambda_7, = \lambda_8,$ $\alpha_2^{(1)}, \alpha_4^{(1)}$ free,	$\chi^2_{(237)} = 429.0$	0.81 0.83 0.77	9,57 9.44 17.56
M 3	Test der „Mittelwert"-Hypothese: Die Mittelwerte der latenten exogenen Variablen sind in allen Gruppen untereinander gleich	$\kappa_1^{(g)} = \kappa_2^{(g)}$ $(g = {}_{2,3})$	$\chi^2_{(235)} = 375.1$	0.82 0.86 0.78	8.35 19.27 29.99
M 4	Test auf gruppeninvariante direkte Effekte der latenten exogenen auf die latenten endogenen Variablen	$\gamma_{11}^{(1)} = \gamma_{11}^{(2)} = \gamma_{11}^{(3)}$ $\gamma_{32}^{(1)} = \gamma_{32}^{(2)} = \gamma_{32}^{(3)}$	$\chi^2_{(239)} = 381.6$	0.82 0.86 0.77	9.49 19.50 31.11
M 5	Test auf gruppeninvariante direkte Effekte der latenten endogenen Variablen untereinander	$\beta_{21}^{(1)} = \beta_{21}^{(2)} = \beta_{21}^{(3)}$ $\beta_{43}^{(1)} = \beta_{43}^{(2)} = \beta_{43}^{(3)}$	$\chi^2_{(243)} = 389.3$	0.81 0.86 0.77	9.45 19.57 30.70

gleich alternativer Modelle am selben Datensatz Hinweise auf deren relative Anpassungsgüte. Die absoluten RMR-Werte sind allerdings nur unter Berücksichtigung der Größenordnungen in der zugrundegeliegenden Ausgangsmatrix interpretierbar.

Auch die chi^2-Prüfgröße wurde von uns in Übereinstimmung mit JÖRESKOG & SÖRBOM (1984) primär zur Beurteilung der relativen Modellanpassung bei ineinander geschachtelten Modellen verwendet. In Tabelle 1 ist die Sequenz untersuchter Modellvarianten unter formalen und inhaltlichen Aspekten zusammengefaßt.

Die vergleichsweise schlechteste Modellanpassung ergab sich für Modell 2 (M2). Ihm liegt die Hypothese zugrunde, daß sich in den Lokalisations- und Ladungsparametern der Schulleistungskonstrukte nicht nur gruppen-, sondern auch zeitinvariante Eigenschaften der verwendeten Meßinstrumente widerspiegeln. Diese Hypothese konnte nicht beibehalten werden, da der resultierende chi^2-Wert bei nur 4 zusätzlichen Freiheitsgraden um 57.7 über dem Wert des Ausgangsmodells (M1) lag. Eine Inspektion der im Anhang aufgeführten Parameterschätzungen zeigt, daß die Mittelwerte der Schulnoten zum zweiten Meßzeitpunkt deutlich abnehmen.

Die in den Modellvarianten M3, M4 und M5 eingeführten Restriktionen führten demgegenüber im Vergleich zum Ausgangsmodell zu kompatiblen Anpassungswerten. Die formulierten Hypothesen bezogen sich auf Erwartungswerte und Effektparameter. In M3 wurde geprüft, ob die Erwartungswerte verbaler und nichtverbaler Fähigkeiten gruppenspezifisch jeweils gleich sind und damit auch die Unterschiede zwischen den Gruppen für diese Konstrukte. Modellvariante M4 lag die Hypothese zugrunde, daß die direkten Effekte der latenten exogenen Variablen, der Fähigkeitskonstrukte, auf die latenten endogenen Variablen, den Schulleistungen, invariant für die drei Gruppen sind. Zusätzlich ging in Modell M5 die Annahme der Gruppeninvarianz bezüglich der direkten Effekte der latenten endogenen Variablen untereinander ein. Die Hypothesensequenz ist im Sinne einer zunehmend sparsameren Modellformulierung aufgebaut. Als vergleichsweise bestes Modell wird mit M5 die Variante mit den meisten Freiheitsgraden, also der restriktivsten Spezifikation ausgewählt, die empirisch am ehesten falsifizierbar ist. Es postuliert strukturell wie numerisch invariante Beziehungen zwischen den latenten Variablen für die drei Schülergruppen unterschiedlichen Leistungsniveaus. Die Parameterschätzungen für Modell M5 finden sich im Anhang.

Tabelle 2 enthält eine Auswahl der inhaltlich interpretierbaren Befunde. Aufgeführt sind die Regressionen im Raum der Konstrukte, ihre Erwartungswerte und die aufgeklärten Varianzanteile der latenten endogenen Variablen.

