Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz

Wolfgang Wagner

vom Promotionsausschuss des Fachbereichs Psychologie der Universität Koblenz-Landau zur Verleihung des akademischen Grades Doktor der Philosophie (Dr. phil.) genehmigte Dissertation

Datum der Disputation: 30. Januar 2008

Vorsitzende des
Promotionsausschusses: Prof. Dr. Annette Schröder

Erstgutachter: Prof. Dr. Andreas Helmke
Zweitgutachter: Dr. Friedrich-Wilhelm Schrader

Inhalt

Zusammenfassung 1
1. Einleitung 2
2. Theoretischer Hintergrund und Fragestellung 6
2.1 Unterricht aus Schülersicht: geteilte und nicht-geteilte Wahrnehmung. Ansätze und empirische Befunde 6
2.1.1 Ansätze der Schul- und Klassenklimaforschung 7
2.1.2 Modelle zur Interpersonalen Wahrnehmung 10
2.1.3 Modelle der kognitiv fundierten Survey-Forschung 12
2.1.4 Reliabilität und Validität von Unterrichtsperzeptionen 15
2.1.4.1 Antwortformat 16
2.1.4.2 Kompositionsmodelle: Theoretische Modelle zur Aggregation von Daten 18
2.1.4.3 Halo-Effekte 21
2.2 Entwicklung der Fragestellung 24
2.3 Fragestellungen 28
3. Datengrundlage und Methoden 30
3.1 Die DESI-Studie (Deutsch Englisch Schülerleistungen International) 30
3.2 Beschreibung der Stichprobe 31
3.3 Beschreibung der Instrumente 32
3.4 Methoden 36
3.4.1 Absolute Übereinstimmung und Reliabilität von Urteilen 36
3.4.2 Mehrebenenanalytische konfirmatorische Faktorenanalysen 42
3.4.3 Simulationsstudie: Spezifikation ordinaler Indikatoren als intervallskaliert in einer konfirmatorischen Zwei-Ebenen-Faktorenanalyse mit je zwei Faktoren auf jeder Ebene 59
4. Ergebnisse 71
4.1 Deskriptive Befunde 71
4.2 Übereinstimmungen von Schülerwahrnehmungen des Unterrichts 78
4.3 Differenziertheit und intraindividuelle Unterschiede (fachspezifische Unterrichtswahrnehmung) 93
4.4 Fachspezifität: Unterscheiden sich die Messmodelle analoger Unterrichtsmerkmale in den Fächern Deutsch und Englisch? 98
4.5 Itemformulierung: Ich- vs. Klassen-Bezug 102
4.6 Determinanten der geteilten und der nicht-geteilten Wahrnehmungskomponenten 114
5. Diskussion 129
5.1 Zusammenfassung der Ergebnisse 129
5.2 Ausblick 134
6. Literatur 142
Verzeichnis der Abbildungen 149
Verzeichnis der Tabellen 150
Anhang A 153
Anhang B 155

Zusammenfassung

In der vorliegenden Untersuchung geht es um methodische Fragen der Unterrichtswahrnehmung aus Schülersicht. Dabei werden theoretische Ansätze zur Urteilsbildung aus der Klassenklima- und der kognitiv fundierten Survey-Forschung sowie der Forschung zur Interpersonalen Wahrnehmung diskutiert. Weiterhin werden Modelle zur inhaltlichen Interpretation von Aggregatmerkmalen (sogenannte Kompositionsmodelle) und zum Einfluss sogenannter „Halo"-Effekte berücksichtigt. Die relevanten Aspekte aus den genannten Theorien sowie empirische Befunde zum Einfluss von Fachleistung, Schulnote und Geschlecht auf die Beurteilung des Unterrichts werden in einem Modell zur Unterrichtswahrnehmung aus Schülersicht zusammengeführt. Daneben werden in der vorliegenden Untersuchung Möglichkeiten und Grenzen verschiedener statistischer Verfahren zur Analyse von Daten zur Wahrnehmung des Unterrichts aus Schülersicht aufgezeigt und diskutiert. Dabei geht es um Fragen der absoluten Übereinstimmung von Urteilern vs. der Reliabilität von Urteilen sowie um Grundlagen und (Effekte von Verletzungen der) Annahmen mehrebenenanalytischer konfirmatorischer Faktorenanalysen.

Datengrundlage der vorliegenden Untersuchung sind im Rahmen des Projekts DESI (Deutsch Englisch Schülerleistungen International) der KMK[1] erhobene Fragebogendaten aus 330 Klassen bzw. Kursen der neunten Jahrgangsstufe, sowie im Längsschnitt erhobene Testleistungen in den Bereichen Deutsch und Englisch. Die Ergebnisse der Analysen bestätigen in großen Teilen das zugrunde gelegte theoretische Modell: Es zeigen sich hohe relative Übereinstimmungen der unterrichtsbezogenen Urteile von Schülerinnen und Schülern. Der Einfluss der Kommunikation mit Mitschülern auf die Urteilsbildung zeigt sich v.a. in der Gruppe der Mädchen. Die theoretischen Unterrichtsmerkmale lassen sich als Faktoren auf beiden Analyseebenen (innerhalb von Klassen und zwischen Klassen) nachweisen, wobei diese – bis auf wenige Ausnahmen – ebenen- sowie fachübergreifend nur im Sinne analoger und nicht „isomorpher" Konstrukte interpretierbar sind. Daneben finden sich deutliche Hinweise auf eine eher Lehrkraft- statt Unterrichtsfach-bezogene Wahrnehmung auf beiden Ebenen. Der Itemformulierung (Ich- vs. Klassen-Bezug) kommt insgesamt betrachtet eine eher geringe Bedeutung zu. Bezüglich der Unterrichtswahrnehmungen finden sich Einflüsse der oben genannten Prädiktoren insbesondere auf der Ebene innerhalb von Klassen. Auf Klassenebene zeigt sich ein möglicher Einfluss einer „milden" Benotung auf die geteilte Unterrichtswahrnehmung.

Die auf beiden Ebenen deutlichsten Effekte auf die Unterrichtswahrnehmung finden sich bei den Globalurteilen bezüglich der jeweiligen Lehrkraft. Zusammengenommen mit den extrem hohen Interkorrelationen der Faktoren *innerhalb eines Fachs* (die sich jeweils auf dieselbe Lehrkraft beziehen) und den – aufgrund von Effekten auf Schulebene etc. erwartungswidrigen – niedrigen bis nicht vorhandenen fachübergreifenden Interkorrelationen spricht dies in hohem Maße für eine Verzerrung der Unterrichtswahrnehmung im Sinne einer globalen Wahrnehmungstendenz. Die globale Wahrnehmung muss allerdings gleichzeitig überwiegend als Ergebnis von Unterrichtswahrnehmungen betrachtet werden: Hier zeigen sich die auf der Basis empirischer Untersuchungen theoretisch postulierten unterschiedlichen Gewichtungen der einzelnen Unterrichtsmerkmale über verschiedene Klassen hinweg.

[1] Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland

1. Einleitung

Die durch die TIMS-Studie[2] angestoßene und durch PISA[3] konsolidierte empirische Wende der deutschen Bildungspolitik hat in den letzten Jahren zu einem enormen Wandel im Bereich der Qualitätssicherung von Schulen geführt. Zu nennen sind insbesondere die dauerhafte Beteiligung Deutschlands an internationalen Surveys (wie TIMSS und PISA), die Initiierung von flächendeckenden Lernstandserhebungen (von der MARKUS-Studie[4] in Rheinland-Pfalz bis hin zu den nunmehr bundesweit flächendeckenden Vergleichsarbeiten, VERA), die Entwicklung von länderübergreifend verbindlichen Bildungsstandards und die Etablierung von Institutionen, zu deren Aufgaben primär die Bestandsaufnahme (*monitoring*), Sicherung und Verbesserung der Schulqualität zählt; hier sind das IQB[5] sowie die Qualitätsagenturen der Bundesländer zu nennen, in Rheinland-Pfalz die AQS[6].

Obwohl die empirische Wende mit Projekten begann, die auf den Ertrag von Schule und Unterricht abzielten (*output*, meist zentriert auf fachliche Kompetenzen und Leistungen), setzt sich zunehmend der Gedanke durch, dass Kompetenzunterschiede (zwischen Klassen, Schulen und auch Systemen) nicht verständlich sind, wenn man nichts über die Prozesse weiß, die dorthin geführt haben. Mit diesen Prozessen ist vor allem der Unterricht gemeint, also das unbestrittene Kerngeschäft der Schule. Dies hatte Konsequenzen für die zweite Generation der großen Vergleichsstudien, die in besonderem Maße den Bereich „Unterrichtsqualität" fokussierten. Vor allem ist hier das Projekt DESI der KMK zu nennen. Dazu kommt: Die in nahezu allen Bundesländern inzwischen eingerichteten Agenturen zur *externen Evaluation* der Schule stimmen darin überein, dass der Unterricht von allen geprüften Qualitätsbereichen der Kernbereich ist. Zur Evaluierung werden neben Unterrichtsbeobachtungen (je nach Bundesland unterschiedlich: durch Inspektoren, Evaluationsteams, Referenten oder Visitatoren) durchweg auch Schülerbefragungen zur Qualität des Unterrichts eingesetzt.

Ein weiterer Trend besteht darin, dass die schulinterne Qualitätsentwicklung und interne *Evaluation des Unterrichts* sich zunehmend derjenigen Instrumente bedient, die für die externe Evaluation entwickelt worden waren bzw. aus den großen Unterrichtsprojekten (wie DESI und MARKUS) abgeleitet wurden. Man spricht hier auch von „Schülerfeedback" (vgl. A. Helmke, 2003). Ein typisches Beispiel sind etwa die von IQESonline[7] (iqesonline.net) angebotenen Instrumente zur Erfassung der Qualität des Deutsch-, Englisch- und Mathematik-

[2] Third International Mathematics and Science Study
[3] Programme for International Student Assessment
[4] Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext
[5] Institut zur Qualitätsentwicklung im Bildungswesen (Berlin)
[6] Agentur für Qualitätssicherung (Mainz)
[7] IQES steht für „Instrumente für die Qualitätsentwicklung und Evaluation in Schulen".

unterrichts aus Schülersicht, an denen der Autor dieser Untersuchung maßgeblich mitgewirkt hat.

Schließlich soll darauf hingewiesen werden, dass die folgenreiche Arbeit der Qualitätsagenturen in den deutschen Bundesländern durch zwei – für die vorliegende Arbeit relevante – Besonderheiten gekennzeichnet ist: (1) Zur Ausbildung und Professionalisierung des Personals werden *Unterrichtsvideos* eingesetzt und mithilfe von kategorienbasierten Rating-Bögen beurteilt. Bei der Auswertung der resultierenden Daten (die dafür erforderlichen Programme wurden vom Autor dieser Untersuchung entwickelt) zeigen sich bemerkenswerte Ergebnisse, insbesondere lassen sich massive urteilerspezifische Urteilsvoreingenommenheiten nachweisen, und je nach Unterrichtsmerkmal gibt es zwischen den Ratern teilweise erheblichen Dissens bei der Beurteilung ein und des gleichen Unterrichts. (2) Die meisten Qualitätsagenturen, so auch die AQS in Rheinland-Pfalz, realisieren ein *mehrperspektivisches* Vorgehen, d.h. sie erheben zum Unterricht sowohl Beobachtungsdaten (Unterrichtsbesuch) als auch Schülerangaben (Fragebögen). Man spricht hier auch von „Triangulation". Bei der Berichterstattung taucht dann gelegentlich das Problem auf, wie unterschiedliche Sichtweisen (z.B. zum „lernförderlichen Klima" oder zum „Klassenmanagement") zu interpretieren sind. Gelegentlich werden sogar zu den äquivalenten Konstrukten neben Beobachtungsdaten und Schülerangaben zusätzlich noch Lehrerangaben erhoben, so im Kanton Luzern (Schweiz).

Die mit einem mehrperspektivischen Vorgehen verbundenen theoretischen und vor allem methodischen Probleme sind jedoch derzeit keineswegs gelöst; vielfach werden sie als Problem nicht einmal erkannt. Das Gleiche gilt für den Sachverhalt, dass die Angaben verschiedener Urteiler (z.B. Schülerinnen und Schüler oder Inspektoren) zum gleichen Unterricht oft erheblich streuen. Früher hat man es sich einfach gemacht und diese Differenzen entweder ignoriert (indem man sich auf ein Mittelwertprofil beschränkte) oder als unvermeidliches „Rauschen" bezeichnet. Im Gegensatz dazu versucht die vorliegende Arbeit, ausgewählte methodische Fragen zu behandeln, deren Beantwortung für das Verständnis der Gütekriterien von unterrichtsbezogenen Urteilen wichtig ist.

Eine Sichtung des aktuellen Forschungsstandes zeigt, dass die Zusammenhänge zwischen Unterrichtsbeurteilungen mit verschiedenen Methoden bzw. aus unterschiedlichen Perspektiven in der Regel gering sind (vgl. z.B. Clausen, 2002; Kunter & Baumert, 2006). Diese Ergebnisse haben zum einen dazu geführt, dass der „Wahrnehmungsperspektive" zunehmend eine eigenständige Validität zugesprochen wird, dass also die Abweichungen zwischen verschiedenen Perspektiven *nicht* als zu minimierende Messfehler zu behandeln sind. Zum ande-

ren wird vorgeschlagen, die Frage nach der optimalen Quelle *bezüglich des jeweils zu erfassenden Konstrukts* in den Vordergrund zu rücken – anstatt generelle Bewertungen der unterschiedlichen Perspektiven der Unterrichtswahrnehmung vorzunehmen (vgl. Kunter & Baumert, 2006).

