Kapitel 12

Mehrebenenanalytische Ansätze zur Erklärung von Schulleistungen

Wolfgang Schneider, Andreas Helmke

12.1 Einleitung

In Modellen zur Beschreibung und Erklärung von Schulleistungen werden Verlauf und Ergebnis schulischen Lernens im wesentlichen auf Merkmale der Schülerpersönlichkeit, des Elternhauses, der Gleichaltrigengruppe und der Schule zurückgeführt (vgl. Weinert & Zielinski, 1977; und die Übersicht bei Haertel, Walberg & Weinstein, 1983). Es gibt unterschiedliche Gründe dafür, daß die zahlreichen empirischen Untersuchungen zur spezifischen Problematik eher zur Verwirrung als zur Klärung bzw. Differenzierung der Sachlage geführt haben: Die Wahl eines allzu einfachen statistischen Analysemodells sowie die Ausklammerung effektrelevanter Schüler- bzw. Unterrichtsmerkmale spielen dabei sicherlich eine wesentliche Rolle (Kühn, 1983; Schneider & Bös, im Druck).

Während Spezifikationsfehler dieser Art im Prinzip durchaus erkannt und gewürdigt werden, fällt auf, daß auf das Problem der hierarchischen Strukturiertheit von Bildungsprozessen in Lehr-Lern-Modellen bislang nur unzulänglich eingegangen worden ist. Dies erstaunt umso mehr, als in der Literatur inzwischen zahlreiche kritische Erörterungen bzw. Diskussionen der Mehrebenenproblematik in Hinblick auf die Bewertung von Bildungseffekten vorliegen (vgl. z.B. Achtenhagen, 1981; Burstein, 1980b; Fend, 1977; v. Saldern, 1982; Treiber, 1980a, b, c). In all diesen Arbeiten wird der Umstand hervorgehoben, daß Bildungsprozesse auf mehreren Ebenen verankert sind: Schüler gehören beispielsweise zu bestimmten Bildungseinheiten (Schulklassen), Schulklassen wiederum sind in Einheiten höherer Ordnung (Schulen) eingebettet usw. . Wie Rachman-Moore & Wolfe (1984) betonen, sollten realistische Modelle des Lehr-Lern-Prozesses diese hierarchische Struktur abbilden, da nicht davon ausgegangen werden kann, daß die hierarchische Verschachtelung der verschiedenen Analyseeinheiten nach Zufallsregeln erfolgt. Es ist eher zu unterstellen, daß spezifische Gruppierungsregeln existieren, denen zufolge Individuen mit bestimmten Charakteristiken in Gruppen mit spezifischen Eigenschaften zusammengefaßt sind. Von daher ist zu erwarten, daß in groß angelegten (large-scale survey) Studien zur Erfassung von Schulleistungen signifikante Schul-, Klassen- und Lehreffekte resultieren. Traditionelle Einebenenanalysen, wie sie auch in der neueren Literatur zur

Schulleistungsprognose (vgl. Parkerson, Lomax, Schiller & Walberg, 1984) noch immer vorgelegt werden, vernachlässigen erklärungsrelevante Analyseebenen, sind mit Spezifikationsfehlern behaftet und führen demzufolge zu verzerrten Effektschätzungen.
Die vorliegende Arbeit setzt sich mit unterschiedlichen Möglichkeiten auseinander, Modelle zur Beschreibung und Erklärung von Schulleistungen mehrebenenanalytisch anzulegen. Zunächst werden die verschiedenen traditionellen Verfahren der Mehrebenenanalyse knapp zusammengefaßt (zur detaillierten Beschreibung s. die oben erwähnte Überblicksliteratur). Auf diesen eher klassischen Ansätzen bauen einige neuere Prozeduren auf, die im folgenden Abschnitt genauer erläutert werden. Daran schließt sich die Beschreibung des eigenen Ansatzes an, der im abschließenden Abschnitt anhand eines Anwendungsbeispiels illustriert wird.

12.2 Grundlegende analytische Modelle

Bei der Durchsicht der Literatur fällt auf, daß unterschiedliche Versuche unternommen worden sind, den Mehrebenencharakter von Modellen zur Analyse von Bildungseffekten angemessen zu berücksichtigen. In Anlehnung an die bei Burstein (1980b) vorfindbare Einteilung gehen wir davon aus, daß diese Versuche im wesentlichen drei Problembereichen gewidmet sind:

(1) dem Problem des ökologischen Fehlschlusses bzw. der Auswahl der geeigneten Analyseebene;
(2) verschiedenen Möglichkeiten der Analyse von Gruppeneffekten;
(3) der Spezifikation von angemessenen analytischen Modellen für Mehrebenendaten.

12.2.1 Das Problem des ökologischen Fehlschlusses bzw. der Auswahl der geeigneten Analyseebene

Das Problem von ebenendifferenten Schlüssen (cross-level inferences) läßt sich allgemein so charakterisieren, daß Vorhersagen für Analyseebenen gemacht werden, die in der empirischen Analyse nicht direkt erfaßt werden. In der Praxis trifft dies besonders häufig auf Studien zu, die auf Schulklassenebene durchgeführt werden, im wesentlichen jedoch Aussagen über Individuen (Schüler) machen wollen. Das Problem besteht jedoch prinzipiell auch für den umgekehrten Fall, daß Analysen für Schüler vorliegen, Schlußfolgerungen jedoch für Gruppen erfolgen. In der Literatur ist der erstgenannte Spezifiktionsfehler als „ökologischer Fehlschluß" und der letztegenannte

(seltenere) als „individueller Fehlschluß" bekannt geworden (vgl. Kap 1 dieses Bandes). Burstein (1980b) diskutiert verschiedene regressionsanalytische Ansätze, in denen das Ausmaß des ökologischen Fehlschlusses dadurch einzugrenzen versucht wird, daß die auf Klassenebene durchgeführte Analyse zusätzlich (hoch-aggregierte) Schülerindikatoren enthält. Obwohl damit das Ausmaß des ökologischen Fehlschlusses in gewissem Umfang abgeschätzt werden kann, bleiben die Konsequenzen in der Praxis unklar. Von daher ist Bursteins (1980b) Schlußfolgerung zuzustimmen, daß es ausgesprochen unklug und riskant ist, Daten auf der Aggregat- (Klassen-) ebene zu analysieren, wenn das eigentliche Ziel der Untersuchung in der Erfassung von Faktoren liegt, die die individuellen Schulleistungen beeinflussen.
Dieses Fazit leitet zum verwandten Problem über, was denn nun eigentlich die angemessene Analyseebene sei. Rogosa's (1978) Beurteilung klärt den Leser schnell darüber auf, daß es sich hier im Grunde um eine verfehlte Fragestellung handelt: „...confining stubstantive questions to any one level of analysis is unlikely to be a productive research strategy" (S. 83).
Ein wesentlicher Grund dafür ist darin zu sehen, daß bei der schon erwähnten hierarchischen Verschachtelung von Schulleistungsdaten davon ausgegangen werden muß, daß korrelative Abhängigkeiten zwischen Beobachtungen innerhalb von Gruppen existieren. Diese Abhängigkeiten haben vor allem negative Konsequenzen für das zugrundegelegte statistische Modell, insbesondere dann, wenn das traditionelle regressionsstatistische Modell

$$Y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij} \qquad (1)$$

zugrundegelegt wird, in dem die Schulleistung des Individuums j in Klasse i (Y(ij)) z.B. über seine Fähigkeit (X(ij)) vorhergesagt wird. In diesem Fall läßt sich die Modellannahme, daß die Fehlerterme (ε(ij)) annähernd normal verteilt und unabhängig voneinander sind, nicht mehr halten. Es ist im Gegenteil davon auszugehen, daß die Fehlerterme (die im übrigen sowohl Spezifikations- wie auch Meßfehler erfassen) für den Fall hierarchisch verschachtelter Daten untereinander korreliert sind. Die Analyse dieser Daten mit den üblichen Regressionsverfahren wird folglich zu fehlerverzerrten Schätzungen führen. Auf der anderen Seite sollte nicht übersehen werden, daß korrelative Abhängigkeiten innerhalb von Gruppen durchaus Informationen über unterschiedliche Unterrichtsprozesse in unterschiedlichen Gruppen liefern können (s. Burstein, 1980b), ein Punkt, auf den noch ausführlicher eingegangen werden soll.
Der direkte Vergleich von Analysen, die unter Verwendung des gleichen Datensatzes sowohl auf der Individual- wie auf der Aggregatebene durchgeführt wurden, macht die Unzulänglichkeit von Einebenenanalysen offensichtlich. Wie Treiber (1980a, b) herausstellte, sind die auf unterschiedlichen Analyseebenen resultierenden Parameterschätzungen in der Regel inkonsi-

stent. Das Erklärungsgewicht individueller Einflußfaktoren wird in Aggregatdatenanalysen überschätzt, während kollektive Effekte in Individualdatenanalysen eher unterschätzt werden. Empirische Belege dafür finden sich in einer Studie von Treiber & Schneider (1978) zur Vorhersage der Mathematikleistungen durch Merkmale der Schülerpersönlichkeit und der Schulklasse. Wenn die Merkmale der Schülerperson und Schulklasse nur auf der Individual- bzw. auf Aggregatebene verknüpft und getrennte Regressionen (Einebenenanalysen) gerechnet wurden, ergaben sich die vorhergesagten Differenzen innerhalb wie zwischen beiden Analysen.
Dies alles unterstreicht nur Rogosa's (1978) Schluß, daß Diskussionen über die Wahl der angemessenen Analyseeinheit absolut unnütz sind, und daß stattdessen, wann immer möglich, mehrebenenanalytische Betrachtungsweisen präferiert werden sollten.