Tabelle 2:

Regressionen im Raum der Konstrukte η_j (j = 1, 2, 3, 4), ihre Erwartungswerte und ihre aufgeklärten Varianzanteile[1]

A-Kurs	B-Kurs	C-Kurs
$\eta_1 = 0^* + 0.10\,\xi_1 + 0.51$	$\eta_1 = -0.75 + 0.10\,\xi_1 + 0.76$	$\eta_1 = -1.70 + 0.10\,\xi_1 + 0.68$
$\eta_2 = 0^* + 0.97\,\eta_1 + 0.24$	$\eta_2 = -0.44 + 0.97\,\eta_1 + 0.35$	$\eta_2 = -0.19 + 0.97\,\eta_1 + 0.32$
$\eta_3 = 0^* + 0.15\,\xi_2 + 0.89$	$\eta_3 = -0.24 + 0.15\,\xi_2 + 0.96$	$\eta_3 = -1.07 + 0.15\,\xi_2 + 1.13$
$\eta_4 = 0^* + 0.89\,\eta_3 + 0.10$	$\eta_4 = -0.71 + 0.89\,\eta_3 + 0.33$	$\eta_4 = -0.83 + 0.89\,\eta_3 + 0.41$
$E(\eta_1) = 0^*$	$E(\eta_1) = -1.30$	$E(\eta_1) = -2.73$
$E(\eta_2) = 0^*$	$E(\eta_2) = -1.69$	$E(\eta_2) = -2.83$
$E(\eta_3) = 0^*$	$E(\eta_3) = -1.10$	$E(\eta_3) = -2.67$
$E(\eta_4) = 0^*$	$E(\eta_4) = -1.69$	$E(\eta_4) = -3.21$
$R^2_{(\eta_1)} = 0.66$	$R^2_{(\eta_1)} = 0.35$	$R^2_{(\eta_1)} = 0.51$
$R^2_{(\eta_2)} = 0.91$	$R^2_{(\eta_2)} = 0.87$	$R^2_{(\eta_2)} = 0.90$
$R^2_{(\eta_3)} = 0.41$	$R^2_{(\eta_3)} = 0.42$	$R^2_{(\eta_3)} = 0.36$
$R^2_{(\eta_4)} = 0.99$	$R^2_{(\eta_4)} = 0.93$	$R^2_{(\eta_4)} = 0.91$

1 die mitgeteilten Ergebnisse beruhen auf der Grundlage von M5 (vgl. Tabelle 1)

Erwartungsgemäß spiegeln sich in den unstandardisierten Regressionskoeffizienten substantielle positive Effekte der latenten Variablen untereinander wider. Die Varianzaufklärung beider Schulleistungskonstrukte liegt zum zweiten Meßzeitpunkt zwischen 87% und 99%, so daß eine nahezu vollständige Vorhersagbarkeit durch die entsprechenden Konstrukte zum ersten Meßzeitpunkt gegeben ist. Generell liegen dabei die Werte für die nichtsprachlichen Leistungen etwas höher. Demgegenüber ist der mit Hilfe der verbalen und nichtverbalen Fähigkeiten erklärbare Anteil der Varianz der sprachlichen und nichtsprachlichen Schulleistungen zum ersten Meßzeitpunkt deutlich geringer (35%–66%). Angesichts von jeweils nur einer in die Regressionsgleichung eingehenden Prädiktorvariablen können diese Werte jedoch immer noch als beachtlich angesehen werden.

Die Mittelwerte der latenten Variablen zeigen die erwarteten deutlichen Niveauunterschiede für beide Schulleistungskonstrukte und Meßzeitpunkte. Dabei nehmen die relativen Unterschiede zwischen den Gruppen bezüglich der nichtsprachlichen Schulleistungen eher noch zu. Es bleibt offen, ob diese auf den Schulnoten basierenden Veränderungen auf schülerspezifischen Entwicklungsprozessen und/oder lehrspezifisch geänderten Bewertungspraktiken beruhen.

5. Die Kreuzvalidierung von Strukturgleichungsmodellen

Nach BENTLER (1980) können im Rahmen des LISREL-Ansatzes drei grundlegende Formen der Kreuzvalidierung unterschieden werden, die zunehmend strengere Anforderungen implizieren:

a) Validierung der Modellstruktur („loose replication"): Die Daten einer zweiten Stichprobe werden derselben Modellstruktur wie diejenigen der Ausgangsstichprobe unterworfen, wobei die zu schätzenden Parameter keinen Restriktionen unterliegen.
b) Validierung der Modellstruktur und einer Teilmenge der zu schätzenden Parameter („moderate replication"): Vorgegeben wird die Modellstruktur der Ausgangsstichprobe sowie die unverändert übernommenen Schätzungen für einzelne, theoretisch bedeutsame Parameter.
c) Validierung aller Parameterschätzungen („tight replication"): Modellstruktur und alle Parameter aus der Ausgangsstichprobe werden unverändert übernommen.

Grundlegend für unsere Vorgehensweise ist, daß die unterschiedlichen Varianten eine Voraussetzungsrelation definieren. Die erfolgreiche Durchführung einer „loose replication" ist eine notwendige Voraussetzung für eine „moderate replication" usw. Für die konkrete Anwendung ist zunächst die Formulierung „Übernahme theoretisch bedeutsamer Parameter" zu präzisieren. Die Vergleichbarkeit zweier strukturgleicher LISREL-Modelle erfordert in einem ersten Schritt die Sicherstellung der Gleichheit ihrer Meßmodelle in dem Sinne, daß die Indikatorvariablen in beiden Stichproben parallele Messungen darstellen. Erst in weiteren Analysen kann dann versucht werden, zusätzliche Parameter unverändert zu übernehmen.