Die Betrachtung der unterschiedlichen Sichtweisen als spezifisch valide ist insbesondere dann einleuchtend, wenn es sich bei den zu erfassenden Konstrukten um sogenannte „Klimawahrnehmungen" handelt. Bei stark verhaltensbezogenen Unterrichtsmerkmalen ist dies dagegen weniger plausibel (vgl. Babad, 1996): Wie häufig beispielsweise eine Lehrkraft zu Beginn einer Unterrichtseinheit eine Übersicht über den Gegenstand bzw. den Ablauf der Stunde gibt, sollte sich eigentlich objektiv feststellen lassen. Der subjektive Anteil der Perspektive könnte eventuell unterschiedliche Auffassungen darüber repräsentieren, was unter „Übersicht" zu verstehen ist (Reicht etwa eine beiläufige Bemerkung bereits aus?) bzw. welcher Referenzzeitraum zugrunde gelegt wird (seit Beginn des Schuljahres, im letzten Halbjahr etc.). Wobei hier wohl nicht nur zwischen, sondern auch innerhalb der jeweiligen Perspektive mit einem gewissen Grad an Variabilität zu rechnen ist. Aber auch solche Interpretationsschwierigkeiten ließen sich vermutlich bis zu einem gewissen Grad durch präzisere Formulierungen ausräumen.

Bei einer Generalisierung der Perspektivenspezifität der Unterrichtswahrnehmungen besteht im Einzelfall die Gefahr, dass Messfehler – oder schlimmstenfalls sogar spezifische Verzerrungstendenzen – als valide Varianzquellen betrachtet werden, anstatt eine Optimierung der jeweiligen Instrumente zu erwägen. Auch für die Interpretation von Ergebnissen und ihre praktische Umsetzung kann die Annahme perspektivenspezifischer Wahrnehmungen problematisch sein: Ein positiver Effekt eines aus Schülersicht gut strukturierten Unterrichts auf ein Zielkriterium kann nicht ohne weiteres auf eine „objektiv" hoch ausgeprägte *Strukturiertheit* zurückgeführt werden. Diese Feststellung stimmt mit dem erweiterten Prozess-Produkt-Paradigma überein, demzufolge Prozessmerkmale des Unterrichts ohnehin keine „direkten" Effekte auf Schülerseite haben; vielmehr sind entsprechende Mediationsprozesse auf Schülerseite zu beachten. Man spricht deshalb auch vom „*mediated processes paradigm*" (Doyle, 1977; Winne & Marx, 1982). Was aber folgt dann aus einem solchen Ergebnis? Lehrkräfte können versuchen, ihren Unterricht – aus ihrer Sicht – strukturierter zu gestalten. Wie aber können Lehrkräfte die Wahrnehmungen ihrer Schülerinnen und Schüler bezüglich der *Strukturiertheit* ihres Unterrichts verändern? Anders ausgedrückt: Was können sie tun, um zu vermeiden, dass ihre didaktischen Aktivitäten (Fragen, Aufforderungen, Anweisungen etc.) nicht versickern oder missverstanden werden? Hierzu wäre eine Theorie der Perspektivenspe-

zifität der Unterrichtswahrnehmung erforderlich, aus der sich entsprechende Handlungsanweisungen ableiten lassen.

Die vorliegende Abhandlung geht den theoretischen und methodischen Besonderheiten bei der Erfassung des *Unterrichts aus Schülersicht* nach. Im Zentrum stehen dabei Fragen der (Interrater-)Reliabilität und Validität der Unterrichtswahrnehmung. Dabei werden auch in der Literatur häufig genannte (verzerrende) Einflussfaktoren auf die Wahrnehmung des Unterrichts von Schülerinnen und Schülern berücksichtigt. Weiterhin wird die „Isomorphie" ausgewählter Konstrukte auf den Analyseebenen innerhalb von Klassen bzw. zwischen Klassen sowie bezüglich der Unterrichtsfächer Englisch und Deutsch untersucht. Dabei geht es um die Frage, ob die jeweiligen Konstrukte inhaltlich auf den verschiedenen Ebenen bzw. in unterschiedlichen Fächern gleichermaßen interpretiert werden können. Die parallele Erfassung identischer Unterrichtsmerkmale in den beiden genannten Unterrichtsfächern bei allen Schülerinnen und Schülern ermöglicht auch die Betrachtung intraindividueller Unterschiede, die ebenfalls Gegenstand der vorliegenden Untersuchung sind. Daneben wird der Frage nachgegangen, welche Bedeutung der Itemformulierung (Ich-Bezug vs. Klassen-Bezug) bei der Erfassung von Unterrichtsmerkmalen zukommt.

2. Theoretischer Hintergrund und Fragestellung

Im Folgenden werden zunächst theoretische Ansätze und empirische Ergebnisse, die sich auf die Unterrichtswahrnehmung aus Schülersicht beziehen bzw. übertragen lassen, dargestellt (Kap. 2.1). Die zentralen Befunde werden anschließend in Kapitel 2.2 zusammengefasst. Die aus diesen Befunden abgeleiteten Fragestellungen der vorliegenden Untersuchung sind Gegenstand des letzten Teilkapitels (Kap. 2.3).

2.1 Unterricht aus Schülersicht: geteilte und nicht-geteilte Wahrnehmung. Ansätze und empirische Befunde

In der vorliegenden Untersuchung geht es um die Erfassung von Unterrichtsqualität. Zur Betonung der spezifischen Beurteiler-Perspektive und einer Vermeidung des stark metaphorischen „Klima"-Begriffs wird hier – in Anlehnung an Clausen (2002) – das Konzept „*Unterrichtswahrnehmungen aus Schülersicht*" verwendet. Um die jeweilige Analyseebene (innerhalb von Klassen, zwischen Klassen) zu kennzeichnen, wird zusätzlich zwischen „*nicht-geteilter*" (subjektiver, idiosynkratischer) und „*geteilter*" (kollektiver) Wahrnehmung unterschieden (vgl. Kenny, 2004).

Üblicherweise werden von Schülerinnen und Schülern wahrgenommene Unterrichtmerkmale mithilfe von Fragebogenskalen (meist Items mit *multiple choice*-Antwortformat) erfasst. Übersichten über eine Vielzahl bereits empirisch erprobter Skalen finden sich beispielsweise bei Gruehn (2000) und Clausen (2002).

Als Vorteile der Erfassung des Unterrichts aus Sicht der Schülerinnen und Schüler werden häufig genannt (vgl. z.B. Clausen, 2002; A. Helmke, 2006):

1. geringer Erhebungsaufwand
2. ökonomische Durchführung
3. geringer Kostenaufwand
4. Einschätzungen basieren auf großer Verhaltensstichprobe
5. hohe Reliabilität durch Aggregation von Daten
6. relativ hohe prädiktive Validität

Daneben wird auf folgende negative Aspekte bei der Unterrichtsbeurteilung aus Schülersicht hingewiesen (s. ebd.):

1. hohes Verzerrungspotential (Halo-Effekte, mögliche Beeinflussung der Wahrnehmungen durch verschiedene Personmerkmale)
2. Überforderung aufgrund mangelnder didaktischer Kompetenz
3. Bewertungsmaßstab ist oft unklar, ebenso der zugrunde gelegte Zeitraum

Häufig werden Skalen zum Unterricht aus Schülersicht in aggregierter Form (z.B. auf Klassenebene) als Prädiktor für verschiedene Zielkriterien (wie z.B. Leistung, Lernmotivation etc.) verwendet. Im Zusammenhang mit der Aggregation solcher Unterrichtswahrnehmungen stellen sich allerdings verschiedene Fragen: Ist die Annahme einer kollektiven Wahrnehmung gerechtfertigt? Wenn ja, welche (z.B. kognitiven, sozialen) Prozesse liegen dieser geteilten Wahrnehmung – für die das Aggregat steht – zugrunde? Welche Prozesse sind verantwortlich für die Entstehung der nicht-geteilten Wahrnehmungen? Diese Fragen sind Gegenstand der folgenden Unterkapitel (Kapitel 2.1.1-2.1.4).

2.1.1 Ansätze der Schul- und Klassenklimaforschung

Aus den zahlreichen Theorien der Schul- und Klassenklimaforschung (s. dazu Dreesmann, 1979, 1982; Saldern, 1987) wird hier der Ansatz von Dreesmann (1979, 1982) ausgewählt, da Dreesmann als erster den Versuch unternommen hat, eine durch explizite theoretische Annahmen geleitete Klassenklimaforschung zu betreiben (Saldern, 1987; Gruehn, 2000). Dreesmann befasst sich dabei auch mit der Frage, wie aus individuellen Klimawahrnehmungen ein kollektives (Klassen-)Klima resultiert.

Bei der Entstehung des Klassenklimas sind nach Dreesmann (1982, S. 52) zwei Prozesse beteiligt: (1) Die kognitive Verarbeitung des Unterrichts durch die individuellen Schülerinnen und Schüler einer Klasse. (2) Ein gruppendynamischer Prozess, bei dem mit zunehmender Kommunikation mit Mitschülern, gemeinsamen Erlebnissen, Abgrenzungsverhalten anderen Klassen gegenüber etc. im Laufe der Zeit gemeinsame Erlebens- und Verarbeitungsstrukturen aufgebaut werden. Das „Unterrichtsklima“ ist also das Resultat dieses gruppendynamischen Prozesses, der immer auch mit den individuellen kognitiven Repräsentationen des Unterrichts verschränkt ist. Dreesmann betont: „Die Schüler werden als aktive Interpreten ihrer Umwelt begriffen, die im sozialen Kontext ihrer Klasse stehen; dieser Kontext beeinflusst die Schüler und umgekehrt sie ihn“ (S. 55).

Bezüglich der individuellen Determinanten der Wahrnehmung des Unterrichtsgeschehens sind Dreesmann (S. 58) zufolge – unter Rückgriff auf ein Modell zur Personenwahrnehmung von Warr und Knapper (1968) – drei verschiedene Komponenten von besonderer Bedeutung:

1. Bei der *attributiven Komponente* spielen vor allem die Art, wie wahrgenommenes Verhalten mit zugrunde liegenden Dispositionen verknüpft wird, sowie die Attribuierungstendenz des Wahrnehmenden – ob er andere und sich selbst als Verursacher oder als von äußeren Faktoren abhängig sieht – eine wichtige Rolle.

2. Die *Erwartungskomponente* bezieht sich auf die den Wahrnehmungsurteilen inhärenten Einstellungen auf zukünftige Interaktionen oder Situationen.
3. Eine besondere Bedeutung kommt der *affektiven Komponente* zu: Sympathie bzw. Antipathie beeinflussen in hohem Maße die wahrgenommenen Eigenschaften einer Person.

Neben diesen individuellen Merkmalen übt auch der *soziale Kontext* innerhalb einer Klasse einen Einfluss auf die Wahrnehmung des Unterrichts aus (Dreesmann, 1982, S. 61). Da der soziale Kontext in einen institutionell vorgegebenen Rahmen eingebettet ist, ist das Verhalten von Schülerinnen und Schülern sowie von Lehrkräften bereits in gewissem Umfang vorgegeben. Damit verbunden sind auch bestimmte soziale *Rollen*: Die Lehrkraft muss als „formale und sachliche Autorität" (ebd.) lehren und erziehen, während von den Schülerinnen und Schülern erwartet wird, dass sie den Anweisungen der Lehrkraft folgen und ihr Lernverhalten auf unterrichtliche Ziele richten.

Betrachtet man die Schüler-Schüler-Beziehungen innerhalb einer Klasse, so ist hier insbesondere in unsicheren Situationen – wenn ein Schüler etwa bei der Beurteilung der Schwierigkeit des Unterrichtsstoffs unsicher ist – eine Beeinflussung der Wahrnehmung durch die Mitschüler zu erwarten (S. 63). Insgesamt sollte, so Dreesmann, die soziale Dynamik innerhalb einer Klasse im Laufe der Zeit zu einer Homogenisierung bezüglich der jeweiligen Unterrichtswahrnehmungen führen (S. 64).

Sind die „Überlappungen" der einzelnen Schülerwahrnehmungen des Unterrichts in Klassen allerdings sehr gering, so schlägt Dreesmann vor, solche Klassen von Analysen auszuschließen, da es in diesen Fällen kein Unterrichtsklima gebe und somit auf Aggregatebene nur Artefakte produziert würden. Gruehn (2000, S. 67) weist jedoch darauf hin, dass es immer ein Unterrichtsklima geben muss, da sowohl die kognitiven als auch die gruppendynamischen Prozesse – die Konstituenten des Klimas – allgegenwärtig sind. Eine Lösung dieses Problems schlägt Satow (1999, S. 42) vor: Er bezieht in seine Analysen sowohl die geteilte Wahrnehmung des Unterrichts (Klassenmittelwerte) als auch die klasseninternen Streuungen ein.

Bezogen auf das Ausmaß der Übereinstimmung der Unterrichtswahrnehmungen von Schülerinnen und Schülern spielen die Iteminhalte eine wesentliche Rolle: Verhaltensbezogene und konkrete (beschreibende) Items werden von Schülerinnen und Schülern eher aus der Sicht „neutraler Beobachter" beantwortet, was zu höheren klasseninternen Übereinstimmungen führen sollte (vgl. Dreesmann, 1982, S. 113). Im Allgemeinen werden jedoch Items mit unterschiedlich hohem Abstraktionsniveau erhoben, so dass Schülerinnen und Schüler zu-

gleich als Beobachter und als Beurteiler fungieren (S. 114). Neben dem Inferenzgrad (niedrig- vs. hoch-inferent) ist die Wahrnehmungsperspektive bei der Itemformulierung von großer Bedeutung (vgl. Eder, 2006). Den Brok, Brekelmans und Wubbels (2006) unterscheiden zwischen dem Adressaten des Lehrerverhaltens (die Klasse oder ein individueller Schüler) und der Wahrnehmungsperspektive (Einschätzung aus Sicht der Klasse oder personalisierte Perspektive).

In Tabelle 1 sind die drei genannten Faktoren (Adressat des Lehrerverhaltens, Wahrnehmungsperspektive und Inferenzgrad) sowie entsprechende Itembeispiele dargestellt. Beim Inferenzgrad ist zu beachten, dass es sich hier um eine kontinuierliche Skala mit den Polen niedrig- und hoch-inferent handelt. Bezüglich der Wahrnehmungsperspektive „aus Sicht der Klasse" ist möglicherweise die Annahme nicht gerechtfertigt, dass Items ohne spezifischen Bezug zum Individuum („Ich finde…") grundsätzlich zu einer Perspektivenübernahme Anlass geben, also Schülerinnen und Schüler durch solche Formulierungen tatsächlich veranlasst werden, bei der Beantwortung die Sichtweise(n) ihrer Mitschülerinnen und -schüler einzubeziehen. Dies ließe sich durch eine Formulierung wie etwa „Wir finden, …" – analog zur individuellen Sicht – sicherstellen.