12.2.2 Möglichkeiten der Analyse von Gruppeneffekten

Verschiedene methodische Zugangswege scheinen geeignet, hierarchisch strukturierte (Schulleistungs-) Daten angemessen zu analysieren. Mehrebenenanalytische Effektzerlegungsmodelle werden üblicherweise im Rahmen des Allgemeinen Linearen Modells (wie Achtenhagen, 1981, zu Recht hervorhebt, mit allen Konsequenzen für die Beurteilung der Datenqualität) spezifiziert. Die Zerlegung der Gesamtvarianz interindividueller Schulleistungsunterschiede in mehrere ineinander hierarchisch verschachtelte additive Varianzkomponenten (z.B. von Schulklassen in Schulen in Schulformen) kann dabei über zwei- oder mehrfaktorielle Auswertungsverfahren kreuzweghierarchischer Varianzanalysen erfolgen (vgl. Treiber, 1980a, b; Treiber & Weinert, 1985). Eine andere Möglichkeit besteht darin, ANCOVA-Modelle zu spezifizieren, in denen Dummyvariablen die Gruppenzugehörigkeit angeben. Das in (1) angegebene einfache Regressionsmodell erweitert sich dann zu

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma_1 G_1 + \ldots + \gamma_{m-1} G_{m-1} + \varepsilon_{ij} \qquad (2)$$

wobei G(1), ... G(m-1) Dummyvariablen der Gruppenzugehörigkeit repräsentieren. Ein Problem von Varianzzerlegungsmodellen besteht darin, daß zwar Gruppeneffekte nachgewiesen werden können, es aber nicht möglich ist, Ursachen von Gruppeneffekten zu bestimmen, bzw. herauszufinden, warum sich einige Gruppen von anderen unterscheiden. Die Literatur zur Analyse von Kontexteffekten (z.B. Boyd & Iversen, 1979) weist demgegenüber auf Ansätze hin, über die sich der Einfluß der Gruppenzugehörigkeit auf individuelle (Schul-) Leistungen direkter erheben läßt.
Gruppeneffekte lassen sich dabei als Kontext- oder Bezugsgruppen-(„frog pond"-) Effekte spezifizieren. Nach Boyd & Iversen (1979) spricht man von

einem Kontexteffekt (z.B. der Fähigkeit), wenn die durchschnittliche Fähigkeit der Gruppe ($\bar{X}$(i.)) auch nach Auspartialisierung der individuellen Fähigkeit (X(ij)) einen unabhängigen Einfluß auf die individuelle Schulleistung (Y(ij)) hat. Im Unterschied dazu wird beim sogenannten Bezugsgruppen- bzw. Vergleichseffekt der relativen Position des Schülers in seiner Klasse besondere Beachtung geschenkt. Rein theoretisch ist es beispielsweise möglich, daß ein Lehrer bei der Beschäftigung mit einzelnen Schülern deren relative Fähigkeit systematisch berücksichtigt.
Kontext- oder Bezugsgruppeneffekte lassen sich auch formal in einem Regressionsmodell abbilden, das folgendermaßen spezifiziert werden könnte:

$$Y_{ij} = \alpha + \beta_1 X_{ij} + \beta_2 \bar{X}_{i.} + \beta_3 (X_{ij} - \bar{X}_{i.}) + \varepsilon_{ij} \qquad (3)$$

In diesem Modell gibt $\beta(1)$ den individuellen, $\beta(2)$ den Kontext- und $\beta(3)$ den Bezugsgruppeneffekt an.
Wie bei Burstein (1980b) und Treiber (1980a, b) herausgestellt wird, läßt sich dieses Modell so nicht schätzen, da die drei Variablen X(ij), $\bar{X}$(i.) und (X(ij) - $\bar{X}$(i.)) linear abhängig sind. Zur Lösung des Dilemmas schlug Firebaugh (1979, 1980) drei Möglichkeiten vor:

a) eine der drei Variablen aus theoretischen Gründen weglassen;
b) direkte Messungen für den Kontext- bzw. Bezugsgruppeneffekt vornehmen

oder

c) unterschiedliche Instrumente zur Erfassung von Kontext- und Bezugsgruppeneffekt verwenden.

Ohne Zweifel stellt die Repräsentation von Gruppeneffekten über $\bar{X}$(i.) bzw. (X(ij) - $\bar{X}$(i.)) eine zu mechanische bzw. distale Messung dar, so daß die unter b) genannte Möglichkeit, direktere Messungen für Gruppenprozesse vorzunehmen, in jedem Falle vorzuziehen ist. Es liegt auf der Hand, daß Kontext- und Bezugsgruppeneffekte dann auch kein wesentliches Problem mehr darstellen.

12.3 Zur Spezifikation angemessener analytischer Modelle für Mehrebenendaten

Bevor kurz auf die bekanntesten Strukturmodelle der Mehrebenenanalyse eingegangen wird, sei der Leser darauf hingewiesen, daß nach wie vor keine Standardprozeduren existieren. Die verfügbaren Methoden lassen sich nach unterschiedlichen Kriterien unterscheiden, etwa danach, ob sie ein eher

einfaches Strukturgleichungsmodell (traditionelles Regressionsverfahren) oder die aus der Ökonometrie stammenden elaborierteren Strukturgleichungsmodelle mit latenten Variablen (z.B. LISREL) zugrundelegen, ob sie sequentielle oder simultane Analysen beinhalten usw. Als wohl herausragendes gemeinsames Merkmal dieser Ansätze kann jedoch ihre Fokussierung auf die mehrebenenanalytische Zerlegung von Relationen (Kovarianzen, Korrelationen, Regressionen) anstatt der von Variationen angesehen werden. Wie schon erwähnt, lassen sich durch Zerlegung von Varianzen entdeckte Kontexteffekte inhaltlich nur schwer interpretieren. Demgegenüber ist die Annahme verbreitet (vgl. Burstein, 1980b; Burstein & Linn, 1982; Burstein, Linn & Capell, 1978; vgl. auch Kap. 3 dieses Bandes), daß über die mehrebenenanalytische Zerlegung von Relationen Schulklassenprozesse direkter erfaßt werden können. Die inzwischen wohl klassisch zu nennende Zerlegung der Relationen wurde von Cronbach (1976) vorgenommen (s.a.Kap 3):

$$
\begin{aligned}
Y_{ij} = {} & \bar{Y}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) && \text{Vorhergesagter Interklasseneffekt} \\
& + (\bar{Y}_{i.} - Y_{..}) - \beta_s(\bar{X}_{i.} - \bar{X}_{..}) && \text{Adjustierter Interklasseneffekt} \\
& + \beta_w(X_{ij} - \bar{X}_{i.}) && \text{Gepoolter Interklasseneffekt} \\
& + (\beta_i - \beta_w)\,(X_{ij} - \bar{X}_{i.}) && \text{Spezifischer Interklasseneffekt} \\
& + \varepsilon_{ij} && \text{Fehlerterm für den einzelnen Schüler}
\end{aligned}
\qquad (4)
$$

Anhand dieser Zerlegung lassen sich auf Schulklassen- und Schülerebene etwa die folgenden Fragen beantworten:

a) Wie sieht die Beziehung zwischen der Klassenzusammensetzung beim Prätest und der Schulleistung (auf Klassenebene) beim Posttest aus (vorhergesagter Interklasseneffekt)?
b) Warum sind für Schulklassen mit gleichen Eingangscharakteristika unterschiedliche Posttest-Schulleistungen zu verzeichnen (adjustierter Interklasseneffekt)?
c) Warum lernen einige Schüler mehr als andere in Schulklassen, wenn man ihre relativen Unterschiede in Hintergrundsvariablen berücksichtigt (gepoolter Innerklasseneffekt)?
d) Warum gibt es unterschiedliche Beziehungen zwischen Prätest (oder anderen Hintergrundsvariablen) und Posttest für unterschiedliche Klassen (spezifischer Innerklasseneffekt)?

Mit den Variationen in den genannten Effektgrößen sind auch ganz bestimmte inhaltliche Deutungsmuster verknüpft (vgl. Achtenhagen, 1981; Burstein, 1980b): Hoch positive Interklassenregressionskoeffizienten (β(b)) können demnach als Indikatoren für Unterrichtsformen gewertet werden, in denen fähigen Schülern großer Spielraum zur Entfaltung ihrer Talente gegeben wird. Umgekehrt deuten niedrige β(s)-Koeffizienten auf kompensatorische Unterrichtseffekte hin. Adjustierte Interklasseneffekte zeigen den spezifischen Lehrer- bzw. Klasseneinfluß auf den im Hinblick auf die Prätestleistung adjustierten mittleren Klassenerfolg. Die Interpretation des gepoolten Innerklassenregressionskoeffizienten (β(w)) entspricht der von β(b), bezieht sich nun aber auf Prozesse innerhalb von Gruppen und nicht auf Gruppenmittelwerte.