Eine andere Vorgehensweise zur Kreuzvalidierung von Strukturgleichungsmodellen schlagen CUDECK & BROWNE (1983) vor. Ausgehend von der Formulierung und Schätzung von k unterschiedlichen Modellen für je einen Datensatz zweier unabhängiger Stichproben A und B resultieren als Werte der zu minimierenden ML-Schätzfunktion (5):

$$\{F_k(S_A ; \Sigma_k|S_A)\} \text{ und } \{F_k(S_B ; \Sigma_k|S_B)\}$$

mit: k = Modellindex ($k = 1, 2, \ldots$)
S_A, S_B = empirische (Kovarianz-)Matrix für die Stichproben A und B
$\Sigma_k|S_A$ bzw. $\Sigma_k|S_B$ = theoretische (Kovarianz-)Matrix bei S_A bzw. S_B als empirischer Ausgangsmatrix

Tabelle 3:

Zusammenstellung der durchgeführten Kreuzvalidierungsvarianten und ihrer Ergebnisse

Kreuzvalidierungsvariante	Kohorte 1			Kohorte 2			Modellanpassung bei simultaner Betrachtung beider Kohorten
	χ^2-Prüfgröße	GFI	RMR	χ^2-Prüfgröße	GFI	RMR	
Validierung der Modellstruktur	$\chi^2_{243} = 389.3$	0.81 0.86 0.77	9.4 19.6 30.7	$\chi^2_{243} = 321.5$	0.85 0.85 0.82	7.0 9.0 17.2	$\chi^2_{486} = 710.8$
Validierung der Modellstruktur bei unverändert übernommenen Ladungsparametern	$\chi^2_{251} = 409.1$	0.80 0.85 0.77	13.4 18.8 36.8	$\chi^2_{251} = 343.1$	0.84 0.84 0.80	15.9 8.6 18.6	$\chi^2_{502} = 752.2$
Validierung der Modellstruktur bei unverändert übernommenen Ladungs- und Effektparametern	$\chi^2_{255} = 431.2$	0.77 0.85 0.77	13.4 18.2 36.7	$\chi^2_{255} = 358.8$	0.84 0.84 0.79	15.9 8.9 18.7	$\chi^2_{510} = 790.0$
Validierung der Modellstruktur bei unverändert übernommenen Ladungs-, Effekt- und Lokalisationsparametern	$\chi^2_{269} = 539.8$	0.71 0.82 0.73	92.1 41.3 64.0	$\chi^2_{269} = 471.9$	0.76 0.82 0.74	93.6 39.1 50.8	$\chi^2_{538} = 1011.7$
Simultane Analyse beider Kohorten bei Annahme kohorteninvarianter Ladungs-, Effekt- und Lokationsparameter	–	0.80 0.85 0.76	20.5 24.7 46.6	–	0.85 0.85 0.80	26.8 18.0 29.5	$\chi^2_{507} = 770.0$

Diese Werte werden in Relation zu den Werten gesetzt, die man bei Übernahme aller Parameterschätzungen („tight replication") aus der jeweils anderen Stichprobe erhält („double cross validation"):

$\{F_k(S_A; \Sigma_k|S_B)\}$ und $\{F_k(S_B; \Sigma_k|S_A)\}$

CUDECK & BROWN bezeichnen diese Werte als Kreuzvalidierungsindizes („cross-validation index") und als optimal valides Modell dasjenige mit dem geringsten Kreuzvalidierungsindex.

Die von CUDECK & BROWN vorgeschlagene Prozedur erscheint aus mehreren Gründen nicht empfehlenswert: (1) Die sich für die Stichproben A und B ergebenden optimal validen Modelle stimmen in der Regel nicht überein; (2) die Auswahl der k verschiedenen Modellvarianten kann nicht begründet werden und (3) besitzt das Verfahren keinerlei formalstatistische Rechtfertigung.

Die Ergebnisse der im Sinne BENTLERs formulierten Sequenz zunehmend strengerer Kreuzvalidierungsvarianten sind in Tabelle 3 zusammengefaßt.

Die Modellanpassung für die Daten einer zweiten unabhängigen Schülerkohorte ergab bei Übernahme der Modellstruktur von Modellvariante 5 (M5) eine chi^2-Prüfgröße von 321.5. Dieser Wert liegt bei der gleichen Anzahl von Freiheitsgraden ($df = 243$) deutlich unter dem für die erste Kohorte resultierenden Anpassungswert von 389.3. Damit ist die Angemessenheit der angenommenen Modellstruktur als replizierbar nachgewiesen. In den weiteren Analysen wurde versucht, Teilmengen der zu schätzenden Parameter wechselseitig zu validieren. Zunächst wurde geprüft, ob die Indikatorvariablen in beiden Kohorten parallele Messungen darstellen, indem die Ladungsparameter unverändert eingesetzt werden. Die sich ergebenden chi^2-Differenzen lagen mit 19.8 für Kohorte 1 bzw. 21.6 für Kohorte 2 in der Größenordnung des kritischen Wertes von 20.1 (bei $df = 8$ und $\alpha = 0.01$), so daß die Ladungsparameter – mit Einschränkungen – als replizierbar angesehen werden können. Als nächste Teilmenge wurden zusätzlich die Parameter der direkten Effekte der latenten Variablen untereinander wechselseitig eingesetzt. Die chi^2-Differenzen von 22.1 für Kohorte 1 und 15.7 für Kohorte 2 lagen nunmehr beide über dem kritischen Wert von 13.3. Demgegenüber zeigten die RMR-Indizes praktisch keine Verschlechterung der Modellanpassung an. Dies änderte sich, wenn als weitere Parameter die Lokalisationsparameter hinzugenommen werden. Alle Kriterien zeigten die Unangemessenheit der Einführung dieser Restriktion an. Dies deutet darauf hin, daß sich die beiden untersuchten Kohorten insbesondere in den Erwartungswerten der Indikatorvariablen voneinander unterscheiden, während ihre Ladungs- und Effektparameter weitgehend übereinstimmen.