Tabelle 1: Adressat des Lehrerverhaltens, Wahrnehmungsperspektive und Inferenzgrad bei Itemformulierungen zur Unterrichtswahrnehmung aus Schülersicht

		Wahrnehmungsperspektive			
		aus Sicht der Klasse		aus individueller Sicht	
Adressat des Lehrerverhaltens	**Inferenzgrad**	niedrig-inferent	hoch-inferent	niedrig-inferent	hoch-inferent
	Klasse	*Beispielitem*: Die Lehrkraft benotet unsere schriftlichen Arbeiten fair.	*Beispielitem*: Die Lehrkraft behandelt uns fair.	*Beispielitem*: Ich finde, die Lehrkraft benotet unsere schriftlichen Arbeiten fair.	*Beispielitem*: Ich finde, dass die Lehrkraft uns fair behandelt.
	individuelle Schülerin / individueller Schüler	*Beispielitem*: Die Lehrkraft benotet meine schriftlichen Arbeiten fair.	*Beispielitem*: Die Lehrkraft behandelt mich fair.	*Beispielitem*: Ich finde, die Lehrkraft benotet meine schriftlichen Arbeiten fair.	*Beispielitem*: Ich finde, dass mich die Lehrkraft fair behandelt.

Theoretisch bezieht sich der Faktor „Adressat des Lehrerverhaltens" auf mögliche Lehrer-Schüler-Interaktionseffekte: Die (individuelle Wahrnehmung der) „Fairness" bezüglich der Benotung schriftlicher Arbeiten kann sich beispielsweise von Schüler(in) zu Schüler(in) unterscheiden. Eine solche Beurteilung würde aber voraussetzen, dass die jeweilige Schülerin bzw. der jeweilige Schüler die schriftlichen Arbeiten aller Mitschülerinnen und Mitschüler selbst gesehen und beurteilt hat. Dies ist allerdings sehr unwahrscheinlich. Bei der Beantwor-

tung einer solchen Frage ist deshalb davon auszugehen, dass Schülerinnen und Schüler auf die (ihnen bekannten) Einschätzungen ihrer Mitschülerinnen und Mitschüler zurückgreifen. In diesem Fall sind demnach der Adressat des Lehrerverhaltens (Klasse) und die Wahrnehmungsperspektive (aus Sicht der Klasse) weitgehend identisch.

Analog dazu wären im obigen Beispiel die individuelle Schülerin bzw. der individuelle Schüler als Adressat des Lehrerverhaltens und die Wahrnehmungsperspektive „aus individueller Sicht" identisch. Denn die Beantwortung der Frage, wie „fair" die Mitschülerinnen und Mitschüler die Benotung der schriftlichen Arbeiten einer bestimmten Schülerin bzw. eines bestimmten Schülers finden, würde ebenfalls voraussetzen, dass diesen alle Arbeiten bekannt sind. Insofern dürfte die Unterscheidung zwischen den Faktoren „Wahrnehmungsperspektive" und „Adressat des Lehrerverhaltens" häufig nur theoretisch möglich sein.

2.1.2 Modelle zur Interpersonalen Wahrnehmung

Modelle der Interpersonalen Wahrnehmung müssen Kenny (2004) zufolge in der Lage sein, drei Ergebnisse der empirischen Konsens-Forschung zu erklären:

1. Der Grad der Übereinstimmung zwischen verschiedenen Ratern bezüglich der Beurteilung einer Person ist relativ gering.
2. Der Konsens nimmt mit zunehmendem Bekanntschaftsgrad mit der zu beurteilenden Person kaum zu.
3. Konsens zwischen verschiedenen Beurteilern lässt sich auch dann nachweisen, wenn diese die zu beurteilende Person kaum kennen.

Auf der Basis zweier bereits von Kenny zuvor entwickelten Modelle – dem *Weighted-Average*- und dem *Social-Relations*-Modell – entwirft Kenny (2004) das sogenannte PERSON-Modell (personality, error, residual, stereotype, opinion, norm). Er unterscheidet insgesamt sechs verschiedene Komponenten der interpersonalen Wahrnehmung, von denen sich jeweils drei auf die *geteilte* bzw. auf die *nicht-geteilte Wahrnehmung* beziehen. Weiterhin differenziert Kenny zwischen *verhaltensbezogenen* und auf *kategoriale*[8] *Variablen bezogenen* (also der physischen Erscheinung, wie z.B. das Geschlecht der zu beurteilenden Person) sowie zwischen *konsistenten* und *inkonsistenten* Urteilen. Im Folgenden werden die einzelnen Komponenten, die bei der Beurteilung einer Person durch verschiedene Beurteiler eine Rolle spielen (also z.B. einer Lehrkraft durch die Schülerinnen und Schüler einer Klasse), kurz dargestellt:

[8] Unter kategorialen Informationen versteht Kenny im Wesentlichen das, was üblicherweise als „Stereotyp" bezeichnet wird. Kenny (2004, S. 268) weist jedoch auf zwei Unterschiede hin: „First, some stereotypes are triggered by behavior, and the stereotype is not information but is an inference. Second, some stereotypes are not immediately learned on the first meeting and are discovered much later on (e.g., that the target is gay)."

1. *Stereotype* bezieht sich auf die geteilte Wahrnehmung bezüglich einer kategorialen Information (also z.B. das Geschlecht der Lehrkraft).
2. *Personality* steht für das verhaltensbasierte Urteil aus Sicht verschiedener Urteiler. Im Prinzip geht es hier um ein überdauerndes Personmerkmal im Sinne eines *traits*.
3. Unter *Norm* versteht Kenny die einzigartige Bedeutung, die Beurteiler übereinstimmend einem Verhaltensakt, also einer Sequenz aus dem „Verhaltensstrom" zuschreiben (einem *state*), die von der allgemeinen Einschätzung des Verhaltens (*personality*, also dem *trait*) abweicht. Eine Lehrkraft kann beispielsweise im Allgemeinen als sehr freundlich wahrgenommen werden, in einer bestimmten Situation aber auch von der Klasse übereinstimmend als unfreundlich eingeschätzt werden.
4. Analog zu *Stereotype* auf der Ebene der geteilten Wahrnehmungen steht *Residual* für den idiosynkratischen Teil der Wahrnehmung einer kategorialen Information.
5. *Opinion* bezieht sich auf den nicht-geteilten Anteil des wahrgenommenen Verhaltens der zu beurteilenden Person im Sinne eines *traits*.
6. *Error* steht für die nicht-geteilte Wahrnehmung eines bestimmten Verhaltensausschnitts, der von der allgemeinen Wahrnehmung des Verhaltens der zu beurteilenden Person (*opinion*) abweicht.

Die geteilte Wahrnehmung ist demnach zusammengesetzt aus den Komponenten *Stereotype*, *Personality* und *Norm*, wobei letztere nur dann als vollkommen geteilt betrachtet werden kann, wenn allen Beurteilern der gleiche Verhaltensausschnitt zu Verfügung steht. Dies ist im Allgemeinen im Kontext von Schülerbefragungen zum Unterricht der Fall. Ausnahmen wären beispielsweise Schüler, die erst im Laufe des Schuljahres zur Klasse hinzugekommen sind bzw. krankheitsbedingt eine Weile den Unterricht nicht besuchen konnten.

Kenny geht nun davon aus (S. 270f.), dass der Effekt der *Personality*-Komponente im Laufe der Zeit stetig zunimmt, während sich der Einfluss der *Stereotype*- und der *Norm*-Komponente verringern. Das heißt, nach einiger Zeit – Kenny nimmt an, dass dies bereits nach 90 beobachteten Verhaltenssequenzen eintritt – basiert die geteilte Wahrnehmung im Wesentlichen auf dem beobachteten Verhalten. Analog dazu steigt auf der Ebene der nicht-geteilten Wahrnehmungen die *Opinion*-Komponente mit der Zeit an, während die *Residual*- und die *Error*-Komponente abnehmen.

Da üblicherweise bei Befragungen von Schülerinnen und Schülern zum Unterricht die Lehrkraft die jeweilige Klasse schon längere Zeit unterrichtet (und nicht erst wenige Minuten oder Stunden), bedeutet dies, dass hier nur die zwei auf das überdauernde Verhalten der Lehrkraft bezogenen Komponenten (*personality* und *opinion*) eine Rolle spielen. Betrachtet man

die individuellen Urteile, so zeigt sich, dass die nicht-geteilte (subjektive) Wahrnehmung hier die dominierende Größe ist. Bei der Aggregation der Urteile der Schülerinnen und Schüler einer Klasse gewinnt mit zunehmender Schülerzahl die *Personality*-Komponente an Gewicht. Kenny (S. 275) weist allerdings darauf hin, dass der „objektivierende" Effekt der Aggregation geschmälert wird, wenn die Beurteiler über die zu beurteilende Person kommunizieren konnten – wenngleich sich die Übereinstimmung der Urteile dadurch erhöhen sollte –, wovon bei Schülerbefragungen zum Unterricht grundsätzlich ausgegangen werden muss.

Dieser „Verzerrungseffekt" der Kommunikation zwischen Beurteilern über die zu beurteilende Person wird darauf zurückgeführt, dass hier eine gemeinsame Sichtweise der Person konstruiert wird, die möglicherweise wenig mit der Realität zu tun hat. Die Ergebnisse aus empirischen Untersuchungen zum Effekt der Kommunikation sind widersprüchlich: Malloy, Agatstein, Yarlas und Albright (1997) fanden in ihren Studien eine Bestätigung für die Kommunikationshypothese, während Funder, Kolar und Blackman (1995) keinen Effekt der Kommunikation nachweisen konnten. Es ist denkbar, dass der Effekt der Kommunikation immer geringer wird, je besser vertraut die Beurteiler mit der zu beurteilenden Person sind, je größer also die Basis der Beurteilung, d.h. die Verhaltensstichprobe, ist.

Kenny (2004) weist darauf hin, dass auch die *opinion*-Komponente valide sein kann, nämlich dann, wenn es um die spezifische Interaktion des Beurteilers mit dem zu Beurteilenden geht. Verhält sich also eine Lehrkraft bezüglich eines zu beurteilenden Merkmals unterschiedlich zu verschiedenen Schülerinnen und Schülern, so wären auch die Abweichungen vom Klassenmittelwert in gewissem Umfang als valide zu betrachten. Dies gilt vor allem dann, wenn sich die jeweiligen Itemformulierungen auf individuelle Schüler als Adressat des Lehrerverhaltens beziehen (vgl. dazu Tabelle 1).

2.1.3 Modelle der kognitiv fundierten Survey-Forschung

Die kognitiv orientierte Survey-Forschung versucht, die an der Beantwortung einer (Survey-)Frage beteiligten kognitiven Prozesse zu bestimmen, um unbeabsichtigte Nebeneffekte bei der Befragung von Probanden zu minimieren. Bezogen auf die kognitiven Aufgaben bei der Beantwortung von Survey-Fragen skizzieren Sudman, Bradburn und Schwarz (1996, S. 56ff.) folgenden Prozess:

Zu Beginn erfolgt eine *Interpretation der Frage*, um deren Bedeutung zu erfassen. Dabei spielt zunächst die *Semantik* eine entscheidende Rolle. Sind Wörter mehrdeutig oder unbekannt, so werden diese vom jeweiligen Befragten (idiosynkratisch) gedeutet. Daneben ist der Bedeutungshorizont von Wörtern oft stark kontextabhängig, d.h. durch die Schaffung eines

bestimmten Kontextes, z.B. durch einen (kurzen) einleitenden Text, lässt sich die Zahl der möglichen Deutungen einschränken.

Neben der Semantik muss auch die *pragmatische* – also die intendierte – *Bedeutung* der Frage erschlossen werden. Die leitende Frage in diesem Prozess lautet: „Was möchte der Untersucher von mir wissen?“ Sind Antwortalternativen vorgegeben, so werden aus diesen Rückschlüsse auf die Bedeutung der Frage gezogen. Bei der Frage „Wie häufig sind Sie wirklich verärgert?“ z.B. würde „wirklich verärgert“ vermutlich sehr unterschiedlich interpretiert werden, abhängig davon, ob die Antwortalternativen im Bereich „nie“ bis „ein paar Mal pro Jahr“ oder „weniger als ein Mal pro Woche“ bis „ein paar Mal pro Woche“ liegen. Ebenso spielen die vorausgehenden und nachfolgenden Fragen eine Rolle im Sinne eines Kontextes. Hier ist v.a. zu beachten, dass eine Tendenz besteht, Neues zu berichten. Das kann dazu führen, dass bei der Beantwortung von Fragen bereits vorher Erfragtes ausgeschlossen wird. Folgt beispielsweise der Frage „Wie zufrieden sind Sie mit Ihrer Arbeit?“ das Item „Wie zufrieden sind Sie mit Ihrer Lebenssituation?“, dann wird bei der Beantwortung des letztgenannten Items der Bereich „Arbeit“ vermutlich ausgeblendet.

Bei Einstellungsfragen wird im Anschluss an die Interpretation der Frage überprüft, ob dazu bereits eine „Meinung“ vorliegt, auf die zurückgegriffen werden kann. Ist dies nicht der Fall, so muss zunächst eine mentale Repräsentation, basierend auf abgerufenen Informationen aus dem Gedächtnis, gebildet werden. Häufig muss anschließend ein Standard im Sinne eines Vergleichsmaßstabs entweder aus dem Gedächtnis abgerufen oder ebenfalls konstruiert werden.