Besondere Beachtung hat in jüngerer Zeit der spezifische Innerklasseneffekt (β(i)) erfahren, was insbesondere auf die Arbeiten von Leigh Burstein und seinen Mitarbeitern (Burstein, 1980b; Burstein & Linn, 1982; Burstein, Linn & Capell, 1978) zurückzuführen ist. Obwohl heterogene β(i)-Werte über unterschiedliche Klassen hinweg prinzipiell auf Stichprobenverzerrungen aufgrund variierender Klassengrößen bzw. Unterschiede in der Klassenzusammensetzung rückführbar sind (vgl. Achtenhagen, 1981; Cronbach, 1976), lassen sie sich auch im Sinne systematischer Unterschiede in den Lehr-Lern-Prozessen interpretieren. Damit ist speziell der Umstand gemeint, daß einige Lehrer relativ größere kompensatorische Effekte als andere aufweisen, was sich in einer unterschiedlichen Verteilung der Schulleistungen beim Posttest (dokumentiert durch flache Regressionssteigungen) im Vergleich zum Prätest ausdrücken kann. Steile Regressionsgeraden (d.h. positive β-Koeffizienten) lassen demgegenüber eher auf traditionelle bzw. individualisierende Unterrichtspraktiken schließen, bei denen sich Schüler mit höheren Eingangsfähigkeiten in Relation zu schwächeren Schülern stärker entfalten können, was in den hohen Posttestwerten zum Ausdruck kommt (vgl. zur detaillierten Beschreibung Burstein, Linn & Capell, 1978).

Burstein und Mitarbeiter (z.B. Burstein, Linn & Capell, 1978) haben diesem Problem besondere Aufmerksamkeit gewidmet und einen mehrebenenanalytischen Auswertungsansatz empfohlen, der unter der Bezeichnung „slopes as outcomes" bekannt geworden ist. In diesem Ansatz wird unterstellt, daß aufgrund unterschiedlicher Kontextbedingungen in unterschiedlichen Klassen die Innerklassenregressionskoeffizienten (z.B. der Vortest-Nachtest-Beziehung) *systematisch* heterogen ausfallen. Burstein (1983) wählt deshalb die Abkürzung SVS (systematically varying slopes).

Das mehrebenenanalytische Verfahren erfolgt in zwei Schritten: Zunächst werden für die einzelnen Schulklassen (als separate „treatment"-Bedingungen) die spezifischen Innerklassenregressionskoeffizienten ermittelt. Im zweiten Schritt werden diese Regressionskoeffizienten als abhängige Variablen in einer Interklassenregressionsanalyse betrachtet, bei der hochaggregierte

Schülervariablen (z. B. durchschnittliche Prätestleistungen und Hintergrundsmerkmale) zusammen mit auf Klassenebene erfaßten Instruktionsmerkmalen als Prädiktoren fungieren.

$$\hat{\beta}_i = \gamma_0 + \gamma_z z_i + \lambda \bar{X}_{.j} + \delta \qquad (5)$$

Ein alternatives mehrebenenanalytisches Modell zur Identifikation von Schüler- und Klasseneffekten wurde von Keesling und Wiley (1974; Wiley, 1976) entwickelt (siehe die Kapitel 3 und 4 dieses Bandes). Die Eleganz dieses kovarianzanalytischen Ansatzes liegt darin, daß einmal der Einfluß der Klassenzugehörigkeit bei der Regression von Schulleistungen auf individuelle Hintergrundsvariablen berücksichtigt und gleichzeitig der Effekt von Gruppen- (Instruktions-) Merkmalen auf die auf Klassenebene erfaßten Schulleistungen um den Einfluß individueller Unterschiede in den Klassen korrigiert wird. Dies sieht konkret so aus, daß zunächst auf Schülerebene eine ANCOVA durchgeführt wird, wobei als abhängige Variable die individuelle Schulleistung Y(ij), als unabhängige Variable die individuelle Klassenzugehörigkeit Z(i), und als Kovariaten die Schülerhintergrundsmerkmale X(ij) fungieren. Über diese gepoolte Innerklassenregression erhält man für jede Schulklasse einen geschätzten Schulleistungswert ($\hat{Y}$), der zusätzlich in die Regression der beobachteten Klassenmittelwerte auf Erklärungsvariablen der Schulklassenebene (Lehrer, Unterricht) (Z(i)) aufgenommen wird:

$$Y_i = \gamma_0 + \gamma_z z_i + \lambda \bar{\hat{Y}}_i + \delta \qquad (6)$$

Es wird davon ausgegangen, daß durch die Berücksichtigung des geschätzten Klassenmittelwerts ($\hat{Y}$(i.)) der als summierter Schülerpersoneneffekt aufzufassen ist, eine verbesserte Schätzung der auf Schulklassenebene aufzuklärenden Klassenunterschiede in der abhängigen Variablen erfolgen kann. Obwohl Burstein (s. Kap. 3) den Umstand beklagt, daß bislang nur wenige empirische Anwendungen des Keesling & Wiley-Modells bekannt geworden sind, liegen im deutschsprachigen Raum inzwischen mehrere Modellerprobungen vor (Treiber, 1980c, 1981; Treiber & Schneider, 1978, 1980, 1981).

12.4 Methodische Probleme der traditionellen Mehrebenenanalysen

Wie von Burstein (1980b) und Treiber (1980b) herausgestellt wurde, gilt das Keesling & Wiley-Modell streng genommen nur unter der Voraussetzung, daß die schulklassenspezifischen Regressionssteigungen $\beta(i)$ parallel verlaufen, m.a.W. nicht signifikant voneinander abweichen. Die Annahme nur zufällig variierender $\beta(i)$-Koeffizienten kann im Modell überprüft und das Modell

beibehalten werden, wenn der in (6) aufgeführte λ-Koeffizient den Wert 1 annimmt. Für den Fall signifikanter Modellabweichungen muß das Modell jedoch neu spezifiziert werden, indem etwa solche Klassenmerkmale aufgenommen werden, die die Unterschiede in β(i) aufklären können. Es liegt auf der Hand, daß solche Neuspezifikationen in der Praxis erhebliche Probleme bereiten können, sei es, daß zusätzlich aufgenommene theoretisch sinnvolle Klassenmerkmale die Unterschiede zwischen Klassen in β(i) auch nicht merklich reduzieren können, oder sei es, daß möglicherweise in diesem Zusammenhang theoretisch sinnvolle Klassenmerkmale überhaupt nicht empirisch erhoben wurden.

Treiber (1980b, S.257) macht einen weiteren Vorschlag zur Problemüberwindung, der das Keesling & Wiley-Modell in die Nähe des „slopes as outcomes"-Ansatzes von Burstein und Mitarbeitern rückt. Der Vorschlag sieht vor, für jede Schulklasse die spezifischen β(i)-Koeffizienten zu schätzen und diese als abhängige Variablen in die sich anschließende Schulklassenanalyse aufzunehmen. Auf den ersten Blick scheint dies eine sinnvolle Maßnahme zu sein. Ihr Wert wird aber dann zweifelhaft, wenn man sich die Kritik am „slopes as outcomes"-Ansatz zu eigen macht, die von den Autoren selbst (vgl. Burstein, 1980b, 1983; Burstein & Linn, 1982) detailliert ausgearbeitet wurde. Grundsätzliche Probleme des „slopes as outcomes"-Ansatzes sind demnach darin zu sehen, daß die Schätzung von spezifischen Innerklassenregressionskoeffizienten auf unzureichenden Stichprobengrößen basiert, wenn man einmal die typische Klassengröße auf 30 ansetzt. Dieser Umstand beeinflußt die Präzision der Schätzung und wirkt sich besonders nachteilig aus, wenn Unterschiede zwischen Innerklassenregressionskoeffizienten auf Signifikanz geprüft werden sollen. Selbst bei günstigen Voraussetzungen (keine „outliers", gut spezifiziertes Modell) kann dies dazu führen, daß inhaltlich sinnvoll interpretierbare Koeffizientenunterschiede über Standardprüfverfahren nicht aufgefunden werden können.

Wenn andererseits diese günstigen Bedingungen nicht vorliegen und die Daten mindere Qualität aufweisen („ill-conditioned data"), sind die Folgen für die Unterschiedsprüfung der β(i)-Koeffizienten nicht mehr vorhersagbar: Es kann in diesem Falle also durchaus sein, daß entweder signifikante oder insignifikante Ergebnisse der Unterschiedsprüfung in die Irre führen.

Sowohl das Keesling & Wiley- wie auch das Burstein et al.-Modell weisen weiterhin das Problem auf, daß Folgen unreliabler Messungen nicht aufgefangen werden können. Die Modellüberprüfungen basieren auf beobachteten Variablen; Aussagen werden aber für zugrundeliegende Konstrukte (latente Variablen) gemacht. Wenn nur ein Indikator zur Repräsentation eines Konstrukts verwendet wird, sind Meßfehler bzw. Reliabilitätsprobleme zu erwarten, die in simplen Regressionsmodellen meist dazu führen, daß die Beziehung zwischen der latenten Prädiktorvariable und der Kriteriumsvariable (z.B. Schulleistung) unterschätzt wird.

Mögliche Auswege aus diesem grundsätzlichen Dilemma der bislang behandelten mehrebenenanalytischen Modelle können darin bestehen, daß entweder robustere Schätzprozeduren eingeführt oder aber komplexere Strukturgleichungsmodelle wie z.B. LISREL zugrundegelegt werden, in denen das Meßproblem explizit angegangen wird. Diese Möglichkeiten sind in einigen neueren Ansätzen exploriert worden, die im folgenden skizziert werden sollen.