Unbefriedigend an der gewählten Vorgehensweise erscheint, daß es sowohl aufgrund theoretisch-inhaltlicher wie auch formal-statistischer Argumente grundsätzlich unangemessen ist, eine Übernahme exakter Parameterwerte zu fordern. Sozialwissenschaftliche Theorien beinhalten durchweg unpräzise Angaben über die numerischen Werte interessieren-

Anhang A1: Momentenmatrizen[1] für die 3 Gruppen von Kohorte 1

	PSB:1+2	PSB:6	AST4:RS	PSB:3	PSB:4	CFT2	Deutsch 1	Engl. 1	Deutsch 2	Engl. 2	Math. 1	Nat. 1	Math. 2	Nat. 2	Konst.
	3285.56														
	2704.55														
PSB: 1+2	2446.88														
	3177.11	3133.10													
	2591.41	2430.30													
PSB: 6	2333.33	2297.40													
	2989.30	2901.48	2766.80												
	2330.30	2245.35	2099.88												
AST4: RS	2044.17	1992.50	1797.40												
	3551.06	3449.30	3249.72	3934.86											
	3054.80	2938.89	2648.99	3547.98											
PSB: 3	2656.77	2571.88	2254.69	2994.27											
	3433.80	3340.49	3133.87	3781.69	3681.69										
	2832.83	2733.33	2458,59	3259.85	3071.97										
PSB: 4	2361.98	2292.19	2010.00	2620.31	2371.88										
	3349.79	3261.13	3070.18	3687.47	3565.35	3515.62									
	2802.63	2702.88	2440.85	3228.79	3001.11	3025.11									
CFT 2	2468.13	2372.92	2063.10	2741.35	2423.96	2594.25									
	469.79	456.69	430.54	512.82	495.49	485.21	68.87								
	355.40	342.48	314.18	406.06	376.06	373.62	48.93								
Deutsch 1	284.69	277.60	244.85	313.96	278.75	290.50	35.17								
	500.63	488.45	458.39	547.39	528.24	517.61	72.52	77.86							
	388.23	374.44	343.01	442.27	412.17	407.62	52.51	58.03							
Englisch 1	294.06	285.83	253.08	323.75	291.25	300.00	36.06	38.60							
	445.85	432.61	408.31	486.20	470.21	460.73	64.97	68.69	61.85						
	331.21	318.74	290.83	379.44	351.52	347.71	44.88	48.68	42.18						
Deutsch 2	274.17	265.31	233.42	301.88	270.00	281.69	33.08	34.27	32.15						
	438.52	427.54	402.03	479.44	462.18	452.20	63.63	67.80	60.25	59.92					
	303.18	292.78	268.54	344.34	321.67	317.21	40.92	44.87	38.10	35.85					
Englisch 2	242.29	232.08	203.75	263.02	235.63	246.25	28.44	29.69	27.54	25.00					
	449.65	437.54	413.30	492.96	476.50	467.35	65.48	69.62	62.21	60.96	64.10				
	352.17	340.71	309.01	408.03	381.41	376.33	47.90	52.22	44.46	40.44	49.82				
Mathem. 1	257.81	246.98	217.85	284.48	254.48	267.00	30.73	31.85	29.67	25.75	29.71				
	470.56	457.25	431.66	514.09	496.55	485.09	68.18	72.62	64.56	63.70	65.76	68.99			
	385.61	371.82	339.52	443.74	411.97	408.82	51.90	56.65	48.38	44.25	52.37	58.11			
Naturl. 1	304.58	293.23	261.77	337.50	298.96	311.71	36.15	37.19	34.90	30.21	33.42	40.35			
	419.44	408.59	385.00	461.69	445.78	437.92	61.07	65.01	58.07	56.92	59.10	61.18	55.58		
	290.25	280.05	254.77	336.92	314.29	310.80	39.24	42.64	36.54	33.15	40.39	43.06	33.97		
Mathem. 2	201.88	195.63	172.06	225.63	200.83	210.56	24.15	24.94	23.27	20.38	22.60	26.06	18.81		
	462.61	449.23	423.06	506.41	490.07	477.75	67.13	71.34	63.66	62.32	64.47	67.41	50.18	67.44	
	360.61	348.84	317.15	415.56	385.96	382.50	48.68	52.86	45.40	41.54	49.20	53.47	40.55	50.79	
Naturl. 2	272.71	262.60	230.85	301.56	266.35	281.85	30.44	33.13	31.25	27.17	30.42	35.29	23.92	33.06	
	56.69	55.21	51.76	62.47	60.28	58.75	8.20	8.76	7.76	7.66	7.87	8.23	7.35	8.11	1.00
	51.52	49.70	44.81	59.09	54.80	54.28	6.87	7.51	6.40	5.85	6.87	7.51	5.65	7.01	1.00
Konst.	48.54	47.19	41.44	54.06	47.92	50.21	5.77	5.94	5.56	4.83	5.21	6.19	4.10	5.52	1.00

1 Die Matrizen sind entsprechend Formel (1) aufgebaut; in der oberen Zeile steht jeweils der Wert für den A-Kurs, in der mittleren Zeile derjenige des B-Kurses und in der unteren Zeile die Werte des C-Kurses.