Ähnliches gilt auch bei verhaltensbezogenen Fragen: Nach der inhaltlichen Deutung der Frage (auf welche Verhaltensweisen bezieht sich die Frage?) folgt der Abruf von Informationen aus dem Gedächtnis. Wird ein bestimmter Zeitabschnitt in der Frage vorgegeben, dann muss entschieden werden, ob die jeweiligen Fälle in das Zeitintervall fallen. Bezieht sich die Frage hingegen auf das „übliche“ Verhalten, dann muss entschieden werden, ob das erinnerte Verhalten repräsentativ ist. Können sich die Befragten nicht ausreichend gut erinnern oder sind sie wenig motiviert, so besteht die Gefahr, dass die Frage aufgrund allgemeiner Vorstellungen über die Häufigkeit solchen Verhaltens beantwortet wird (vgl. dazu auch Tourangeau, Rips und Rasinski (2000, S. 173f.) sowie den Abschnitt zur *systematic distortion hypothesis* in Kap. 2.1.4.3). Eine weitere „Fehlerquelle“ stellt der Rückgriff auf temporär verfügbar gemachte Informationen durch eine vorangegangene Frage dar, wenn sich die aktuelle Frage auf einen ähnlichen Inhalt bezieht.

Bei der Erfassung von Verhalten oder Einstellungen durch Fremdeinschätzungen (*proxy reporting*) sind weitere Besonderheiten zu berücksichtigen (Sudman et al., 1996, S. 227ff.). Verglichen mit selbstbezogenen Informationen werden Informationen über andere beim Encodieren weniger stark elaboriert, was die Abrufbarkeit aus dem Gedächtnis erschwert. Fremdauskünfte basieren entsprechend auf einer weniger breiten Basis an (aus dem Gedächtnis abrufbaren) Informationen. Ein Vorteil bei der Erfassung von Merkmalen mithilfe von Fremdeinschätzungen ist in geringeren Verzerrungstendenzen bezüglich sozialer Erwünschtheit und Selbstdarstellung zu sehen.

Bei der Beurteilung der *Einstellungen* anderer Personen spielen drei Informationsquellen eine wichtige Rolle (vgl. dazu auch Hoch, 1987):

1. die eigene Einstellung des Beurteilers zum jeweiligen Gegenstand
2. die vom Beurteiler wahrgenommene Ähnlichkeit oder Unähnlichkeit mit der anderen Person
3. andere relevante Informationen z.B. aus Gesprächen bzw. beobachtetem Verhalten

Auf der Grundlage der eigenen Einstellungen entwirft der Beurteiler eine Einschätzung, die entsprechend der wahrgenommenen Ähnlichkeit mit der jeweiligen Person mehr oder weniger stark gewichtet wird. Zuletzt wird dieses Urteil durch Informationen aus Gesprächen etc. ergänzt. Häufig wird auch versucht, Einstellungen bzw. Verhalten aus eingeschätzten Persönlichkeitsmerkmalen (*traits*) abzuleiten. Es ist naheliegend, dass in diesem Beurteilungsprozess von großer Bedeutung ist, wie gut sich der Beurteiler und die jeweilige Person kennen.

Bezogen auf die Unterrichtswahrnehmungen von Schülerinnen und Schülern aus Sicht der Klasse (vgl. dazu das Konzept „Wahrnehmungsperspektive“ in Kap. 2.1.1) ergeben sich aus den oben genannten Überlegungen folgende Besonderheiten:

1. Das Urteil entspricht sowohl einer Selbst- als auch einer Fremdauskunft, da Schülerinnen und Schülern hier sowohl nach ihrer eigenen Einschätzung als auch nach der Einschätzung ihrer Mitschüler gefragt werden. Es ist möglich, dass dadurch die Selbsteinschätzung – die ohnehin auch eine wichtige Rolle bei der Einschätzung der Einstellungen anderer Personen spielt – besonders akzentuiert wird.
2. Da sich die Fremdauskunft hier auf mehrere Personen (die Mitschüler) bezieht, muss zunächst ein Modell eines „repräsentativen Mitschülers“ entworfen werden. Aus der

wahrgenommenen Ähnlichkeit mit diesem Modell wird dann eine entsprechende Gewichtung bezüglich der eigenen Position vorgenommen.

3. Der Einbezug weiterer relevanter Informationen aus Gesprächen etc. hängt nun einerseits von deren Verfügbarkeit (von welchen Mitschülern liegen relevante Informationen bezüglich des zu beurteilenden Unterrichtsmerkmals in welchem Umfang vor?) und der entsprechenden Aggregation dieser Daten (im Sinne des „repräsentativen Mitschülers“) ab.

Die Qualität der Einschätzung der Einstellung der Mitschüler hängt also von der Qualität der wahrgenommenen Ähnlichkeit mit einem „repräsentativen Mitschüler“ und der Qualität und Quantität der relevanten Informationen ab.

2.1.4 Reliabilität und Validität von Unterrichtsperzeptionen

Eine sinnvolle Interpretation von auf empirischen Daten basierenden Ergebnissen setzt den Nachweis der Reliabiltät, der Validität sowie der Objektivität der jeweils eingesetzten Untersuchungsinstrumente voraus. Reliabilität bezieht sich auf die Messgenauigkeit, mit der ein Merkmal erfasst wird. Validität hingegen bezieht sich auf die Gültigkeit der Messung (wird wirklich das gemessen, was gemessen werden soll?). Objektivität ist dann gegeben, wenn die Messung als unabhängig vom jeweiligen Beobachter angesehen werden kann. Bei sogenannten *paper-and-pencil*-Verfahren (z.B. Fragebögen) ist das letztgenannte Kriterium im Allgemeinen erfüllt, v.a. wenn es sich um *multiple choice*-Antwortformate handelt (die oft maschinell eingelesen und ausgewertet werden). Mit dem Antwortformat von sogenannten Likert-Skalen sind allerdings bestimmte Validitäts- und Reliabilitätsproblematiken verbunden (vgl. dazu Kap. 2.1.4.1).

Während bezüglich des statistischen Nachweises der Reliabiltät bzw. Validität eines Instruments im Falle nicht-hierarchisch organisierter Daten verschiedene Standardverfahren existieren (z.B. Cronbachs α als interner Konsistenz-Index, Korrelation mit Außenkriterien), sind im Falle hierarchischer Datenstrukturen, bei denen Konstrukte auf Individual- und Aggregatebene betrachtet werden sollen, solche Standards noch nicht etabliert. Eine besondere Schwierigkeit in Bezug auf die Validität stellen hier häufig die unterschiedlichen Bedeutungen der Konstrukte auf den verschiedenen Ebenen dar (vgl. dazu Kap. 2.1.4.2). Betrachtet man die Reliabilität von solchen Konstrukten, so lässt sich neben den Reliabilitäten der Skalen auf beiden Ebenen (in Strukturgleichungsmodellen spricht man hier von Messmodellen und entsprechenden Indikator-Reliabilitäten) auch die (relative) Übereinstimmung von Indi-

viduen innerhalb von Gruppen als Reliabilität überprüfen. Ob eine solche Überprüfung überhaupt sinnvoll ist, hängt von theoretischen Annahmen über das zu erfassende Konstrukt auf der Aggregatebene ab.

Die Problematik der Übereinstimmungsreliabilität soll im Folgenden anhand von zwei Beispielen verdeutlicht werden. Das erste Beispiel bezieht sich auf die interne Konsistenz von Skalen. Besteht eine Skala aus Items, die alle sehr niedrige Trennschärfen aufweisen, dann ist diese Skala – selbst wenn sie aufgrund einer großen Zahl an Items eine hohe Reliabilität aufweist – inhaltlich schwer interpretierbar, da nicht davon ausgegangen werden kann, dass diese Skala auch wirklich das intendierte Konstrukt erfasst, also valide ist. Analog zu der Aggregation verschiedener Items bei einer Skala kann bei der Aggregation verschiedener Urteile zu einem bestimmten Item (z.B. Schülerurteile in Klassen zu einem bestimmten Unterrichtsmerkmal) eine niedrige gemeinsame Varianz (d.h. eine niedrige relative Übereinstimmung zwischen Beurteilern, gemessen z.B. in Form einer geringen Intraklassenkorrelation) die Interpretation der Variable auf Aggregatebene im Hinblick auf die Validität erschweren.

Dies lässt sich auch an einem zweiten Beispiel verdeutlichen. Bei der Auswertung von Unterrichtsvideografien werden häufig sogenannte hoch-inferente Ratings eingesetzt. Dazu werden Beurteiler auf der Basis gut ausgearbeiteter Ratingbögen, oft mit umfangreichen Beschreibungen der zu erfassenden Konstrukte und Hinweisen auf Besonderheiten, geschult. Im Anschluss an die Schulung erfolgen meist mehrere statistische Überprüfungen bezüglich der Übereinstimmung der jeweiligen Urteile verschiedener Beurteiler, wobei bestimmte Mindestkriterien erfüllt werden müssen. Oft sind eine oder mehrere „Nachschulungen“ erforderlich, um hinreichende Übereinstimmungen zu erzielen. Vergleicht man diesen Prozess mit der Erfassung von Unterricht aus Schülersicht anhand von Fragebögen, so wird deutlich, dass dort mit wesentlich niedrigeren Übereinstimmungen gerechnet werden muss. Selbstverständlich sind die hohen Übereinstimmungen von Beurteilern auch deshalb erforderlich, weil im Allgemeinen nur ein einzelner Rater eine bestimmte Lehrkraft beurteilt – und nicht wie bei Schülerbefragungen eine gesamte Klasse. Aber selbst bei einer großen Anzahl von Urteilen pro Lehrkraft wäre die Validität des Aggregats von Unterrichts-Ratings bei einer zu geringen Übereinstimmung zweifelhaft, da möglicherweise zum Teil nicht beabsichtigte „Störfaktoren“ hier eine gewisse Rolle spielen könnten, wie z.B. die Qualität des Filmmaterials, die äußere Erscheinung der Lehrkraft etc.

2.1.4.1 Antwortformat

Was das Antwortformat von Items betrifft, so besteht hier mittlerweile der Konsens, dass die Zahl der Antwortalternativen nicht zu klein sein sollte (mindestens fünf Kategorien), da

sonst möglicherweise mit einer geringeren Reliabilität (interne Konsistenz sowie Retest-Reliabilität) bzw. Validität der zu Skalen zusammengefassten Items zu rechnen ist (vgl. dazu Weng, 2004; Preston & Colman, 2000). Dabei spielen allerdings auch die kognitiven Voraussetzungen der Befragten eine Rolle. Weng (2004) schlägt als Optimum sechs oder sieben Kategorien vor, wenn sich die Befragung – wie in seiner Studie – auf College-Studenten bezieht. Bei der Wahl der Antwortalternativen sollten auch die Bearbeitungsschwierigkeit, die bei einer geringeren Zahl an Antwortkategorien niedriger eingeschätzt wird, sowie die Möglichkeit, die individuelle Einstellung aufgrund einer ausreichenden Zahl an Alternativen adäquat ausdrücken zu können, berücksichtigt werden (vgl. Preston & Colman, 2000). Ob jede einzelne Antwortalternative verbal verankert wird oder nur die Pole, ist offenbar ohne Bedeutung für die Reliabiltität der aus solchen Items resultierenden Skalen. Weng (2004) schlägt jedoch aus Gründen der Interpretierbarkeit von Item- bzw. Skalenmittelwerten[9] vor, alle Kategorien zu benennen.

Auf eine interessante Variante bezüglich der Vorgabe von Antwortalternativen weist auch Barnette (2000) hin: Wenn mit möglichen Verzerrungen bei der Beantwortung der Fragen gerechnet werden muss – z.B. aufgrund von Unaufmerksamkeit bei der Bearbeitung –, dann lässt sich dem durch eine gelegentliche Umkehr der Reihenfolge der Antwortkategorien entgegenwirken. So können beispielsweise jeweils die Hälfte der Items einer Skala (alle gleichgerichtet formuliert) mit unterschiedlichen Antwort-Reihenfolgen erfasst werden. Dieses Verfahren ist Barnette (ebd.) zufolge dem gemischten Einsatz von positiv und negativ formulierten Items bei konstanter Reihenfolge der Antwortalternativen, was häufig zu Problemen bezüglich der internen Konsistenz bzw. der Faktorstruktur etc. führt, deutlich vorzuziehen.

Bezüglich der Zahl der Antwortalternativen ist auch die Reliabilität der zu erfassenden Skala von großer Bedeutung: Skalen mit niedriger Reliabilität „profitieren“ von einer höheren Zahl an Antwortkategorien wesentlich deutlicher als Skalen mit hoher Reliabilität (vgl. Weng, 2004).

[9] Da es sich hierbei um ordinale Daten handelt, ist die Bildung von Mittelwerten streng genommen ohnehin fraglich.

2.1.4.2 Kompositionsmodelle: Theoretische Modelle zur Aggregation von Daten

Die Notwendigkeit der Überprüfung von klasseninternen Übereinstimmungen bei der Aggregation von Individualdaten lässt sich anhand der von Chan (1998) vorgeschlagenen Unterscheidung fünf theoretischer Kompositionsmodelle verdeutlichen:

1. Im *additiven Modell* (*additive model*) ist die Streuung innerhalb der Gruppen ohne theoretische Bedeutung. Sie stellt lediglich ein messtheoretisches Problem im Sinne zufälliger Messfehler und Verzerrungstendenzen dar.
2. Im *Modell unmittelbarer Übereinstimmung* (*direct consensus model*) hingegen ist die Übereinstimmung der Urteile innerhalb der Gruppen theoretisch bedeutsam, da die Bedeutung des Merkmals auf Aggregatebene (also z.B. auf Klassenebene) einen funktionalen Zusammenhang mit dem Merkmal auf Individualebene aufweist: Das Aggregat steht für die geteilte, die Abweichungen innerhalb der Gruppe für die nicht-geteilte Wahrnehmung „isomorpher" Konstrukte. Neben der Übereinstimmung innerhalb der Gruppen spielt auch die Variabilität des Merkmals zwischen Gruppen eine wesentliche Rolle, da mangelnde Streuungen auf der Aggregatebene die Validität des dort hypothetisch formulierten Konstrukts in Frage stellen.
3. Das *Modell der Übereinstimmung mit Perspektivenverschiebung* (*referent-shift consensus model*) unterscheidet sich vom Modell unmittelbarer Übereinstimmung dadurch, dass Individuen (z.B. aufgrund von Itemformulierungen wie etwa „In unserer Klasse…") die Sicht von Gruppenmitgliedern übernehmen. Hier ist zu beachten, dass sich auch die Konstrukte auf der Aggregatebene bezüglich der beiden Modelle theoretisch unterscheiden. Chan differenziert zwischen *organizational collective climate* (*referent-shift consensus model*) und *organizational climate* (*direct consensus model*).
4. Ein weiteres Kompositionsmodell, das sogenannte *Dispersionsmodell* (*dispersion model*), liegt vor, wenn die gruppeninternen Streuungen als Konstrukt auf Aggregatebene betrachtet werden sollen. Da es sich hier um ein Konstrukt handelt, das nur auf der Aggregatebene existiert – und somit keine theoretischen Zusammenhänge zu dem jeweiligen Merkmal auf der Individualebene existieren –, ist dessen theoretische Beschreibung und Einbettung in ein nomologisches Netzwerk unabdingbar.
5. Während sich die zuvor genannten Modelle auf den Status bezüglich eines Merkmals beziehen, geht es im *Prozess-Modell* (*process model*) um Entwicklungsverläufe (wenn etwa die Entwicklung der Streuungen innerhalb von Klassen bezüglich des Unterrichtsklimas im Laufe der Zeit untersucht werden soll).