12.3. Neuere Ansätze der Mehrebenenanalyse

Es ist an dieser Stelle nicht möglich, die in neuerer Zeit präsentierten mehrebenenanalytischen Entwicklungen umfassend darzustellen; (einige dieser Ansätze sind schon an anderer Stelle in diesem Band angemessen dokumentiert). Von daher werden exemplarisch Verfahrensweisen herausgegriffen, die eine größere Nähe zum eigenen Ansatz aufweisen. Innerhalb des klassischen Regressionsmodells stellt u.E. die Arbeit von Rachman-Moore & Wolfe (1984) eine interessante mehrebenenanalytische Variante dar, die über robuste, nichtlineare Schätzverfahren dem Problem der geringen Stichprobengröße bei den Innerklassenregressionskoeffizientenschätzungen begegnet. Dieses Verfahren soll deshalb kurz beschrieben werden.
Eine vielversprechende Erweiterung des klassischen mehrebenenanalytischen Modells stellen weiterhin Konzeptionen dar, in denen Strukturgleichungsmodelle mit latenten Variablen (LISREL) eingesetzt werden. Wisenbaker & Schmidt (siehe Kapitel 8) gehen ausführlich auf diese Variante ein. Probleme und Möglichkeiten von LISREL bei der Spezifikation mehrebenenanalytischer Modelle sollen hier deshalb nur kurz am Beispiel des Ansatzes von Rock (1985) erörtert werden.

12.3.1 Robuste Schätzung des SVS-Modells

Der Ansatz von Rachman-Moore und Wolfe (1984) zur mehrebenenanalytischen Prüfung des SVS-Modells kann als Elaboration des von Burstein und Mitarbeitern entwickelten „slopes as outcomes"-Modells angesehen werden, das um die bei Keesling & Wiley (1984) vorfindbare Adjustierungslogik erweitert und mit einer adäquaten Schätzprozedur ausgestattet wurde.
Die Autoren gehen bei der Entwicklung des statistischen Modells für Mehrebenendaten vom SVS-Ansatz aus. Für den Fall, daß Y die abhängige Variable auf Schülerebene und X einen Satz von Prädiktorvariablen auf Schülerebene angibt, resultiert für i = 1, ..., n(k) Schüler in k Klassen folgendes allgemeine Modell:

$$Y_{ki} = \alpha_k + \sum_{\mu=1}^{r} \eta_k^u X_{ki}^u + E_{ki} \qquad (7)$$

Hierbei sind $\alpha(k)$ (Intercept für Klasse k) und $\eta(u, k)$ (spezifischer Regressionskoeffizient für Schülermerkmal u innerhalb Klasse k) die zu schätzenden Parameter, während E(ki) als Zufallsvariable aufzufassen ist, die in jeder Klasse den Erwartungswert 0 hat.
Da dieses Modell in der Realität kaum zu schätzen ist (bei mehreren Prädiktorvariablen $\eta(1,k)$, ...,$\eta(r,k)$, würden sicherlich inakkurate Werte bzw. unteridentifizierte Modelle resultieren), führen die Autoren eine Restriktion ein, die die Anzahl der Schätzparameter reduziert. Es wird angenommen, daß Veränderungen der spezifischen Regressionskoeffizienten in unterschiedlichen Schulklassen für die erfaßten Prädiktorvariablen proportional erfolgen müssen, wenn sie auf zugrundeliegende, klassenspezifische Unterrichtsprozesse rückführbar sein sollen. Über die Restriktion

$$\eta_k^u = \eta_k \beta_u \qquad (8)$$

wird die Proportionalität der verschiedenen Regressionskoeffizienten im Prädiktorsatz sichergestellt, was schließlich dazu führt, daß pro Klasse eine einzige zusammengesetzte Erklärungsvariable auf Schülerebene

$$\tilde{X}_{ki} = \sum_{\mu=1}^{r} \beta_u X_{ki}^u \qquad (9)$$

resultiert. In diesem Ansatz wird offensichtlich der Versuch gemacht, die einzelnen Prädiktoren als multiple Indikatoren eines „wahren" Merkmals (z.B. Schülerfähigkeit) zu betrachten: Nur die Regression der Schulleistung auf das „wahre" Merkmal sollte zwischen den einzelnen Schulklassen variieren. Dies stellt zweifellos einen beträchtlichen Unterschied zu den oben diskutierten SVS-Modellen dar. Die Autoren verheimlichen nicht, daß die theoretischen Grundlagen für ihre Annahme dürftig sind und rechtfertigen die Vorgehensweise hauptsächlich mit pragmatischen Vorteilen, die in der Tat nicht von der Hand zu weisen sind.
Eine weitere vorgenommene Restriktion scheint ungleich problematischer: Es wird angenommen, daß die Variation von Regressionskoeffizienten und Intercepts zwischen den Klassen *ausschließlich* auf einen Satz von S Prädiktorvariablen auf Klassenebene (Z) zurückgeführt werden kann:

$$\eta_k = 1 + \sum_{\nu=1}^{s} \delta_\nu z_k^\nu \qquad (10)$$

$$\alpha_k = \mu + \sum_{\nu=1}^{s} \gamma_\nu z_k^\nu + \lambda \bar{\tilde{x}}_{k.} + E_k \qquad (11)$$

Die neu eingeführten Parameter beschreiben den Effekt der Prädiktoren (auf Klassenebene) auf den allgemeinen Innerklassenregressionskoeffizienten (δ) bzw. auf den Klassenintercept (γ) sowie den Effekt der aggregierten zusammengesetzten Prädiktorvariablen auf den Klassenintercept (λ). Mit E(k) und E(ki) sind im statistischen Modell Fehlerkomponenten auf Klassen- wie auf Schülerebene vorhanden. Das restringierte Modell schreibt sich vollständig als

$$Y_{ik} = \mu + \sum_{\nu=1}^{s} \gamma_\nu z_k^\nu + \lambda \bar{\tilde{X}}_{k.} + (1 + \sum_{\nu=1}^{s} \delta_\nu z_k^\nu) \tilde{X}_{ki} + E_k + E_{ki} \qquad (12)$$

Zur Schätzung des in (12) spezifizierten Modells dienen robuste Prozeduren (nicht-lineare Kleinstquadratschätzungen), über die iterativ und in mehreren Stufen die Parameterwerte ermittelt werden.
Robuste Schätzprozeduren sind gerade angesichts des Problems von kleinen Stichproben und sog. „Ausreißern" („outliers") besonders angemessen und stellen einen spezifischen Vorzug des Ansatzes von Rachman-Moore und Wolfe dar. Weitere Vorteile liegen darin, daß komplexe Fehlerstrukturen berücksichtigt werden und die iterative Schätzung einigen Schutz vor verzerrten Parameterwerten bietet. Probleme des Verfahrens liegen sicherlich in der wohl allzu starken Annahme, daß sich Unterschiede in den Innerklassenregressionskoeffizienten ausschließlich auf Erklärungsvariablen der Klassenebene zurückführen lassen. Die Konsequenzen dieser Maßnahme für die Parameterschätzung sind nicht klar. Trotz des Versuchs, über die zusammengesetzte Schülerprädiktorvariable so etwas wie einen „true score" zu etablieren, ist das Problem der adäquaten Repräsentation von latenten Variablen durch beobachtete Indikatoren nicht gelöst.

12.3.2 Mehrebenenanalyse anhand von LISREL

Die Frage nach der angemessenen multiplen Indikatorisierung von latenten Konstrukten, also die Frage nach der Lösung des Meßproblems, wird in Modellen zur Analyse komplexer Kovarinazstrukturen direkt und explizit angegangen. Strukturgleichungsmodelle mit latenten Variablen wie LISREL VI (Jöreskog & Sörbom, 1984) unterscheiden zwischen einem *Meßmodell*, in dem die Beziehung zwischen den beobachteten Indikatoren und den durch sie definierten latenten Variablen spezifiziert wird, und einem *Strukturmodell*, in dem die zwischen den latenten Variablen angenommenen Beziehungen festgelegt werden. Es wird damit möglich, sog. Meßfehler (unreliable Messungen der latenten Variablen) von sog. Spezifikationsfehlern (unvollständige Formulierung des Strukturmodells) zu unterscheiden. Wie schon erwähnt, sind beide Fehlerquellen im klassischen Regressionsmodell konfundiert. Die

Schätzung des Modells erfolgt (in der Regel) über Maximum-Likelihood-Prozeduren, und ein Modelltest liefert verschiedene Indizes zur Beantwortung der Frage, wie genau die beobachtete Kovarianzstruktur durch das spezifizierte Modell reproduziert werden kann.
Vorweg muß betont werden, daß nicht alle theoretisch interessanten Mehrebenenmodelle mit LISREL geschätzt werden können. Nehmen wir etwa als Beispiel das von Mason et al. (1983) formulierte allgemeine SVS-Modell:

$$Y_{ij} = \gamma_{00} + \gamma_{0z}z_j + \gamma_{10}x_{ij} + \gamma_{1z}x_{ij}z_j + (\alpha_{0j} + \alpha_{1j}x_{ij} + \varepsilon_{ij}) \quad (13)$$