Anhang A2: Momentenmatrizen[1] für die 3 Gruppen von Kohorte 2

	PSB:1+2	PSB:6	AST4:RS	PSB:3	PSB:4	CFT2	Deutsch 1	Engl. 1	Deutsch 2	Engl. 2	Math. 1	Nat. 1	Math. 2	Nat. 2	Konst.
	3405.16														
	2675.50														
PSB: 1+2	2027.40														
	3406.35	3459.13													
	2623.50	2637.50													
PSB: 6	2064.90	2200.00													
	3030.48	3034.13	2750.46												
	2277.60	2258.70	2022.15												
AST4: RS	1721.35	1813.17	1576.37												
	3728.57	3744.84	3337.80	4145.64											
	3020.25	3011.00	2609.25	3546.75											
PSB: 3	2412.50	2506.73	2077.40	3032.21											
	3689.68	3696.83	3308.18	4076.19	4070.24										
	2809.00	2796.50	2429.00	3273.50	3090.50										
PSB: 4	2170.19	2253.37	1862.12	2702.89	2480.29										
	3444.44	3455.00	3087.60	3808.10	3782.54	3577.06									
	2641.75	2630.30	2281.09	3079.60	2869.35	2758.68									
CFT 2	1987.40	2074.90	1715.48	2481.06	2234.23	2092.65									
	469.05	469.84	420.35	515.32	511.03	478.54	66.02								
	353.45	351.10	307.66	407.55	379.60	358.50	48.64								
Deutsch 1	253.65	264.71	222.52	308.75	278.17	257.23	34.15								
	492.46	493.49	441.75	540.56	537.62	501.87	68.70	72.60							
	357.60	353.95	311.36	408.70	381.60	358.19	48.20	49.98							
Englisch 1	242.79	255.19	214.00	295.10	261.73	242.94	31.98	31.96							
	474.29	476.83	426.57	522.78	518.10	484.24	66.24	69.40	67.78						
	334.85	333.25	230.50	386.35	361.65	339,26	45.65	45.75	43.80						
Deutsch 2	238.27	249.71	210.37	292.98	263.46	241.83	31.29	29.71	29.92						
	442.06	443.25	395.56	486.51	482.54	450.11	61.50	64.68	62.40	58.51					
	296.95	294.45	256.87	340.15	319.00	297.99	40.33	40.75	38.38	34.49					
Englisch 2	179.04	188.75	157.54	218.46	196.06	180.42	23.44	22.73	22.04	17.10					
	448.89	450.00	402.37	494.76	493.40	460.76	62.33	65.52	63.18	58.65	60.60				
	305.00	303.20	264.83	352.95	330.25	311.86	41.44	41.98	39.45	34.81	37.79				
Mathem. 1	198.65	207.60	171.02	246.73	224.23	207.04	25.87	24.44	24.40	18.21	22.10				
	478.02	479.13	429.00	527.46	522.46	488.98	66.24	69.51	67.30	62.60	63.67	69.25			
	354.55	353.15	304.92	410.35	379.20	361.16	47.77	47.43	44.95	39.72	41.14	60.18			
Naturl. 1	242.89	252.60	209.27	297.98	268.75	249.64	31.69	29.81	29.67	22.10	24.89	32.15			
	419.44	419.68	377.19	463.89	461.59	432.92	58.40	61.14	59.22	54.98	56.40	59.83	53.51		
	278.95	277.55	241.89	325.55	302.90	287.56	37.99	38.04	36.18	31.86	33.44	38.41	31.80		
Mathem. 2	166.25	171.35	141.62	208.08	188.75	175.39	21.60	20.50	20.21	15.27	18.25	20.67	16.04		
	471.43	473.25	423.11	520.64	516.43	482.40	65.46	68.71	66.54	61.79	62.87	67.22	58.91	66.76	
	350.35	347.55	302.04	406.75	377.95	357.55	47.24	47.14	44.87	39.52	40.96	48.32	38.16	48.34	
Naturl. 2	229.71	237.69	197.67	283.08	255.77	234.79	29.56	27.92	28.25	21.02	23.48	29.39	19.56	28.75	
	58.02	58.18	51.85	64.05	63.41	59.25	8.05	8.44	8.16	7.59	7.71	8.24	7.22	8.13	1.00
	51.00	50.70	43.99	59.15	54.90	51.82	6.90	6.90	6.54	5.77	5.97	6.92	5.48	6.86	1.00
Konst.	43.94	45.96	38.33	54.33	48.75	45.08	5.69	5.39	5.39	3.98	4.48	5.50	3.73	5.21	1.00

1 Die Matrizen sind entsprechend Formel (1) aufgebaut; in der oberen Zeile steht jeweils der Wert für den A-Kurs, in der mittleren Zeile derjenige des B-Kurses und in der unteren Zeile die Werte des C-Kurses.