Bei der Aggregation von Schülerwahrnehmungen des Unterrichtsklimas handelt es sich in der Regel um das zweite bzw. das dritte Kompositionsmodell (*direct consensus* bzw. *referent-shift model*). Dementsprechend ist die Überprüfung der klasseninternen Übereinstimmungen hier zwingend erforderlich. Ob dies jedoch auf der Ebene einzelner Klassen erfolgen sollte, ist allerdings eine bisher weitgehend ungeklärte Frage (vgl. dazu auch Lüdtke, Trautwein, Kunter & Baumert, 2006). Hier müsste wohl eher *theoretisch* begründet werden, warum bestimmte Klassen von Analysen ausgeschlossen werden (z.B. weil kurz vor der Erhebung ein Lehrerwechsel stattgefunden hat und der Unterricht der neuen Lehrkraft deshalb nur unzureichend beurteilt werden kann).

Hohe Heterogenität bezüglich der individuellen Unterrichtswahrnehmungen in Klassen kann auch ein Zeichen dafür sein, dass es relevante „Subgruppen" (z.B. Cliquen) gibt (vgl. Chan, 1998; Lüdtke et al., 2006; Satow, 1999). Es ist denkbar, dass sich die Wahrnehmungen in bestimmten Schülergruppen deshalb stärker ähneln, weil innerhalb solcher Gruppen in wesentlich höherem Ausmaß kommuniziert wird als zwischen Schülerinnen und Schülern aus unterschiedlichen Gruppen. Satow (1999) weist darauf hin, dass aufgrund unterschiedlicher sozialer Strukturen innerhalb von Schulklassen grundsätzlich nicht davon auszugehen ist, dass das Klima in Klassen von allen Schülerinnen und Schülern übereinstimmend erlebt wird. In einer Studie von Lubbers (2003) zeigte sich beispielsweise, dass die sozialen Netzwerke von Mädchen und Jungen (Altersdurchschnitt: 13 Jahre) innerhalb von Klassen fast vollständig getrennt sind. Die vielfach empirisch bestätigte These, wonach in der Kindheit und Jugend das Geschlecht bei der Bildung von Dyaden ein entscheidendes Selektionskriterium darstellt (ebd.), lässt sich also auch auf die Bildung sozialer Netzwerke im Klassenkontext übertragen.

Zu berücksichtigen ist auch die eventuelle Abhängigkeit individueller Unterrichtswahrnehmungen von Personmerkmalen. Bei der Wahrnehmung des Unterrichtstempos etwa spielt vermutlich die Fachkompetenz eine wichtige Rolle: Leistungsschwache Schülerinnen und Schüler sollten das Unterrichtstempo höher einschätzen als leistungsstarke. Das hieße, dass in Klassen mit großer Leistungsheterogenität auch mit einer größeren Streuung bezüglich der individuellen Wahrnehmung des Unterrichtstempos zu rechnen ist, was aber für die Erfassung des Konstrukts auf Klassenebene ohne Bedeutung ist.

Der Einbezug der klasseninternen Streuungen des wahrgenommenen Klimas – wie bei Satow (1999) – entspricht dem Dispersions-Modell. Hier muss theoretisch begründet werden, welche Bedeutung dem – im Falle verschiedener Klimamerkmale: *jeweiligen* – Konstrukt auf Klassenebene zukommt.

In ähnlicher Weise wie Chan (1998) unterscheidet auch Bliese (2000) verschiedene theoretische Modelle bei der Aggregation von Individualdaten. Die beiden Pole eines Kontinuums von Aggregationsmodellen stellen dabei folgende Modelle dar:

1. Das *Umwandlungsprozess-Modell* (*compilation process model*) bezieht sich auf die Aggregation von Individualdaten, bei der das jeweilige Merkmal auf der Individual- und der Aggregatebene theoretisch nicht verwandt sind. Solche Modelle beziehen sich oft auf Streuungen innerhalb von Gruppen oder Zusammensetzungen von Gruppen (wie z.B. die Geschlechtszusammensetzung von Klassen; hier ist offensichtlich, dass es sich um völlig verschiedene Variablen auf Individual- und Klassenebene handelt).
2. Bei *Kompositionsprozess-Modellen* (*composition process models*) wird angenommen, dass die auf Individualebene erfassten Phänomene jeweils innerhalb der Gruppen *und gleichzeitig* mit dem Merkmal auf Aggregatebene isomorph sind. Im Idealfall würde dies bedeuten, dass es keine Streuungen innerhalb von Gruppen gibt.

In den meisten Fällen, so Bliese, handelt es sich bei der Aggregation von Individualmerkmalen um einen Prozess, der zwischen diesen beiden Polen liegt. Bliese spricht in diesem Zusammenhang von *fuzzy composition process models*. Hier sind die Merkmale auf Individual- und Aggregatebene theoretisch mehr oder weniger eng miteinander verbunden, also weder vollkommen isomorph, noch völlig unabhängig voneinander. In solchen Fällen besteht die Möglichkeit, dass das Konstrukt auf der Aggregatebene zusätzlich Kontexteffekte beinhaltet, die das individuelle Merkmal nicht aufweist. Als Beispiel nennt Bliese hier die durchschnittliche Absenz am Arbeitsplatz, die theoretisch als „Absenz-Kultur“ bezeichnet werden kann.

Werden auf Individualebene Merkmale erfasst, die einen direkten Bezug zur Aggregatebene aufweisen – etwa die Unterrichtswahrnehmung aus Sicht der Klasse –, so ist die Annahme berechtigt, dass die Variable auf Aggregatebene einer zuverlässigeren Messung des (objektiven) Merkmals entspricht als das entsprechende (subjektiv eingefärbte) Individualmerkmal. Bezieht sich das auf Individualebene erfasste Merkmal hingegen nicht direkt auf das Aggregatmerkmal, so ist diese Annahme nicht ohne weiteres gerechtfertigt. Hier besteht zwar eine nomologische Beziehung zwischen den Merkmalen auf beiden Ebenen, aber zusätzlich können hier kontextuelle Faktoren bei der Aggregatvariable eine Rolle spielen: Bei der Aggregation der auf individuelle Merkmale bezogenen Einschätzungen werden möglicherweise Effekte des Kontextes sichtbar, die die jeweiligen Gruppen in ihrer Gesamtheit betreffen (wie z.B. die jeweilige materielle Ausstattung einer Schule), die über das individuelle Merkmal hinausgehen (vgl. dazu das Beispiel zur „Absenz-Kultur“ oben).

Eine Überprüfung der Übereinstimmungen innerhalb der Gruppen ist zwar bei *fuzzy composition*-Modellen Bliese zufolge nicht grundsätzlich erforderlich. Im Falle der Erfassung von Merkmalen, die sich auf die geteilten Wahrnehmungen beziehen, sollte ein solcher Nachweis allerdings erbracht werden, da eine gewisse Übereinstimmung theoretisch erwartet wird.

2.1.4.3 Halo-Effekte

Bezüglich der Validität von Unterrichtwahrnehmungen aus Schülersicht sind sogenannte Halo-Effekte (auch *correlational bias* genannt) von großer Bedeutung (vgl. z.B. Clausen, 2002). Halo-Effekte resultieren aus der mangelnden Fähigkeit von Beurteilern, verschiedene Aspekte einer zu beurteilenden Person klar zu trennen. Im Allgemeinen werden drei verschiedene Erklärungsmodelle für Halo-Effekte unterschieden (vgl. Lance, LaPointe & Steward, 1994; Feeley, 2002):

1. Im *general impression model* wird angenommen, dass ein globaler Eindruck der zu beurteilenden Person die einzelnen zu beurteilenden Konstrukte „kontaminiert". Die Tatsache beispielsweise, dass ein bestimmter Schüler seine Lehrkraft besonders mag (aus welchen Gründen auch immer), sollte demnach die Urteile dieses Schülers bezüglich verschiedener Unterrichtsmerkmale in gleicher Weise beeinflussen.
2. Im *salient dimension model* hingegen wird die Beeinflussung der Urteile aufgrund einer oder mehrerer salienter Dimensionen postuliert. So könnte etwa die physische Attraktivität (als eine saliente Dimension) einer zu beurteilenden Person einen Einfluss auf andere zu beurteilende Dimensionen ausüben.
3. Das *inadequate discrimination model* postuliert, dass Beurteiler nicht in der Lage sind, die verschiedenen zu erfassenden Dimensionen adäquat zu unterscheiden. Dieser Halo-Effekt sollte insbesondere bei abstrakt formulierten Items bzw. bei geringer Anstrengung während der Bearbeitung der Items auftreten.

Lance et al. (1994) konnten nachweisen, dass das *general impression model* unabhängig von verschiedenen Kontextbedingungen, die u.a. das Auftreten von Effekten im Sinne der beiden anderen Erklärungsmodelle begünstigen sollten, die Daten am besten erklärte. Der Halo-Effekt sollte, so wird häufig angenommen (vgl. z.B. Feeley, 2002; Kenny & Bergman, 1980), bei der Aggregation von Daten mehrerer Beurteiler allerdings abgemildert werden. Kenny und Bergman (1980) argumentieren, dass zwar innerhalb von Beurteilern die auf Halo-Effekten basierenden Fehlerterme korreliert sind. Zwischen verschiedenen Beurteilern seien diese Fehler jedoch unkorreliert.

Auf die Unterrichtswahrnehmungen von Schülerinnen und Schülern bezogen hieße das, dass die einzelnen Unterrichtsmerkmale zwar innerhalb von Klassen aufgrund des Halo-Effektes hoch miteinander korrelieren sollten, nicht jedoch auf der Klassenebene. Dies steht jedoch – zumindest auf den ersten Blick – im Widerspruch zu empirischen Ergebnissen: Dort zeigt sich meist auf der Klassenebene eine einfachere Faktorstruktur als innerhalb von Klassen (vgl. z.B. Baumert, Kunter, Brunner, Krauss, Blum & Neubrand, 2004; Helmke, A., Schrader, Wagner, Klieme, Nold & Schröder, in Druck-a; Klieme, Jude, Eichler & Willenberg, in Druck), d.h. die geteilten Wahrnehmungen sind höher korreliert als die „subjektiven" Anteile dieser Merkmale. Für diese hohen Zusammenhänge auf Klassenebene gibt es mindestens zwei konkurrierende Erklärungen (vgl. dazu auch Marsh, 1982): (1) Es ist denkbar, dass diese hohen Zusammenhänge „realistisch" sind, dass also z.B. Lehrkräfte mehr oder weniger „gut" unterrichten – und zwar bezüglich verschiedener (theoretisch) unterscheidbarer Unterrichtsdimensionen gleichzeitig. (2) Es handelt sich um einen systematischen *bias* im Sinne eines Halo-Effektes. Geht man etwa entsprechend dem *general impression*-Modell davon aus, dass in die Unterrichtswahrnehmungen von Schülerinnen und Schülern jeweils ein Gesamteindruck der Lehrkraft (z.B. der Schüler mag seine Lehrkraft) eingeht, dann ist es möglich – wenn nicht gar wahrscheinlich –, dass auch diesbezüglich eine gewisse Übereinstimmung zwischen Schülerinnen und Schülern einer Klasse herrscht. Eine solche Übereinstimmung würde aber bedeuten, dass dieser Effekt bei der Aggregation der Daten *nicht* beseitigt wird.

Bei der Aggregation von Schülerdaten zur Unterrichtswahrnehmung *ohne* Berücksichtigung von Kontrollvariablen wird – zumindest implizit – davon ausgegangen, dass Schülerinnen und Schüler einer Klasse bezüglich ihrer Beurteilungen jeweils repräsentativ bezüglich der Gesamtheit der Schülerinnen und Schüler sind. Bei einer solchen Aggregation erhöht sich zwar die Reliabilität, nicht jedoch zwingend die Validität, wenn ein konstanter *bias* vorliegt (vgl. Haladyna & Hess, 1994; Dreesmann, 1979, S. 72; Babad, 1996). Haladyna und Hess (ebd.) konnten bei Studenten u.a. einen Effekt des Geschlechts nachweisen: Studentinnen beurteilten den Unterricht deutlich „milder", vergaben also höhere Beurteilungen, als ihre männlichen Kommilitonen. Bei der einfachen Aggregation der Daten bliebe dieser *bias* zumindest teilweise erhalten, sofern sich die Klassenzusammensetzungen bezüglich des Geschlechts unterscheiden. Die Forschungsergebnisse zum Effekt des Geschlechts sind Aleamoni (1999) zufolge – zumindest bezogen auf studentische Beurteiler – eher uneinheitlich. Bei Schülerinnen und Schülern wurden u.a. Effekte des Geschlechts und der Benotung bzw. der auf Lehrerangaben basierenden Leistung (Gentry, Gable & Rizza, 2002; Babad, 1996) bezüglich der Unterrichtswahrnehmung gefunden.