Mit LISREL ist es nicht möglich, den Interaktionsterm (der 4. Term) sowie die für unterschiedliche Gruppen als unterschiedlich angenommenen Fehlerstrukturen (der 5. Term), m.a.W. die Heteroskedastizität von Fehlern, angemessen zu verarbeiten. Weiterhin ist zu beachten, daß die bei LISREL-Schätzungen vorwiegend benutzten Maximum-Likelihood-Prozeduren nur für relativ große Stichproben anwendbar sind, also nicht richtig funktionieren, wenn - wie beim SVS-Modell üblich - einzelne Schulklassen verglichen werden (vgl. auch Burstein, 1983). Dies macht deutlich, daß es nicht einfach ist, das SVS-Modell mit LISREL zu kombinieren.
Bevor wir näher auf unseren eigenen Ansatz eingehen, der eine solche Verknüpfung versucht, soll kurz auf ein neueres Anwendungsbeispiel eingegangen werden, in dem LISREL zur mehrebenenanalytischen Kontextanalyse eingesetzt wird (Rock, 1985).
In der Studie von Rock geht es darum, die Mathematikleistungen von High-School-Studenten über Schülerhintergrundsvariablen und Kontextmerkmale der Schule vorherzusagen. Das theoretisch interessante Modell zur Abbildung von Kontext- und Bezugsgruppeneffekten (vgl. Gleichung (3)) wird zunächst spezifiziert, dann aber wegen der nicht zu umgehenden Kollinearitätsprobleme verworfen. Stattdessen wird eine Modellvariante geprüft, die das Kollinearitätsproblem dadurch zu lösen versucht, daß Bezugsgruppen- und Kontexteffekte durch unterschiedliche Merkmale repräsentiert werden (Methode (c) nach Firebaugh, 1979). Rock spezifiziert das Modell als

$$Y_{ij} = \beta_0 + \beta_1\bar{X}_{j.} + \beta_2(X_{ij} - \bar{X}_{j.}) + \beta_3 Z_{j.} + \varepsilon_{ij} \quad (14)$$

Es wird also angenommen, daß die individuelle Mathematikleistung (Y(ij)) abhängt von dem allgemeinen Leistungsstand der Schule ($\bar{X}$(i.)), der relativen Position des Schülers in der Schule (X(ij) - $\bar{X}$(i.)) und dem Einfluß von Kontextmerkmalen, wie den durchschnittlichen Einkommensverhältnissen im Schuldistrikt (Z(i.)).
Da für alle Merkmale multiple beobachtete Indikatoren vorhanden sind, ist die Schätzung und Testung des Modells mit LISREL vorteilhaft: Sie empfiehlt sich umso mehr, als simultane Modellvergleiche für drei ethnische Schüler-

gruppen geplant sind, um die allgemeine Gültigkeit des gewählten Modells zu überprüfen. Als wichtigstes Ergebnis kann gezeigt werden, daß sowohl der Leistungsstand der Schule wie auch die relative Position des Schülers in seiner Schule einen signifikanten Erklärungswert für seine Mathematikleistung hat. Dieser Befund hat für alle untersuchten Schülergruppen Gültigkeit.
Die Vorteile von LISREL konnten in der Arbeit von Rock nicht zuletzt deshalb voll ausgenützt werden, weil eine ausgesprochen große Stichprobe (etwa 12000 Schüler aus mehreren hundert Schulen) zur Verfügung stand. Wie sieht es nun aber mit den Einsatzmöglichkeiten von LISREL aus, wenn die Datenbasis nicht so üppig ist? Dies soll im folgenden Abschnitt anhand eigener Untersuchungen genau erörtert werden.

12.4 Der eigene Ansatz

12.4.1 Grundlegende Überlegungen

Die Zielrichtung der eigenen Analysen ist bei Burstein (1983) treffend zusammengefaßt. Angesichts der (oben beschriebenen) Probleme von LISREL bei der Analyse von SVS-Modellen kommt Burstein zu dem Schluß, daß die Möglichkeiten von LISREL bei der Analyse von Mehrebenendaten am besten genutzt werden können, wenn

(a) auf irgendeine Weise Cluster von Schulklassen gebildet werden;
(b) geprüft wird, ob die Kovarianzstrukturen innerhalb dieser Cluster invariant sind;
(c) im Falle einer Bestätigung von (b) die Daten für jedes resultierende Cluster gepoolt werden und
(d) die Invarianz der Kovarianzstrukturen zwischen den Clustern getestet wird.

Für den Fall, daß Makro-Merkmale (z.B. unterschiedliche Unterrichtsprozesse) als Grundlage der Clusterbildung dienen, kann die signifikante Variabilität der Kovarianzstrukturen zwischen den Clustern auf die Makro-Merkmale zurückgeführt werden.
Im eigenen Ansatz wurde auf die schon erörterte Annahme von Wiley (1970) und Burstein (1980b; Burstein, Linn & Capell, 1978) zurückgegriffen, derzufolge Unterschiede in den Innerklassenregressionskoeffizienten vom Post- auf den Prätest substantielle Unterschiede in Unterrichtsprozessen widerspiegeln können. So mögen beispielsweise fähigere Schüler in Schulklassen mit reichhaltigem Lernangebot ihre Anlagen im Vergleich zu schwächeren Schülern besser nutzen, was die Schüler-Leistungsrangfolge im Zeitraum zwischen Vor-

und Nachtest stabilisieren kann. Als Folge sind (hoch) positive Regressionskoeffizienten für die Posttest-Prätest-Regression zu erwarten. Umgekehrt können in Schulklassen mit eindeutig remedialer/kompensatorischer Instruktionsorientierung schwächere Schüler davon profitieren, daß ihnen vergleichsweise mehr Unterstützung durch den Lehrer zukommt als den besseren Schülern.
In solchen Klassen ist nicht auszuschließen, daß die Prätestleistung die Schulleistung im Nachtest nur noch minimal vorhersagt, es werden also relativ niedrige Innerklassenregressionskoeffizienten resultieren.
In der Arbeit von Schneider und Treiber (1984) interessierte insbesondere die Frage, ob traditionelle Schulleistungsmodelle über unterschiedliche Gruppen bzw. Typen von Schulklassen hinweg allgemeine Gültigkeit besitzen oder ob sie nur lokal valide sind. Das Schulleistungsmodell wurde so konzipiert, daß Hintergrundsmerkmale des Schülers (intellektuelle Fähigkeiten) und Instruktionsmerkmale als exogene (unabhängige) Variablen die Leistungen in einem Mathematiktest vorhersagen sollten, der innerhalb eines Schuljahres insgesamt viermal vorgegeben worden war. An der Untersuchung nahmen insgesamt 113 Schulklassen der 6. Klassenstufe teil.
Aus dieser Gesamtstichprobe wurden über auf Klassenebene durchgeführte Posttest-Prätest-Regressionen Extremgruppen von jeweils vier Schulklassen mit äußerst hohen (High-Slope-) oder flachen (Low-Slope-) Innerklassenregressionskoeffizienten selegiert. Da das erwähnte Schulleistungsmodell multiple Indikatoren für die einzelnen Konstrukte aufwies, bot sich die Analyse mit LISREL auf Individualebene an. Zunächst wurde in einem Simultanvergleich der beiden Extremgruppen (High- vs. Low-Slope-Klassen) geprüft, ob das gleiche Schulleistungsmodell an die Daten beider Gruppen angepaßt werden kann. Da die Indices der Anpassungsgüte unbefriedigende Werte ergaben, wurden in der Folge separate Analysen für beide Klassentypen durchgeführt. Als wichtigstes Ergebnis zeigte sich, daß lediglich für die High-Slope-Klassen eine befriedigende Modellanpassung erzielt werden konnte. Demgegenüber schien das spezifizierte Modell für die Erklärung der Mathematikleistungen in Low-Slope-Klassen mit prinzipiell remedial-kompensatorischer Instruktionsausrichtung nicht angemessen zu sein. Die ungenügenden Anpassungswerte für die Daten dieses Klassentypus legten es nahe, nach alternativen Modellkonstruktionen zu suchen. Die Ergebnisse wurden als Bestätigung der These von Snow (1977) gewertet, wonach Schulleistungsmodelle nur „lokal" anwendbar sind.
In einer nachfolgenden Untersuchung (Schneider & Helmke, 1985) wurde der Versuch unternommen, diese Befunde an einer unabhängigen Schülerstichprobe zu validieren und gleichzeitig einige der bei Schneider & Treiber (1984) vorfindbaren Probleme zu überwinden. Eines dieser Probleme ist zweifellos darin zu sehen, daß die LISREL-Analysen auf Extremgruppen begrenzt wurden und es unterlassen wurde, das Schulleistungsmodell für die Gesamt-