Anhang A3:

Parameterschätzungen für Modellvariante 5 (M5) in beiden Kohorten auf der Grundlage gleicher Modellstrukturen

	Kohorte 1			Kohorte 2		
Parameter	A-Kurs	B-Kurs	C-Kurs	A-Kurs	B-Kurs	C-Kurs
ν_1		56.95 (.84)			58.18 (.73)	
ν_2		55.06 (.87)			58.27 (.88)	
ν_3		51.39 (.91)			52.02 (.80)	
ν_4		62.59 (.65)			64.21 (.74)	
ν_5		59.98 (.79)			63.15 (.77)	
ν_6		59.07 (.81)			59.03 (.90)	
μ_1		8.21 (.13)			8.00 (.13)	
μ_2		8.78 (.13)			8.46 (.14)	
μ_3		7.83 (.12)			8.04 (.13)	
μ_4		7.60 (.13)			7.62 (.12)	
μ_5		7.92 (.17)			7.73 (.12)	
μ_6		8.16 (.13)			8.19 (.15)	
μ_7		7.34 (.14)			7.26 (.13)	
μ_8		8.15 (.15)			8.07 (.11)	
λ_1		0.91 (.08)			0.96 (.08)	
λ_2		0.83 (.08)			0.92 (.08)	
λ_3		0.72 (.08)			0.66 (.07)	
λ_4		0.88 (.10)			0.93 (.08)	
λ_5		0.96 (.07)			0.71 (.05)	
λ_6		0.84 (.06)			0.75 (.04)	
λ_7		0.63 (.06)			0.74 (.08)	
λ_8		0.74 (.06)			0.70 (.06)	
γ_{11}		0.10 (.01)			0.13 (.02)	
γ_{32}		0.15 (.02)			0.09 (.02)	
β_{21}		0.97 (.08)			0.71 (.07)	
β_{43}		0.89 (.07)			1.25 (.18)	
$\kappa_1 = \kappa_2$	0*	−5.72 (.90)	−10.68 (1.24)	0*	−7.91 (.86)	−14.34 (1.16)
α_1	0*	−0.75 (.16)	− 1.70 (.23)	0*	−0.53 (.22)	− 1.34 (.32)
α_2	0*	−0.44 (.14)	− 0.19 (.24)	0*	−0.80 (.16)	− 1.39 (.25)
α_3	0*	−0.24 (.24)	− 1.07 (.34)	0*	−1.07 (.20)	− 2.12 (.28)
α_4	0*	−0.71 (.16)	− 0.83 (.27)	0*	0.42 (.36)	0.53 (.61)
$\phi_{11} - \kappa_1^2$	55.73 (12.36)	32.80 (7.52)	52.79 (14.84)	22.14 (6.07)	35.45 (7.85)	53.70 (14.70)
$\phi_{12} - \kappa_2^2$	24.60 (6.58)	30.27 (7.42)	32.09 (10.65)	27.18 (7.23)	27.07 (6.89)	48.17 (13.41)
$\phi_{21} - \kappa_1 \kappa_2$	21.78 (6.43)	5.39 (4.56)	24.22 (9.12)	11.18 (4.55)	7.72 (4.63)	3.81 (8.94)
ψ_{11}	0.26 (.11)	0.57 (.16)	0.46 (.20)	0.63 (.20)	0.45 (.18)	0.73 (.31)
ψ_{22}	0.06 (.08)	0.12 (.11)	0.10 (.18)	0.15 (.14)	0.29 (.11)	0.09 (.13)
ψ_{33}	0.79 (.25)	0.93 (.29)	1.27 (.44)	0.06 (.10)	0.12 (.15)	0.16 (.21)
ψ_{44}	0.01 (.15)	0.11 (.16)	0.17 (.33)	0.07 (.16)	0.32 (.21)	0.00 (.23)
ψ_{31}	0.44 (.12)	0.58 (.15)	0.28 (.17)	0.09 (.08)	0.12 (.09)	0.32 (.14)
δ_1^2	22.55 (5.46)	24.23 (5.23)	43.13 (11.21)	17.05 (4.68)	40.44 (7.37)	62.63 (14.83)
δ_2^2	46.04 (8.81)	37.63 (6.56)	33.38 (8.94)	50.86 (10.11)	39.65 (7.10)	35.72 (9.76)
δ_3^2	29.74 (6.96)	62.66 (10.52)	49.96 (13.17)	36.27 (7.82)	48.21 (8.53)	57.03 (14.07)
δ_4^2	22.52 (4.50)	40.63 (6.68)	49.31 (11.28)	35.94 (7.03)	35.03 (5.75)	52.09 (11.24)
δ_5^2	29.14 (6.41)	38.43 (7.62)	44.09 (11.76)	21.36 (5.92)	47.40 (8.75)	55.07 (13.45)
δ_6^2	37.07 (7.26)	54.21 (9.08)	54.50 (13.04)	40.88 (8.63)	50.44 (8.80)	24.11 (7.87)
ϵ_1^2	0.80 (.15)	0.86 (.16)	0.85 (.23)	0.69 (.15)	0.64 (.12)	1.01 (.23)
ϵ_2^2	0.54 (.13)	0.85 (.17)	1.62 (.39)	0.34 (.17)	1.19 (.24)	1.19 (.35)
ϵ_3^2	0.77 (.15)	0.61 (.13)	0.53 (.16)	0.85 (.17)	0.55 (.10)	0.57 (.13)
ϵ_4^2	0.63 (.15)	0.85 (.18)	0.88 (.26)	0.31 (.15)	0.38 (.12)	0.28 (.14)
ϵ_5^2	0.78 (.23)	0.99 (.26)	0.36 (.35)	0.67 (.16)	1.86 (.31)	1.68 (.38)
ϵ_6^2	0.82 (.15)	1.39 (.22)	1.68 (.36)	1.37 (.26)	2.22 (.33)	2.00 (.40)
ϵ_7^2	0.37 (.18)	0.78 (.22)	0.44 (.34)	0.88 (.22)	0.96 (.24)	1.45 (.40)
ϵ_8^2	1.42 (.26)	0.95 (.17)	1.44 (.33)	0.68 (.14)	0.94 (.16)	1.90 (.40)
$\epsilon_{1,3}$	0.46 (.12)	0.24 (.11)	0.22 (.15)	0.18 (.12)	0.15 (.08)	0.15 (.13)
$\epsilon_{2,4}$	0.07 (.10)	0.22 (.13)	−0.21 (.23)	−0.06 (.12)	0.14 (.12)	0.05 (.17)
$\epsilon_{5,7}$	0.02 (.16)	0.20 (.20)	−0.54 (.28)	0.20 (.16)	0.37 (.22)	1.05 (.35)
$\epsilon_{6,8}$	0.31 (.15)	0.34 (.15)	0.44 (.26)	0.24 (.14)	0.68 (.18)	0.91 (.33)