Was die Benotung betrifft, so gibt es einen breiten Konsens bezüglich der sogenannten *grading leniency* und der Beurteilung der Instruktionsqualität (vgl. Aleamoni, 1999; Greenwald, 1997). Marsh und Roche (1997, 2000) weisen jedoch darauf hin, dass bei der Beurteilung des Zusammenhangs von Noten und Unterrichtsbeurteilungen mindestens drei konkurrierende Hypothesen in Frage kommen:

1. Im Rahmen der *Validitätshypothese* wird davon ausgegangen, dass der positive Zusammenhang zwischen Noten (Polung: höhere Note entspricht besserer Note) und Unterrichtsbeurteilungen als Hinweis auf die Validität des Instruments verstanden werden kann, dass also „besserer" Unterricht zu besseren Leistungen (also besseren Noten) führt.
2. Eine alternative Hypothese (die sogenannte *prior characteristics hypothesis*) besteht darin, dass verschiedene vorausgehende Person- oder Situationsmerkmale (z.B. Fachinteresse oder Klassengröße) sowohl das Lernen – und somit die Noten – als auch die Unterrichtswahrnehmung beeinflussen.
3. Die *grading leniency*-Hypothese postuliert, dass der Unterricht von Lehrkräften, die bezogen auf die tatsächliche Leistungsfähigkeit der Klasse zu gute Noten vergeben, als besser eingeschätzt wird. Das heißt, nicht die Noten per se sind hier von Bedeutung, sondern die verglichen mit der jeweiligen Leistung unangemessene Benotung.

Betrachtet man die oben im Zusammenhang mit dem *general-impression*-Halo-Effekt aufgeführte generelle Einstellung zur Lehrkraft („Ich mag meine Lehrkraft gerne.") näher, so lässt sich auch ein (gleichzeitiger) Zusammenhang in umgekehrter Richtung formulieren. Es ist denkbar, dass die Einstellung zur Lehrkraft insbesondere aus verschiedenen Merkmalen des Unterrichts der jeweiligen Lehrkraft resultiert, dass also die Einstellung zur Lehrkraft wesentlich von deren Unterricht abhängt. Dies ist insofern plausibel, da Schülerinnen und Schüler ihren Lehrkräften im Allgemeinen überwiegend in Unterrichtssituationen begegnen, die Einstellung zur Lehrkraft also kaum unabhängig von ihrem Verhalten im Unterricht zustande kommen kann. Demnach ließe sich die Einstellung zur Lehrkraft auch – zumindest zum Teil – im Sinne eines Globalurteils bezüglich des Unterrichts interpretieren.

Um zu einem solchen Globalurteil zu gelangen, müssen die verschiedenen berücksichtigten Dimensionen gewichtet werden. Carkenord und Stephens (1994) konnten hierbei Diskrepanzen bezüglich der selbsteingeschätzten und der tatsächlich vorgenommenen Gewichtungen von Beurteilern nachweisen. Daneben unterscheiden sich verschiedene Beurteiler in ihren individuellen Gewichtungen der einzelnen Komponenten. Interessanterweise hängen diese Ge-

wichtungen nicht nur vom jeweiligen Beurteiler, sondern auch von der jeweils zu beurteilenden Person ab (McIntyre & James, 1995). Die Globalurteile bezüglich verschiedener Personen setzen sich also aus unterschiedlichen Anteilen einzelner Merkmale zusammen (und sind demnach nicht miteinander vergleichbar, da sich sozusagen das Messmodell unterscheidet).

Bezogen auf die Stabilität von Halo-Effekten kommen Murphy und Anhalt (1992) auf der Basis zweier Studien zu dem Schluss, dass Halo-Effekte wenig stabil sind, wenn entweder die zu beurteilende Person oder deren Verhalten (z.B. der Inhalt einer Unterrichtseinheit) verändert wird. Moderat stabile Halo-Effekte ergeben sich, wenn sowohl die zu beurteilende Person als auch das Verhaltenssegment konstant gehalten werden.

Eine andere Quelle systematischer Verzerrungen konnten Renaud und Murray (2005) im Rahmen einer Untersuchung identifizieren. Ausgangspunkt ihrer Überlegungen war die *systematic distortion hypothesis*, die eine Erklärung für künstlich überhöhte Zusammenhänge zwischen verschiedenen Konstrukten liefert. Urteile basieren dieser Theorie nach u.a. auf impliziten Persönlichkeitstheorien, also bestimmten Vorstellungen darüber, welche Verhaltensweisen im Allgemeinen mehr oder weniger häufig gemeinsam auftreten. Interessanterweise sind niedrig-inferente Ratings (wenngleich im Allgemeinen reliabler; vgl. dazu Kap. 2.1.1) von solchen Verzerrungstendenzen stärker betroffen, da diese stärker gedächtnisbasiert sind. Bezogen auf die Beurteilung des Unterrichts durch Schülerinnen und Schüler wäre es denkbar, dass sich die impliziten Persönlichkeitstheorien aufgrund unterschiedlicher Erfahrungen mit Lehrkräften im Laufe der schulischen Sozialisation (z.B. über verschiedene Bildungsgänge hinweg) unterscheiden.

2.2 Entwicklung der Fragestellung

In den vorangegangenen Kapiteln wurde gezeigt, dass bei der Unterrichtswahrnehmung aus Schülersicht zwei Ebenen unterschieden werden müssen, nämlich die Individual- und die Klassenebene. Die Individualebene steht dabei für die nicht-geteilte, die Klassenebene für die geteilte Wahrnehmung des Unterrichts. Kenny (2004) unterscheidet im Rahmen seines PERSON-Modells (vgl. Kap. 2.1.2) wiederum jeweils drei verschiedene Komponenten der Wahrnehmung auf beiden Ebenen. Für die Unterrichtswahrnehmung spielen aber aufgrund der hohen Anzahl an beobachteten Verhaltensakten nur die auf überdauerndes Verhalten bezogenen Komponenten eine Rolle, deren Gewicht im Laufe der Zeit immer größer wird. Dies setzt allerdings eine gewisse Konstanz des konkret zu erfassenden Merkmals voraus. Möglicherwei-

se gibt es hier Unterschiede zwischen verschiedenen Lehrkräften und Unterrichtsmerkmalen, auch in Abhängigkeit vom jeweiligen Unterrichtsstoff.

Sowohl Kenny als auch Dreesmann (1979, 1982; vgl. Kap. 2.1.1) gehen davon aus, dass die Urteilsbildung durch Kommunikation beeinflusst wird. Dabei sollte der Einfluss der Kommunikation umso stärker ausfallen, je *unsicherer* sich die jeweilige Schülerin bzw. der jeweilige Schüler in ihrem bzw. seinem eigenen Urteil ist. Kenny (2004) sieht in der Kommunikation auch entsprechend eine Quelle für Verzerrungen, die dem objektivierenden Einfluss der Aggregation von Individualdaten entgegenwirkt, da sich in diesem Fall auf Aggregatebene wahre Werte und Meinungen vermischen. Hier wird bereits deutlich, dass sich die Übereinstimmung zwischen Beurteilern auf die *Reliabilität* eines Konstrukts auf Aggregatebene bezieht. Bei einer sehr niedrigen Übereinstimmung wird jedoch u.U. – je nach zugrunde gelegtem Kompositionsmodell (vgl. Kap. 2.1.4.2) – auch die Validität von Konstrukten in Zweifel gezogen.

Unsicherheiten im Urteilsprozess können entsprechend der Befunde der kognitiv fundierten Survey-Forschung (vgl. Kap. 2.1.3) beispielsweise dann auftreten, wenn sich Beurteiler nicht ausreichend an – bezogen auf die jeweilige Frage – relevante Informationen erinnern können, oder wenn sie zur Beantwortung der Frage(n) wenig motiviert sind. In solchen Situationen steigt die Gefahr, dass auf allgemeine Vorstellungen z.B. über die Auftretenshäufigkeit bestimmter Verhaltensweisen zurückgegriffen wird. Dies ist insbesondere bei den stärker gedächtnisbasierten, niedrig-inferenten Urteilen der Fall.

Bei der Befragung von Schülerinnen und Schülern zum Unterricht aus Sicht der Klasse sind insbesondere die Konstruktion eines „repräsentativen Mitschülers“, die wahrgenommene Ähnlichkeit mit diesem, sowie die Verfügbarkeit von relevanten Informationen von Mitschülern von Bedeutung. Da innerhalb von Klassen verschiedene soziale Netzwerke bestehen, ist davon auszugehen, dass sich unterschiedliche Beurteiler auf verschiedene Referenzen beziehen.

Ausgangspunkt bei der Einschätzung der Wahrnehmungen von Mitschülern ist zunächst das eigene Urteil, dass entsprechend der wahrgenommenen Ähnlichkeit mit den Mitschülern mehr oder weniger stark gewichtet wird. Dementsprechend ist es fraglich, ob die Konstrukte hier auf Individual- und auf Klassenebene – wie in Chans (1998) Kompositionsmodell – als vollkommen isomorph angesehen werden können (dass das Konstrukt auf Aggregatebene also lediglich eine reliablere Messung des Konstrukts auf Individualebene darstellt). Hier scheint Blieses (2000) *fuzzy composition process model*, das bezüglich der Konstrukte auf Individual-

und Aggregatebene nur eine Ähnlichkeit – aber keine vollkommene Isomorphie – annimmt, angebrachter zu sein.

Eine wichtige Rolle bei der Urteilsbildung spielen möglicherweise sogenannte Halo-Effekte (vgl. Kap. 2.1.4.3), die als Erklärungsansatz für überhöhte Zusammenhänge zwischen theoretisch unterscheidbaren Dimensionen herangezogen werden. Empirisch bewährt hat sich insbesondere das *general impression*-Modell, wonach der generelle Eindruck eines Beurteilers hinsichtlich einer zu beurteilenden Person (im Sinne eines Globalurteils) einen Einfluss auf die jeweiligen Urteile hinsichtlich verschiedener Konstrukte ausübt. Dieses Modell lässt sich vermutlich auch auf die Klassenebene übertragen, wenn man davon ausgeht, dass es eine übereinstimmende globale Wahrnehmung der Lehrkraft durch die Schülerinnen und Schüler einer Klasse gibt. Demnach wären auch die Zusammenhänge auf Klassenebene aufgrund eines globalen Eindrucks der Lehrkraft verzerrt.

Da Schülerinnen und Schüler ihre Lehrkräfte v.a. im Kontext von Unterrichtssituationen erleben, ist – umgekehrt – anzunehmen, dass der generelle Eindruck von Lehrkräften auch von deren Verhalten im Unterricht abhängt. Aufgrund der Personenabhängigkeit der Gewichtungen, mit denen unterschiedliche Merkmale in das Globalurteil einfließen, ist davon auszugehen, dass sich die Kompositionen der generellen Eindrücke von Lehrkraft zu Lehrkraft unterscheiden.

Im Sinne eines verzerrenden Einflusses auf die Unterrichtswahrnehmung werden neben dem generellen Eindruck der zu beurteilenden Person insbesondere das Geschlecht des Beurteilers sowie die Milde bzw. Strenge der Benotung durch die zu beurteilende Lehrkraft diskutiert. Mädchen beurteilen offensichtlich Unterricht tendenziell milder (also positiver) als Jungen. Was die Benotung anbelangt, so wird – im Rahmen der sogenannten *grading leniency*-Hypothese – angenommen, dass eine verglichen mit dem tatsächlichen Leistungsniveau unangemessene Benotung zu besseren Beurteilungen der Unterrichtsqualität führt.

In Abbildung 1 werden die zentralen Befunde nochmals grafisch veranschaulicht. Ausgangspunkt ist das (beobachtbare) Verhalten der Lehrkraft. Dieses wird – über verschiedene individuelle Vermittlungsprozesse (Einflussfaktoren wie Leistung, Motivation bei der Beantwortung der Fragen etc.) – aggregiert zu verschiedenen Unterrichtswahrnehmungen sowie zu einem auf die Lehrkraft bezogenen Globalurteil.

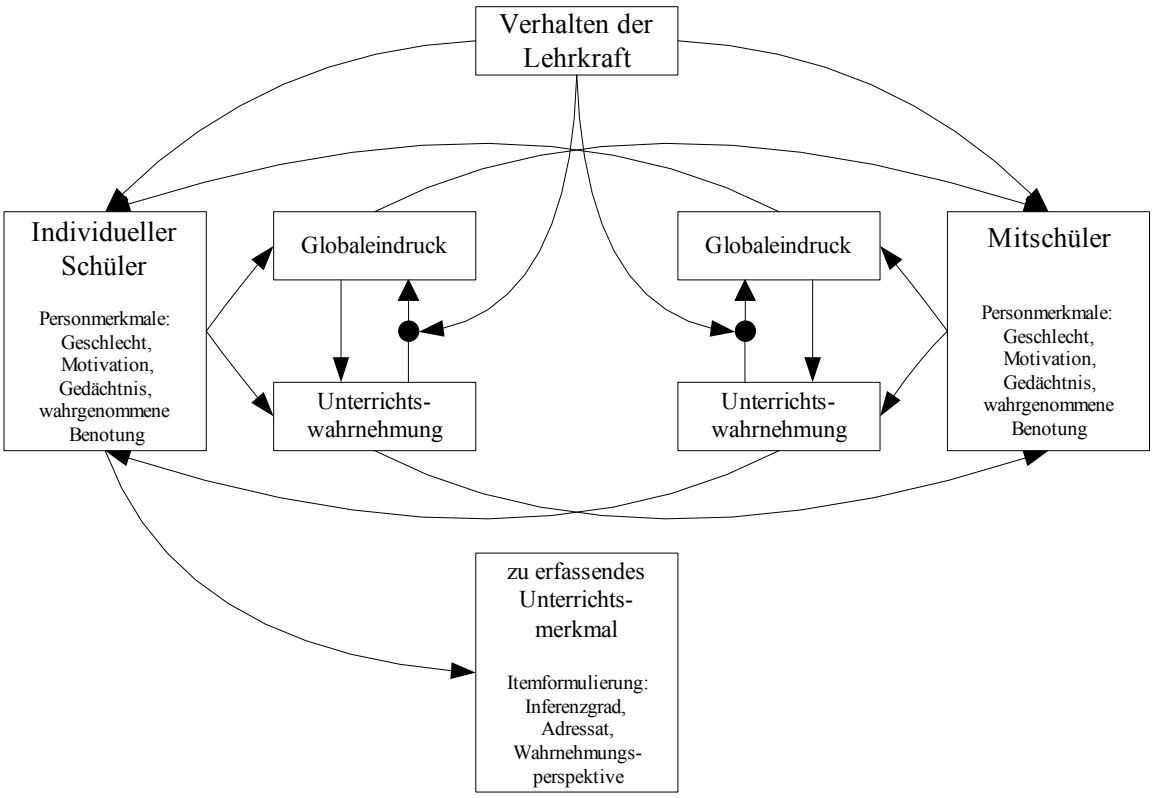

Abbildung 1: Einflussgrößen der individuellen Wahrnehmung von Unterricht

Die Wahrnehmung eines spezifischen Unterrichtsmerkmals (U^*) lässt sich Anderson zufolge (1996, S. 114) theoretisch in eine kontextfreie Komponente (U) – d.h. eine vom jeweiligen Globalurteil bezüglich der Lehrkraft unabhängigen – und eine Globalurteil-Komponente (G) zerlegen, die jeweils mit dem relativen Gewicht g bzw. 1-g in die Gleichung eingehen (vgl. (1)). In Abbildung 1 ist dies durch den vom Globaleindruck ausgehenden Pfeil auf die Unterrichtswahrnehmungen angedeutet. Umgekehrt beeinflussen die Unterrichtswahrnehmungen wiederum das Globalurteil bezüglich einer Lehrkraft, wobei das relative Gewicht hier abhängig von der jeweiligen zu beurteilenden Person variieren sollte (dafür steht der Pfeil von „Verhalten der Lehrkraft“ auf den Pfeil zwischen „Unterrichtswahrnehmung“ und „Globaleindruck“).