stichprobe zu schätzen und zu testen. Weiterhin erschien der Innerklassenregressionskoeffizient als *alleiniges* Gruppenbildungskriterium fragwürdig, da er z.B. keine Information über unterschiedliche Leistungszuwächse in unterschiedlichen Schulklassentypen gibt. Die Studie von Schneider & Helmke benutzte deshalb als Gruppierungskriterium die Kombination von Post-Prätest-Regressionskoeffizienten und Leistungszuwachs im untersuchten Zeitraum (ca. acht Monate). Zur Identifikation von Schulklassentypen wurden sowohl nicht-hierarchische Clusteranalyse-Algorithmen wie auch ein simpleres Median-Split-Verfahren eingesetzt, um eine vollständige Klassifikation der einbezogenen Schulklassen zu gewährleisten.
Das gewählte Schulleistungsmodell war dem von Schneider und Treiber in struktureller Hinsicht sehr ähnlich, unterschied sich aber in der Indikatorisierung von Unterrichtsmerkmalen: Individuellen Schülerperzeptionen des Unterrichts und der Klassenführung wurde nunmehr der Vorzug vor Beobachtungsdaten gegeben, da letztere nicht auf Individualebene zu verankern sind. Im Unterschied zur Vorgängerstudie gingen aufgrund von Kollinearitätsproblemen (Verwendung *identischer* Mathematiktests zu allen drei Zeitpunkten) lediglich zwei Erhebungen der Mathematikleistung (Vor- und Nachtest) in die Analyse ein. Insgesamt 632 Schüler aus 34 Klassen der 5. Klassenstufe in Hauptschulen nahmen an der Untersuchung teil. Die Daten entstammen dem deutschen Beitrag zur „Classroom-Environment Study: Teaching for Learning (CES)" der IEA (vgl. Weinert & Helmke, 1984; Helmke et al., 1985).
Bei der Suche nach Schulklassentypen zeigte sich zunächst als Problem, daß die Clusteranalyse Ergebnisse lieferte, die inhaltlich nur schwer zu interpretieren waren. Eine besondere Schwierigkeit lag darin, die zu bevorzugende Anzahl von Clustern festzulegen (vgl. zur Diskussion dieses Problems auch Aitkin et al., 1981). Da die Entscheidung für eine spezifische Cluster-Lösung demzufolge arbiträr wirkte, wurde stattdessen der Median-Split-Prozedur der Vorzug gegeben, über die insgesamt vier Schulklassentypen ermittelt wurden, die sich systematisch im Hinblick auf Innerklassenregressionskoeffizienten und Leistungszuwachs (jeweils hoch vs. niedrig) unterschieden. Für die vier resultierenden Gruppen wurde die Vorgehensweise der Vorgänger-Studie repliziert: Auf Schülerebene wurden zunächst Simultanvergleiche via LISREL durchgeführt, die wiederum unzureichende Modellanpassung anzeigten. Die im Anschluß für jede Gruppe getrennt durchgeführten LISREL-Analysen konnten die Befunde von Schneider & Treiber (1984) im wesentlichen bestätigen. Wiederum konnten befriedigende Modellanpassungen nur für die High-Slope-Klassen erzielt werden, wobei betont werden muß, daß dieses Ergebnis nun nicht mehr lediglich für Extremgruppen Gültigkeit hat. Die zusätzliche Berücksichtigung des Leistungszuwachses als Gruppierungskriterium erwies sich insofern als nützlich, als gezeigt werden konnte, daß innerhalb der High- bzw. Low-Slope-Klassen eine wesentlich bessere Daten-

anpassung jeweils für die Gruppe mit hohem Leistungszuwachs erzielt werden konnte. Wesentlich scheint schließlich der Befund zu sein, daß die „overall"-Analyse, bei der alle Schüler ohne Berücksichtigung des Klassenkontextes „gepoolt" wurden, gute Modellanpassungswerte ergab. Dies macht deutlich, wie leicht die Vernachlässigung der Mehrebenenproblematik zu falschen bzw. irreführenden Schlüssen führen kann.
Trotz der insgesamt positiven Befunde macht auch die Studie von Schneider & Helmke deutlich, daß das spezifierte Schulleistungsmodell für die Low-Slope-Klassen nicht adäquat zu sein scheint. Bei der Durchsicht von Deskriptivstatistiken fiel auf, daß sich Selbstkonzept- und Motivationskennwerte in den einzelnen Schulklassengruppen deutlich unterschieden. Eine Erklärungsmöglichkeit für den schlechten Daten-Fit bei Low-Slope-Klassen könnte darin bestehen, daß hier Motivationsmerkmale eine ungleich wichtigere Funktion bei der Erklärung von Schulleistungen haben.
Im folgenden Illustrationsbeispiel wird dieser Hypothese gezielt nachgegangen, indem das zugrundegelegte Schulleistungsmodell um das exogene Merkmal „Motivation" ergänzt wird.

12.4.2 Illustration des eigenen Ansatzes am empirischen Beispiel

Um die Auswirkungen der Modellerweiterung möglichst genau erfassen zu können, wurde im Anwendungsbeispiel auf die Stichprobe und den Variablensatz von Schneider & Helmke (1985) zurückgegriffen. Als Basis dienten demnach wieder die 632 Fünftkläßler aus 34 (bayerischen) Schulklassen; zu Details der zugrundeliegenden Untersuchung vgl. Weinert & Helmke (1984). Das allgemeine Schulleistungsmodell wurde so konzipiert, daß die allgemeine Intelligenz und die Motivation der Schüler sowie die von ihnen wahrgenommene Unterrichtsqualität als unabhängige (exogene) Faktoren bzw. Prädiktoren für die zu zwei Zeitpunkten gemessene Leistung im Fach Mathematik dienten. Jedes der genannten Konstrukte war durch mehrere beobachtete Variablen repräsentiert. Im einzelnen wurden folgende Merkmale erfaßt:

(a) Die Messung der *allgemeinen Intelligenz* erfolgte über den Kognitiven Fähigkeitstest (KFT 4-6) von Heller, Gaedeke & Weinläder (1976) wobei die Subtests Wortbedeutung, Zahlenreihen und Figurenanalogien ausgewählt wurden, die im wesentlichen Sprachverständnis, schlußfolgerndes Denken und Raumvorstellung erfassen sollten.

(b) Der Bereich *Motivation* wurde durch zwei verschiedene Aspekte abgedeckt, und zwar durch Aspekte des Selbstbildes der Begabung, das in theoretischer Hinsicht für die Anstrengungskalkulation und damit für *Lern-*

prozesse von Bedeutung ist sowie durch Facetten der Leistungsangst, die sich auf Störungen bei der Umsetzung des Gelernten im Kontext schulischer Leistungssituation und damit auf *Leistungsprozesse* bezieht. Im einzelnen handelt es sich um die folgenden manifesten Variablen:

(1) das Selbstkonzept der eigenen Begabung für das Fach Mathematik;
(2) die subjektive Einschätzung der eigenen Leistungsfähigkeit in Mathematik, verglichen mit der Klasse (sozialer Vergleichsmaßstab);
(3) die Intensität der Leistungsangst, die in mündlichen und schriftlichen Leistungssituationen in der Schule empfunden wird und
(4) die Häufigkeit von aufgabenirrelevanten, selbstzentrierten (Worry-) Reaktionen bei schriftlichen Arbeiten und Tests.

(c) Der Bereich *Unterrichtsqualität* umfaßte je zwei Komponenten der Klassenführung und der Instruktionsqualität im eigentlichen Sinne. Gemeinsam ist allen Variablen, daß sie aus der Sicht der Schüler erhoben wurden, also Perzeptionen darstellen. Folgende manifeste Variablen sollten das Konstrukt abdecken:

(1) Die Effizienz der Klassenführung, d.h. Angaben darüber, ob der Unterrichtsverlauf reibungslos ist oder durch disziplinarische oder andere Ereignisse häufig unterbrochen wird;
(2) Das Ausmaß der Nutzung der vorhandenen Unterrichtszeit für die Behandlung des Unterrichtsstoffs;
(3) die Klarheit und Verständlichkeit des Mathematikunterrichts aus Schülersicht
(4) der Grad der Individualisierung des Mathematikunterrichts.

Diese Skala umfaßt sowohl Aspekte der Passung zwischen Höhe der unterrichtlichen Anforderungen und subjektiver Bewältigbarkeit (appropriateness) als auch die Häufigkeit remedialen Eingreifens im Falle von Verständnisschwierigkeiten von Schülern.

(d) Die *Mathematikleistung* im Vor- und im Nachtest wurde mit identischen Testaufgaben erhoben. Zur Abbildung des Konstrukts Mathematikleistung dienten zu jedem Zeitpunkt jeweils zwei verschiedene Testsummenwerte, die den Grad der Beherrschung der vier Grundrechenarten (Test 1) sowie den Erfolg bei Textaufgaben wiedergaben, die weniger auf die Beherrschung von Algorithmen als, vielmehr auf das Verständnis und die praktische Anwendung der Grundrechenarten abzielten (Test 2).
Die von Schneider & Helmke (1985) durchgeführte sequentielle Strategie ließ sich verkürzen, da die dort ermittelten vier Gruppen von Schulklassen (gebildet nach dem Median des Innerklassenregressionskoeffizienten und des

Leistungszuwachses) für die LISREL-Analysen aus Gründen der Vergleichbarkeit mit der vorliegenden Studie unverändert übernommen wurden. Der erste Analyseschritt bestand demnach darin, für die vier Schulklassentypen im Rahmen eines simultanen Mehrgruppenvergleichs zu prüfen, ob das zugrundegelegte Schulleistungsmodell generalisierbar war. Der im Vergleich zu den Freiheitsgraden hohe χ^2-Wert sowie unbefriedigende Werte für die übrigen goodness-of-fit-Indices legten es nahe, diese Annahme aufzugeben. Stattdessen wurden im zweiten Schritt separate Analysen für die beiden Varianten der High-Slope- und Low-Slope-Klassen (mit jeweils niedrigem oder hohem Leistungszuwachs) durchgeführt, um die für die einzelnen Gruppen unterschiedliche Anpassungsgüte des gewählten Modells an die Daten zu explorieren.

Wie schon bei der Vorgängerstudie stellte sich heraus, daß die Pfade zwischen Instruktions- bzw. Fähigkeitsfaktor (mit Ausnahme von Gruppe 3) und Posttest-Leistung ohne nennenswerten Informationsverlust weggelassen werden konnten. Gleiches galt für den neu hinzugekommenen Motivationsfaktor, der ebenfalls keine Bedeutung für die Posttest-Leistung hatte. Weiterhin machte eine Inspektion der ersten Ableitungen der Anpassungsfunktion deutlich, daß in allen vier Schulklassentypen eine Meßfehlerkorrelation zwischen den für Vor- und Nachtest korrespondierenden Testwerten zur Beherrschung der Grundrechenarten die Modellanpassung erheblich verbesserte. Die resultierenden Modellschätzungen sind in den Abbildungen 12.1 bis 12.4 angegeben.