der Parameter. Auch formal wäre es angemessener, bei der Kreuzvalidierung Intervallrestriktionen für die einzelnen Parameter zu berücksichtigen, beispielsweise der Art „exakte (Punkt-)Parameterschätzung ± Standardfehler". Diese Möglichkeit besteht jedoch in dieser Form im LISREL-Programm bislang nicht. Mit dem Ziel einer ersten Abschätzung der durch die Übernahme exakter Parameterschätzungen induzierten Anforderungen an die Daten wurde abschließend eine simultane Analyse beider Kohorten durchgeführt, deren Ergebnis in der untersten Zeile von Tabelle 3 zu finden ist. Hierbei wurden unter Ausnutzung der in LISREL vorhandenen Option sog. „constrained"-Parameter kohorteninvariante Ladungs-, Effekt- und Lokalisationsparameter spezifiziert. Dies entspricht einer Approximation von Intervallrestriktionen für diese Parameter. Die Kriterien für die Güte der Modellanpassung zeigen eine deutlich bessere Anpassung: Bei einem Vergleich entsprechender Modellvarianten anhand der beiden rechts unten zu findenden chi^2-Werte ergibt sich ein Differenzwert von 241.7, der bei 31 Freiheitsgraden hochsignifikant ausfällt.

6. Zusammenfassung

Anhand eines einfachen Illustrationsbeispiels aus der pädagogisch-psychologischen Forschung wurden die vielfältigen Möglichkeiten von LISREL-Modellen und ihrer Kreuzvalidierung bei der simultanen Analyse von Paneldaten mehrerer Gruppen aufgezeigt.

Demonstriert wurde, daß neben regressions- und faktorenanalytischen Verfahren auch varianzanalytische Modelle als spezielle Strukturgleichungsmodelle formulierbar sind. Die Vorteile der integrativen Behandlung unterschiedlicher statistischer Modelle liegen insbesondere in einer größeren Flexibilität und Anwendungsbreite. In einem einheitlichen terminologischen und schätztheoretischen Rahmen können sowohl Hypothesen über komplexe Einflußstrukturen als auch mittelwertbezogene Hypothesen analysiert werden. Anhand des ausgewählten Schulleistungsmodells konnten dabei – trotz seiner offenkundigen Restriktionen – einige der in der Literatur berichteten Befunde repliziert werden.

Darüber hinaus wurde ausführlich auf die Probleme bei der Testung von Strukturgleichungsmodellen eingegangen. Im Hinblick auf die Frage, welches Kriterium für die statistische Modellbeurteilung herangezogen werden sollte, wird häufig vorgeschlagen, daß eine simultane Betrachtung aller verfügbaren Informationen die Methode der Wahl sein sollte. Wie sich jedoch aus den vorgelegten Analysen ablesen läßt, verändern sich die verschiedenen Kriterien (chi^2-Prüfgröße, GFI, RMR) nicht unbedingt gleichsinnig, wenn unterschiedliche Modellvarianten vergleichend gegenüber-