$$U^* = g \cdot U + (1-g) \cdot G \tag{1}$$

Darüber hinaus werden sowohl die auf die Lehrkraft bezogenen globalen Wahrnehmungen als auch die Unterrichtswahrnehmungen der Mitschüler in die jeweiligen individuellen Wahrnehmungen integriert. Dies wird in Abbildung 1 durch die Pfeile von „Globaleindruck“ bzw. „Unterrichtswahrnehmung“ der Mitschüler auf die individuelle Schülerin bzw. den indi-

viduellen Schüler verdeutlicht. Hierbei spielen die jeweils zur Verfügung stehenden relevanten Informationen von Mitschülern (es ist davon auszugehen, dass in der Regel die vorliegenden Informationen über die Mitschüler hinweg variieren) sowie die wahrgenommene Ähnlichkeit mit diesen eine bedeutende Rolle: Je mehr Informationen zur Verfügung stehen und je höher die wahrgenommene Ähnlichkeit mit dem jeweiligen Mitschüler, desto stärker sollte sich der Einfluss des jeweiligen Mitschülers auf die Urteilsbildung auswirken.

Bezogen auf die konkrete Erfassung eines Unterrichtsmerkmals aus Schülersicht ist zudem die Itemformulierung von Bedeutung. Die Beantwortung eines Items sollte von Inferenzgrad, Adressatenbezug und Wahrnehmungsperspektive abhängen.

2.3 Fragestellungen

Basierend auf den theoretischen Annahmen und empirischen Ergebnissen zur „Unterrichtswahrnehmung aus Schülersicht“ soll in der vorliegenden Untersuchung folgenden Forschungsfragen nachgegangen werden:

1. *Übereinstimmung*: Sind die Übereinstimmungen der Urteile von Schülerinnen und Schülern innerhalb von Klassen bezüglich verschiedener Unterrichtsmerkmale ausreichend hoch, um von „geteilter Unterrichtswahrnehmung“ sprechen zu können? Ausgehend von der Annahme, dass die sozialen Netzwerke von Jungen und Mädchen nahezu vollständig getrennt sind, sollten sich bei geschlechtsspezifischen Analysen (also getrennte Analysen für Jungen bzw. Mädchen) – aufgrund der stärkeren Kommunikation innerhalb dieser Gruppen – höhere Übereinstimmungen zeigen. Dieser Effekt sollte sich insbesondere bei Unterrichtsmerkmalen zeigen, die eine Perspektivenübernahme erfordern, da dort der Rückgriff auf Informationen von Mitschülern besonders bedeutsam ist.
2. *Differenziertheit*: Wie differenziert beurteilen Schülerinnen und Schüler den Unterricht? Lassen sich theoretische Konstrukte empirisch nachweisen und hinreichend unterscheiden (konvergente und diskriminante Validität)?
3. *Fachspezifität*: Unterscheidet sich die Unterrichtswahrnehmung aus Schülersicht in den Unterrichtsfächern Deutsch und Englisch? Es wird erwartet, dass bei Unterrichtsmerkmalen, die mithilfe niedrig-inferenter Items erfasst werden, keine Unterschiede bestehen sollten, da dort die Wahrnehmung (weitgehend) unbeeinflusst von Personmerkmalen (wie z.B. Fachinteresse) sein sollte. Bei einer Erfassung des Unterrichts mittels hoch-inferenter Indikatoren, die sich zudem auf mit der Person der beurteilenden Schülerin bzw. des beurteilenden Schülers verwobene Aspekte beziehen (z.B.

fachbezogene Lernmotivation etc.), ist hingegen mit unterschiedlichen Wahrnehmungen in den Fächern Deutsch und Englisch zu rechnen.

4. *Intraindividuelle Unterschiede*: Wie stark unterscheiden Schülerinnen und Schüler zwischen dem Unterricht im Fach Deutsch und dem Unterricht im Fach Englisch? Hier geht es um die Frage von Beurteiler-Tendenzen (Milde- vs. Strenge-Tendenz): Hohe Zusammenhänge würden hier bedeuten, dass Schülerinnen und Schüler Unterricht mit einer *generellen Tendenz* (also in beiden Fächern gleichermaßen) beurteilen.
5. *Itemformulierung*: Welche Rolle spielen Ich-Bezug vs. Klassen-Bezug bei der Formulierung der Items? Beim Adressatenbezug ist von einer Konfundierung mit der Wahrnehmungsperspektive auszugehen. Entsprechend sollte bei Items mit Klassen-Bezug die Übernahme der Klassenperspektive eine gewisse Rolle spielen: Da der kognitiv fundierten Survey-Forschung zufolge auch bei der Beantwortung einer Frage aus Sicht der Mitschülerinnen und Mitschüler zunächst das eigene Urteil zugrunde gelegt wird, und weiterhin mit einem generellen (d.h. von der Itemformulierung unabhängigen) Einfluss der Urteile der Mitschülerinnen und Mitschüler auf die Unterrichtswahrnehmung zu rechnen ist, werden hier relativ geringe Effekte aufgrund der Itemformulierung erwartet. Zumindest tendenziell sollten sich bei klassenbezogenen Items höhere Übereinstimmungen zeigen. Weiterhin wird erwartet, dass der Adressatenbezug eine eigenständige Varianzquelle darstellt (im Sinne eines sogenannten Methodenfaktors), die sich von den jeweiligen Unterrichtsinhalten der Items isolieren lässt.
6. *Determinanten der Unterrichtsperzeption*: Beurteilen Mädchen den Unterricht positiver? Führen höhere Fachleistungen und bessere Schulnoten zu einer günstigeren Unterrichtsbeurteilung? Variieren die Einflüsse der verschiedenen Unterrichtswahrnehmungen auf die Globalurteile bezüglich der Lehrkräfte über Klassen hinweg?

3. Datengrundlage und Methoden

Gegenstand dieses Kapitels ist zunächst eine kurze Beschreibung der DESI-Studie, in deren Rahmen die der vorliegenden Untersuchung zugrunde liegenden Daten erhoben wurden. Es folgen Beschreibungen der Stichprobe sowie der hier verwendeten Instrumente. Im letzten Teilkapitel wird auf die methodischen Grundlagen und Besonderheiten bei der Analyse von Daten zur Unterrichtswahrnehmung aus Schülersicht eingegangen.

3.1 Die DESI-Studie (Deutsch Englisch Schülerleistungen International)

Die Schulleistungsstudie „Deutsch Englisch Schülerleistungen International“ (DESI) wurde in den Jahren 2001 bis 2006 von einem interdisziplinären Konsortium (Bildungsforscher sowie Deutsch- und Englischdidaktiker) unter der Federführung des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) auf der Grundlage einer Ausschreibung der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) durchgeführt (vgl. dazu Beck & Klieme, 2006; DESI-Konsortium, 2006; Beck, Bundt & Gomolka, in Druck). Nach einer umfangreichen Pilotierungs- und Validierungsstudie im Herbst 2002 wurden im Rahmen der Hauptuntersuchung in der Zeit von September 2003 bis Juni 2004 die sprachlichen Leistungen von etwa 11000 Neuntklässlern aus 219 Schulen sowie Bedingungen der Kompetenzentwicklung in den Fächern Deutsch und Englisch in einem Längsschnittdesign (mit zwei Messzeitpunkten) untersucht.

DESI stellt eine bundesweite Untersuchung dar, an der sich allgemein bildende Schulen aus sämtlichen Ländern der Bundesrepublik Deutschland beteiligten. Die Stichprobe ist somit repräsentativ für die Zielpopulation der Neuntklässler an allgemein bildenden Schulen in der Bundesrepublik Deutschland – vorausgesetzt es werden die aufgrund unterschiedlicher Ziehungswahrscheinlichkeiten (in der Stichprobe sind beispielsweise bilinguale Klassen deutlich überrepräsentiert) notwendigen Populationsgewichte verwendet. Zudem gewährleisten hohe Beteiligungsraten (Leistungstests: 94% zum ersten, 95% zum zweiten Messzeitpunkt; Lehrerfragebögen: 86% bzw. 95%; Schülerfragebögen: 92% zu beiden Messzeitpunkten; Elternfragebögen: 69%) die Repräsentativität der Studie.

Die Untersuchung von Unterrichtsmerkmalen im Hinblick auf die Leistungsentwicklung von Schülerinnen und Schülern stellte – neben der Erfassung des Kompetenzstands von Schülerinnen und Schülern am Ende des neunten Schuljahres im Deutschen und im Englischen – einen Schwerpunkt des DESI-Projekts dar[10]. Um solche Zusammenhänge optimal analysieren

[10] Die besondere Bedeutung des Unterrichts im Projekt DESI wird auch durch die begleitende Videostudie des Englischunterrichts in 105 Klassen unterstrichen, die zwischen den beiden Erhebungszeitpunkten für die Leis-

zu können, wurden bei der Stichprobenziehung komplette Klassen (soweit vorhanden: zwei Klassen pro gezogener Schule) zugrunde gelegt.

3.2 Beschreibung der Stichprobe

An DESI nahmen Schülerinnen und Schüler aus insgesamt 427 Klassen teil. In einigen Fällen handelt es sich dabei aber lediglich um vollständige Deutsch- bzw. Englisch-Klassen, d.h. die Schülerinnen und Schüler bildeten nur in Bezug auf jeweils eines der beiden untersuchten Unterrichtsfächer eine Klasse, während sie sich im jeweils anderen Fach auf verschiedene Klassen verteilten. Insgesamt liegen Daten von je 388 vollständigen Englisch- bzw. Deutsch-Klassen vor. In einigen Fällen bekamen die Schülerinnen und Schüler dieser Klassen im Laufe des Schuljahres eine neue Deutsch- bzw. Englischlehrkraft. Die Unterrichtswahrnehmungen aus Schülersicht, die zum zweiten Messzeitpunkt erhoben wurden, sollten allerdings auf einer hinreichend großen Verhaltensstichprobe der jeweils unterrichtenden Lehrkraft basieren. Deshalb reduziert sich die Stichprobe für solche Fragestellungen auf 379 Englisch- bzw. 377 Deutsch-Klassen, die zumindest bereits vor der Vergabe der Halbjahreszeugnisse von derselben Lehrkraft (im jeweiligen Unterrichtsfach) unterrichtet wurden.

Da in der vorliegenden Untersuchung Unterrichtswahrnehmungen bezüglich der Fächer Deutsch und Englisch parallel analysiert werden, beziehen sich die Auswertungen auf 8423 Schülerinnen und Schüler aus 330 Klassen in 177 Schulen, die von derselben Deutsch- bzw. Englischlehrkraft bereits vor Vergabe der Halbjahreszeugnisse unterrichtet wurden. In 24 Klassen wurden dabei die Schülerinnen und Schüler in beiden Fächern von derselben Lehrkraft unterrichtet, in 280 Klassen von unterschiedlichen Lehrkräften. In den verbleibenden 26 Klassen ist unklar, ob die Schülerinnen jeweils von derselben Lehrkraft oder von unterschiedlichen Lehrkräften in den Fächern Deutsch und Englisch unterrichtet wurden[11].

Im Fach Englisch gaben insgesamt 149 Lehrkräfte an, die Klasse bereits vor Beginn der neunten Klassenstufe unterrichtet zu haben, 119 Lehrkräfte haben die Klasse zu Beginn, fünf im Verlauf der neunten Klassenstufe übernommen. Von 67 Lehrkräften liegen keine diesbezüglichen Angaben vor. Ähnlich verhält es sich auch im Fach Deutsch: Dort gaben 142 Lehrkräfte an, die Klasse bereits vor Beginn der neunten Klassenstufe unterrichtet zu haben, 117 Lehrkräfte haben die Klasse zu Beginn, drei im Verlauf der neunten Klassenstufe übernommen, und von 68 Lehrkräften liegen keine Angaben hierzu vor.

---

tungstests und Fragebögen durchgeführt wurde (vgl. dazu T. Helmke, A. Helmke, Schrader, Wagner, Nold & Schröder, in Druck).

[11] Die Einteilung basiert auf Schülerangaben. Dabei wurden Klassen, bei denen mindestens 80% der Angaben übereinstimmten, als von derselben Lehrkraft bzw. von unterschiedlichen Lehrkräften unterrichtet eingestuft.

3.3 Beschreibung der Instrumente

Für die vorliegende Untersuchung wurden nur solche Skalen zum Unterricht aus Schülersicht ausgewählt, die bei *allen* Schülern sowohl für das Fach Englisch als auch für das Fach Deutsch erhoben wurden. Um der Frage des Item-*wordings* (Ich- vs. Klassen-Bezug) nachgehen zu können, wurden sämtliche Skalen einbezogen, die jeweils sowohl Items mit Ich-Bezug als auch mit Klassen-Bezug enthalten. Diese drei Skalen (*Thematische Motivierung*, *Verständlichkeit*, *Schülerorientierung*) wurden in DESI (Klieme et al., in Druck; A. Helmke et al., in Druck-a) aufgrund ihrer hohen Interkorrelationen sowohl im Fach Deutsch als auch im Fach Englisch auf Klassenebene jeweils einem Faktor zweiter Ordnung (Prozessqualität) zugeordnet.