Ein Vergleich dieser Befunde mit denen von Helmke & Schneider (1985) macht schnell klar, daß der erhoffte Effekt der Modellerweiterung ausgeblieben ist: Der Einbezug des Motivationskonstrukts erbrachte in keinem einzigen Falle eine verbesserte Modellanpassung, was insbesondere an den goodness-of-fit (GFI-) und root-mean-square-residual-(RMSR-) Indices abzulesen ist. Die latente Motivationsvariable beeinflußt die Prätestleistung zwar in allen vier Schulklassentypen, doch sind die Strukturkoeffizienten mit Werten zwischen .10 und .20 relativ schwach ausgeprägt. Zwar klärt der Einbezug der Motivationsmerkmale im Durchschnitt etwa 5-10% mehr Varianz im Mathematik-Vortest auf, doch reicht dies nicht aus, um den unbefriedigenden Daten-Fit der Low-Slope-Klassen entscheidend zu verbessern. Die nur „lokale" Gültigkeit des zugrundegelegten Schulleistungsmodells wird damit ein weiteres Mal unterstrichen.

Der abschließend durchgeführte globale Prüftest, in den alle Probanden ohne Rücksicht auf die Klassenzugehörigkeit in eine gepoolte Analyse einbezogen wurden, machte jedoch wiederum die Relevanz der mehrebenenanalytischen Betrachtung deutlich. Ähnlich wie in der Untersuchung von Schneider & Helmke (1985) brachte diese Globalanalyse überzeugende Modellanpassungswerte (z.B. einen GFI-Wert von 0.95 und einen RMSR-Wert von 0.04). Dieser Befund suggeriert eine optimale Passung des Modells für alle einbezo-

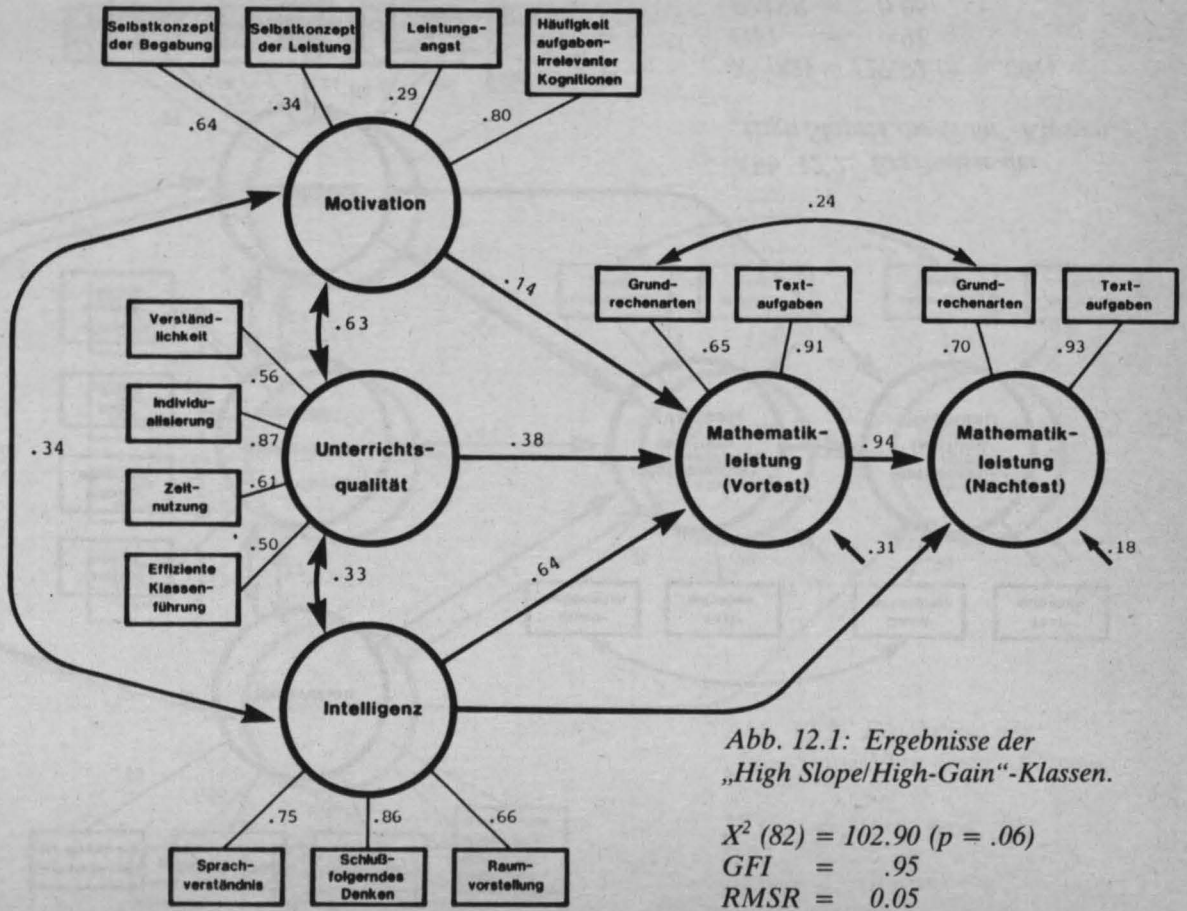

Abb. 12.1: Ergebnisse der „High Slope/High-Gain"-Klassen.

X^2 *(82) = 102.90 (p = .06)*
GFI = .95
RMSR = 0.05

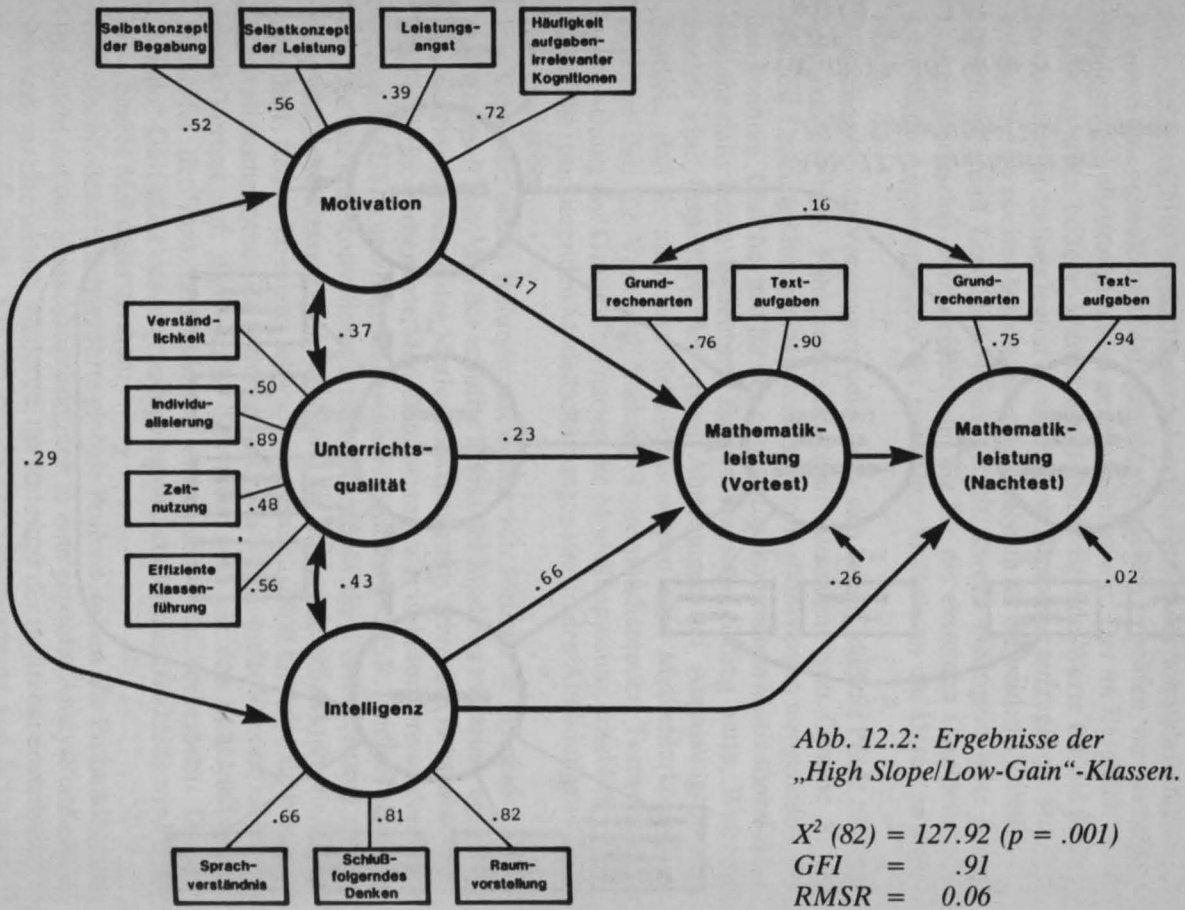

Abb. 12.2: Ergebnisse der „High Slope/Low-Gain"-Klassen.