gestellt werden. Dies bestätigt den Eindruck, daß bei der Modellbeurteilung allzuviele Freiheitsgrade vorhanden sind, was die Gefahr einer zu liberalen Handhabung dieser zentralen Frage birgt. Entscheidet man sich etwa dafür, den globalen chi^2-Wert als Richtschnur der Modellbeurteilung zu nehmen, bieten sich gezielte Modellmodifikationen an, um die Anpassungsgüte an die empirischen Daten zu verbessern. Ohne anschließende Kreuzvalidierung des dann in explorativer Weise gewonnenen Modells läßt sich aber nicht entscheiden, ob dieses Vorgehen lediglich zu einer Überanpassung an die zugrundeliegenden Stichprobendaten geführt hat („data overfit") oder nicht. In diesem Zusammenhang wurden verschiedene im LISREL-Ansatz realisierbare Kreuzvalidierungsvarianten demonstriert. Dabei zeigte sich, daß das postulierte Schulleistungsmodell in wesentlichen Aspekten auf eine zweite unabhängige Schülerkohorte derselben Grundgesamtheit übertragen werden kann. Während die Ladungs- und Effektparameter strukturell wie numerisch weitgehend kompatibel waren, kamen die Vorteile der zusätzlichen Berücksichtigung von Mittelwertsinformationen darin zum Ausdruck, daß für beide Kohorten insbesondere unterschiedliche Lokalisationsparameter nachgewiesen werden konnten. Abschließend wurde die Angemessenheit der bei Kreuzvalidierungen üblichen Übernahme exakter Parameterwerte im Fall sozialwissenschaftlicher Modelle kritisch diskutiert. Weitaus angemessener erscheint die Möglichkeit, Intervallrestriktionen für die Parameter vorzugeben. Diese Vorgehensweise dürfte der mangelnden Präzision sozialwissenschaftlicher Theorien eher gerecht werden und sollte eine Option in künftigen LISREL-Versionen werden.

7. Literatur

Bentler, P. M.: Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 1980, 31, 419–456.

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

Boomsma, A.: The robustness of LISREL against small sample sample sizes in factor analysis models. In: K. G. Jöreskog & W. Wold (eds.), Systems under indirect observation (Part I). Amsterdam: North-Holland, 1982, 149–174.

Cattell, R. B. & Weiss, R. H.: Grundintelligenztest CFT2. Braunschweig: Westermann, 1972.

Cudeck, R. & Browne, M. W.: Cross-validation of covariance structures. Multivariate Behavioral Research, 1983, 18, 147–167.

Fippinger, F.: Allgemeiner Schulleistungstest für 4. Klassen (AST4). Weinheim: Beltz, 1967.

Green, B. F.: Parameter sensitivity in multivariate methods. Journal of Multivariate Behavioral Research, 1977, 12, 263–287.

Horn, W.: Prüfsystem für Schul- und Bildungsberatung (PSB). Göttingen: Hogrefe, 1969.

Jöreskog, K. G.: Estimation and testing of simplex models. Brit. J. of Math. and Stat. Psychology, 1970, 23, 121–145.

Jöreskog, K. G.: Statistical estimation of structural models in longitudinal-development investigations. In: J. R. Nesselroade & P. B. Baltes (eds.), Longitudinal research in the study of behavior and development. N.Y.: Academic Press, 1979, 303–351.

Jöreskog, K. G. & Sörbom, D.: Advances in factor analysis and structural equation models. Cambridge, Mass.: Ast Books, 1979.

Jöreskog, K. G. & Sörbom, D.: LISREL VI. Analysis of linear structural relationships by the method of maximum likelihood. User's Guide. Mooresville, Ind.: Scientific Software, 1984.

Kühn, R.: Bedingungen für Schulerfolg. Zusammenhang zwischen Schülermerkmalen, häuslicher Umwelt und Schulnoten. Göttingen: Hogrefe, 1983.

Long, J. S.: Covariance structure models: An introduction to LISREL. Beverly Hills, Cal.: Sage, 1983.

Möbus, C.: Analyse von Rohdaten mit LISREL: Allgemeines Lineares Modell, stochastische Differenzen- und Differentialgleichungssysteme. In: C. Möbus & W. Schneider (Hrsg.), Strukturmodelle für Längsschnittdaten und Zeitreihen. Bern: Huber, 1986, 57–126.

Möbus, C. & Schneider, W. (Hrsg.): Strukturmodelle für Längsschnittdaten und Zeitreihen. Bern: Huber, 1986.

Morrison, D. F.: Multivariate statistical methods. New York: McGraw-Hill, 1967.

Munck, I. M. A.: Model building in comparative education. Stockholm: Almquist & Wiksell, 1979.

Opwis, K.: LISREL zur Analyse multivariater Mittelwertstrukturen. In: C. Möbus & W. Schneider (Hrsg.), Strukturmodelle für Längsschnittdaten und Zeitreihen. Bern: Huber, 1986, 221–231.

Parkerson, J. A., Lomax, R. G., Schiller, D. P. & Walberg, H.: Exploring causal models of educational achievement. Journal of Educational Psychology, 1984, 76, 638–646.

Schneider, W. & Treiber, B.: Classroom differences in the determination of achievement changes. American Educational Research Journal, 1984, 21, 195–211.

Sörbom, D.: Structural equation models with structured means. In: K. G. Jöreskog & H. Wold (eds.), Systems under indirect observation. (Part I). Amsterdam: North-Holland, 1982, 183–196.

Timm, N. H.: Multivariate analysis with applications in education and psychology. Monterey, Cal.: Brooks/Cole, 1975.

Werts, C. E., Linn, R. L. & Jöreskog, K. G.: A simplex model for analyzing academic growth. Educational and Psychological Measurement, 1977, 37, 745–756.