X^2 (82) = 127.92 (p = .001)
GFI = .91
RMSR = 0.06

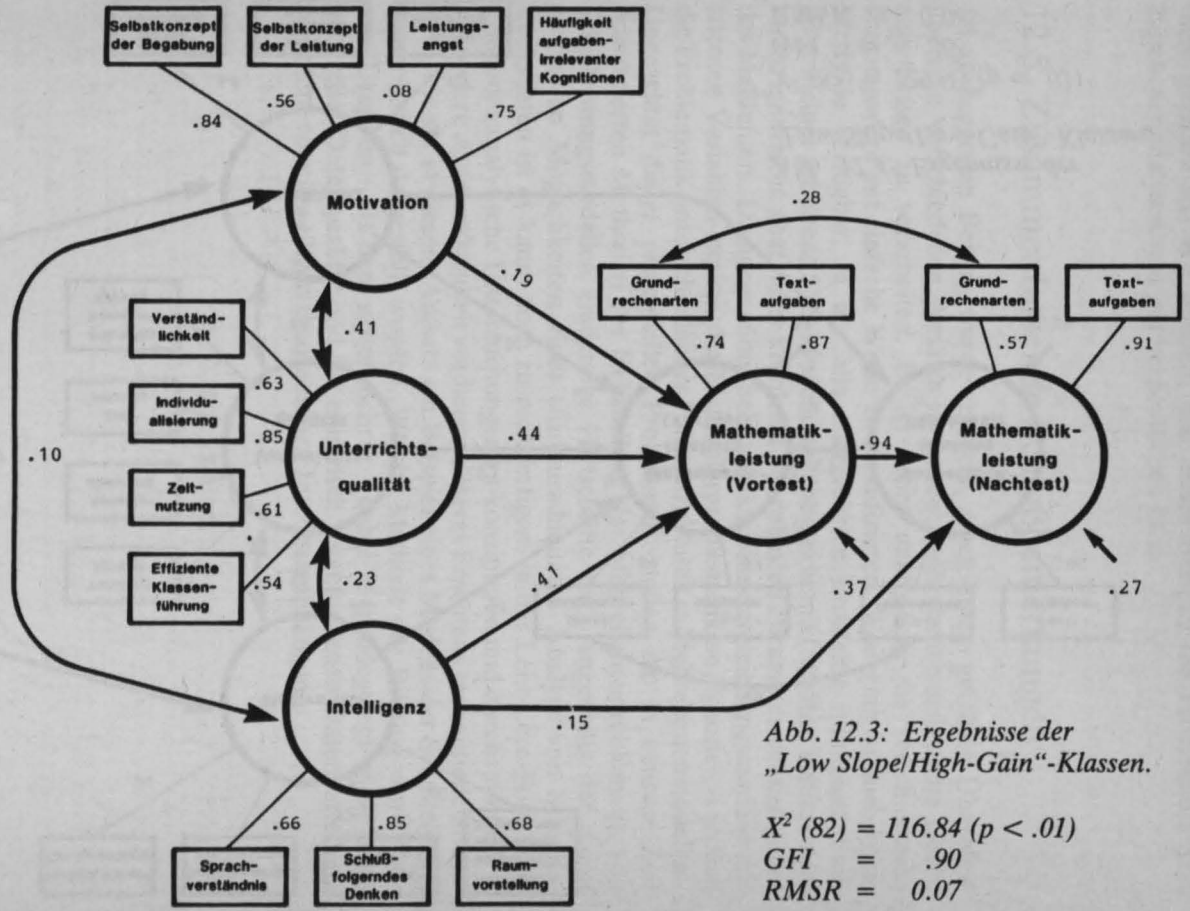

Abb. 12.3: Ergebnisse der „Low Slope/High-Gain"-Klassen.

X^2 (82) = 116.84 ($p < .01$)
GFI = .90
RMSR = 0.07

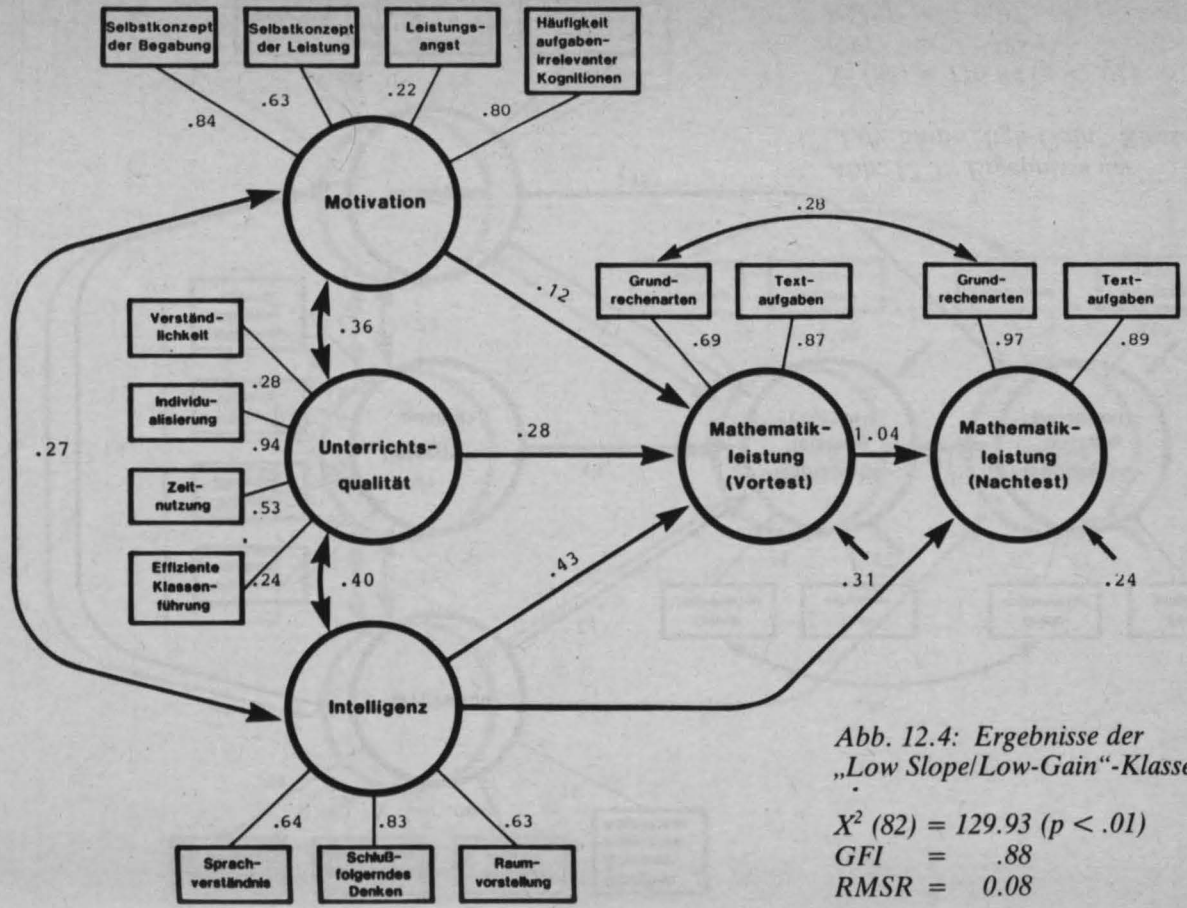

Abb. 12.4: Ergebnisse der „Low Slope/Low-Gain"-Klassen.

X^2 (82) = 129.93 ($p < .01$)
GFI = .88
RMSR = 0.08

genen Probanden und damit eine generelle Gültigkeit, die de facto jedoch nicht gegeben ist. Wir meinen, daß dieses Beispiel die Problematik von Einebenenanalysen sehr gut verdeutlichen kann.

12.5. Zusammenfassende Schlußbemerkungen

Im vorliegenden Beitrag wurde versucht, einen knapp gefaßten Überblick über die verschiedenen Ansätze zu geben, mehrebenenanalytische Daten angemessen zu verarbeiten. Es ist sicher unmöglich, die verschiedenen Zugangswege auf einfache Weise zu klassifizieren. Weiterhin ist auch keine Methode verfügbar, in der alle aufgeführten Probleme minimiert sind. Allgemeiner gehaltene SVS-Modelle arbeiten zwar mit (Schätz-) Fehlerstrukturen, ignorieren aber meist Probleme der robusten Parameterschätzung und des Meßfehlers. Demgegenüber nehmen sich Strukturgleichungsmodelle mit latenten Variablen explizit der Meßfehlerproblematik an, ignorieren jedoch die Problematik unterschiedlicher Fehlerstrukturen in Mehrebenenmodellen. Ungeachtet dieser prinzipiellen Problematik stellen die in neuerer Zeit präsentierten Methoden zur Behandlung der Mehrebenenproblematik von Schulleistungsmodellen eindeutige Fortschritte dar. Angesichts der damit eröffneten Möglichkeiten einer strukturadäquaten Analyse von Schulleistungsdaten ist es kaum noch zu rechtfertigen, in der Lehr-Lern-Forschung einebenenanalytische Untersuchungen zu konzipieren und durchzuführen. Besondere Aufmerksamkeit verdienen unseres Erachtens sequentielle Strategien, wie sie in diesem Aufsatz am Beispiel eines Modells der Schulleistung exemplarisch dargestellt wurden. Welche Methode der Bildung von (Klassen-)Gruppen - als dem ersten Schritt - dabei zugrundegelegt wird, sollte neben der Datenqualität und der empirisch vorgefundenen Datenverteilung vor allem vom jeweiligen theoretischen Interesse abhängen.