Zusätzlich wurde die Skala *Strukturiertheit*, die im Fach Deutsch einem gleichnamigen Faktor zugeordnet werden kann (vgl. Klieme et al., in Druck), im Fach Englisch ein eigenständiges Merkmal auf Klassenebene darstellt (vgl. A. Helmke et al., in Druck-a), sowie die Skala *Klassenführung*, die im Fach Englisch einem gleichnamigen Faktor auf Klassenebene zugeordnet werden kann (ebd.), während diese im Fach Deutsch auf den Faktor *Strukturiertheit* lädt (Klieme et al., in Druck), ausgewählt. Diese Skalen wurden in die vorliegenden Analysen einbezogen, um ein möglichst hohes Maß an Heterogenität bezüglich der Faktorstruktur auf Klassenebene sicherzustellen.

In Tabelle 2 sind die der vorliegenden Untersuchung zugrunde liegenden Unterrichtsmerkmale und die entsprechenden Items (Indikatoren) sowie Angaben zur Herkunft der Skalen aufgeführt. Bezüglich des Adressatenbezugs enthalten – wie oben bereits erwähnt – die ersten drei Unterrichtsmerkmale (*Thematische Motivierung*, *Verständlichkeit* und *Schülerorientierung*) jeweils sowohl Items mit Ich- als auch mit Klassen-Bezug, während die Skalen *Strukturiertheit* und *Klassenführung* durchgängig aus klassenbezogenen Items bestehen. Als Kriterium für ichbezogene Itemformulierungen wird hier der *ausdrückliche* Bezug („ich", „mich") auf die individuelle Schülerin bzw. den individuellen Schüler herangezogen, während Formulierungen sowohl mit (z.B. „unsere") als auch ohne expliziten Bezug zur Klasse als klassenbezogen behandelt werden.

Was die Wahrnehmungsperspektive der Items anbelangt, ist aufgrund der Formulierung „mein Lehrer/ meine Lehrerin" durchgängig von einer individuellen Sichtweise[12] auszugehen. Einzige Ausnahme stellt das erste Item der Skala *Verständlichkeit* dar, das sich auf „den Unterricht" bezieht. Allerdings ist hier die Annahme der Klassenperspektive wohl wenig sinn-

[12] Dies lässt sich auch verdeutlichen, wenn man den expliziten Hinweis auf die Klassenperspektive ergänzt: „Wir finden, mein Englischlehrer…" Eine derart ungewöhnliche Satzergänzung dürfte wohl kaum implizit vorgenommen werden.

voll, da in diesem Kontext (ob eine individuelle Schülerin bzw. ein individueller Schüler die Aufgabenstellungen versteht) eine Beurteilung aus Sicht der Klasse kaum in Frage kommt.

Bei den klassenbezogenen Items der Unterrichtsmerkmale *Thematische Motivierung* und *Verständlichkeit* ist eine auf die Mitschülerinnen und Mitschüler bezogene Beurteilung kaum möglich (bzw. nur indirekt über das Verhalten der Mitschülerinnen und Mitschüler): Ob beispielsweise die Unterrichtsinhalte für die Mitschülerinnen und Mitschüler „interessant" sind (zweites Item der Skala *Thematische Motivierung*) lässt sich nur indirekt erschließen (z.B. aufgrund der Beteiligung am Unterrichtsgeschehen). Hier ist es naheliegender, von einer Urteilsbildung auf der Basis vorliegender Informationen der Mitschülerinnen und Mitschüler auszugehen. Eine solche Urteilsbildung wird aber wiederum durch die individuelle Wahrnehmungsperspektive der Itemformulierungen eingeschränkt.

Mit Ausnahme des ersten Items der Skala *Thematische Motivierung*, das sich (aufgrund der Kombination von „manchmal" und der auf Ablehnung bzw. Zustimmung bezogenen Antwortkategorien indirekt) auf Häufigkeiten bezieht, sind alle Items im Sinne von Intensitäten zu interpretieren. Allen Items liegt dasselbe vierstufige Antwortformat zugrunde: stimmt gar nicht, stimmt eher nicht, stimmt eher, stimmt ganz genau.

Betrachtet man die verschiedenen Unterrichtsmerkmale, so fällt auf, dass sich die Konstrukte *Thematische Motivierung*, *Verständlichkeit* und *Schülerorientierung* zwar einerseits auf die weitgehend durch die Lehrkraft vorgegebene Unterrichtssituation beziehen; andererseits sollten hier bestimmte Einstellungen und (Lern-)Voraussetzungen der Schülerinnen und Schüler eine große Rolle spielen. Im Gegensatz dazu lassen sich die beiden Unterrichtsmerkmale *Strukturiertheit* und *Klassenführung* eher „objektiv" (aus der Perspektive eines „unbeteiligten" Beobachters) beurteilen. Im Falle der *Klassenführung* ist jedoch die Objektivität der Beurteilung dadurch eingeschränkt, dass zwar das Lehrerverhalten relativ gut beobachtbar sein dürfte, nicht hingegen das Schülerverhalten. Für Schülerinnen und Schüler ist es möglicherweise schwierig einzuschätzen, ob die Lehrkraft „alles mitbekommt, was in der Klasse passiert", da dies voraussetzt, dass die Schülerinnen und Schüler über das Verhalten ihrer Mitschüler immer informiert sind.

Tabelle 2: Skalen und Items zum Unterricht aus Schülersicht

Unterrichts-merkmal / Items	Itemtext (Fach Englisch[13])	Quelle[14]
Thematische Motivierung		
Item 1 (Ich-Bezug)	Mein Englischlehrer/ meine Englischlehrerin kann mich manchmal richtig für die Unterrichtsthemen begeistern.	MARKUS (adaptiert)
Item 2 (Klassen-Bezug)	Mein Englischlehrer/ meine Englischlehrerin kann auch trockene Themen wirklich interessant machen.	MARKUS (adaptiert)
Verständlichkeit[15]		
Item 1 (Ich-Bezug)	Die Aufgabenstellungen im Englischunterricht sind für mich klar und verständlich.	DESI
Item 2 (Klassen-Bezug)	Wenn mein Englischlehrer/ meine Englischlehrerin etwas erklärt, dann gibt er/ sie dazu anschauliche Beispiele.	DESI, angelehnt an SALVE
Item 3 (Klassen-Bezug)	Mein Englischlehrer/ meine Englischlehrerin drückt sich klar und deutlich aus.	DESI, angelehnt an SALVE
Item 4 (Ich-Bezug)	Mein Englischlehrer/ meine Englischlehrerin erklärt so, dass ich es verstehen kann.	SALVE (adaptiert)
Schüler-orientierung[16]		
Item 1 (Klassen-Bezug)	Mein Englischlehrer/ meine Englischlehrerin geht auf unsere Vorschläge ein.	DESI, angelehnt an MARKUS, SALVE
Item 2 (Klassen-Bezug)	Mein Englischlehrer/ meine Englischlehrerin ermutigt uns, unsere eigene Meinung auszudrücken.	DESI
Item 3 (Klassen-Bezug)	Wenn jemand eine gute Idee hat, dann geht mein Englischlehrer/ meine Englischlehrerin darauf ein.	DESI
Item 4 (Ich-Bezug)	Mein Englischlehrer/ meine Englischlehrerin gibt mir Gelegenheit, meine Meinung zu sagen.	angelehnt an PISA 2000, PISA 2003
Item 5 (Ich-Bezug)	Mein Englischlehrer/ meine Englischlehrerin interessiert sich für das, was ich zu sagen habe.	PISA 2003
Strukturiertheit		
Item 1	Zu Beginn der Stunde gibt mein Englischlehrer/ meine Englischlehrerin eine Übersicht, worum es geht.	DESI
Item 2	Am Ende der Stunde fasst mein Englischlehrer/ meine Englischlehrerin das Wichtigste zusammen.	DESI, angelehnt an MARKUS, PISA 2000
Item 3	Mein Englischlehrer/ meine Englischlehrerin gibt Hinweise, worauf es in der Unterrichtsstunde besonders ankommt.	DESI
Klassenführung		
Item 1	Mein Englischlehrer/ meine Englischlehrerin sorgt dafür, dass die Schüler die ganze Stunde über aufpassen.	SALVE (adaptiert)
Item 2	Mein Englischlehrer/ meine Englischlehrerin bekommt alles mit, was in der Klasse passiert.	SALVE (adaptiert)

[13] Aus Platzgründen sind hier nur die Itemformulierungen für das Fach Englisch angegeben. Ersetzt man jeweils „Englisch" durch „Deutsch", so erhält man die analogen Formulierungen für das Fach Deutsch.

[14] Bei den Quellenangaben werden die gängigen Akronyme verschiedener Projekte verwendet (in Klammern wird hier auf Publikationen verwiesen, die die Itemtexte beinhalten): MARKUS: „Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext" (A. Helmke, Hosenfeld, Schrader & Wagner, 2002); PISA (2000, 2003): „Programme for International Student Assessment" (Kunter, Schümer, Artelt, Baumert, Klieme, Neubrand, Prenzel, Schiefele, Schneider, Stanat, Tilmann & Weiß, 2002; PISA-Konsortium Deutschland, 2006); SALVE: „Systematische Analyse des Lernverhaltens und des Verständnisses in Mathematik: Entwicklungstrends und Fördermöglichkeiten" (A. Helmke, Hosenfeld & Schrader, 2002)

[15] Die DESI-Skala *Verständlichkeit* wurde um das Item „Mein Englisch(Deutsch-)lehrer/ meine Englisch (Deutsch-)lehrerin erklärt so, dass ich es verstehen kann." ergänzt.

[16] Das erste Item der DESI-Skala *Schülerorientierung* („Mein Englischlehrer/ meine Englischlehrerin lässt uns bei Gruppenarbeit Themen oder Aufgaben auswählen.") wurde aufgrund seiner konditionalen Formulierung – die Beantwortung der Frage hängt davon ab, *ob* Gruppenarbeit stattfindet – hier nicht berücksichtigt.

Zusätzlich zu den oben beschriebenen Unterrichtsmerkmalen wurden verschiedene Prädiktoren der individuellen und kollektiven Unterrichtswahrnehmungen in die Analysen einbezogen. Neben dem Geschlecht der Schülerinnen und Schüler wurden deren Fachleistungen in Form von Schulnoten (Schülerangaben: Note Ende 8. Klassenstufe, erwartete Note am Ende der 9. Klassenstufe) bzw. Leistungstests – Textrekonstruktion[17] (C-Test; Englisch) und Sprachbewusstheit[18] (Deutsch) – als unabhängige Variablen betrachtet. Die beiden Leistungstests wurden aus der Gesamtheit der in DESI zur Verfügung stehenden Subtests ausgewählt, weil sie (1) im Längsschnitt erhoben worden sind (was nicht für alle DESI-Subtests gilt) und (2) weil sie „zentrale" Kompetenzen (d.h. sie weisen jeweils fachspezifisch hohe Interkorrelationen mit anderen Sub*scores* auf) der Deutsch- bzw. Englischleistung darstellen (vgl. Jude, Klieme, Eichler, Lehmann, Nold, Schröder, Thomé & Willenberg, in Druck).

Bei den Leistungs*scores* handelt es sich um sogenannte *plausible values* (PVs; vgl. Mislevy, Beaton, Kaplan & Sheehan, 1992). Diese basieren auf der Überlegung, dass im Falle nicht vollständig reliabler Test*scores* bezüglich des Mittelwerts, der Verteilung und der Zusammenhänge mit anderen Merkmalen – die im sogenannten Hintergrundmodell repräsentiert sind – optimale Schätzer (die weitgehend die latenten Schätzer reproduzieren) aus einer Verteilung gezogen werden können. Dazu werden bei der Ziehung der PVs neben dem Test*score* gleichzeitig die zuvor geschätzten latenten Populationscharakteristika berücksichtigt. Erzielen beispielsweise Mädchen in einem Leistungstest höhere Werte als Jungen, dann wird – bei identischem Testscore – für Mädchen aus einer Verteilung, die oberhalb derer der Jungen liegt, gezogen. Da die so erzeugten Werte – analog zu imputierten Werten – mit einer gewissen Unsicherheit versehen sind (es handelt sich um Ziehungen), werden sinnvollerweise mehrere Werte pro Person gezogen (in DESI wurden jeweils fünf PVs erzeugt). Analysen müssen entsprechend mehrfach wiederholt und die Ergebnisse zusammengeführt werden[19].

Besonders vorteilhaft sind PVs beispielsweise auch bei der Analyse von Veränderungen, da Veränderungsmessungen oft mit besonders schwierigen Reliabilitätsproblemen verbunden sind (v.a. wenn die verschiedenen Messungen hoch korreliert sind). In DESI wurden dazu Differenzwerte aus jeweils zwei PVs (erster PV des jeweiligen Leistungstests am Ende des Schuljahres minus erster PV des jeweiligen Leistungstests zu Beginn des Schuljahres) gebildet, die aus einem gemeinsamen (zweidimensionalen, also pro Messzeitpunkt eine Dimension) Modell stammten (vgl. Hartig, Jude & Wagner, in Druck).

[17] Zum mithilfe des C-Tests erfassten theoretischen Konstrukt vgl. Harsch und Schröder (2006).
[18] Zum Konzept *Sprachbewusstheit* in DESI vgl. Eichler und Nold (2006). Zum mithilfe des Tests „Sprachbewusstheit-Deutsch" erfassten Konstrukt vgl. Eichler (2006).
[19] Viele Statistikprogramme bieten mittlerweile Optionen für die Auswertungen mit Imputationen bzw. *plausible values* an, so dass sich der Mehraufwand bei solchen Analysen für den Anwender in Grenzen hält.

Neben den genannten Vorteilen sind im Zusammenhang mit PVs auch einige Nachteile zu nennen: (1) Das Modell, mit dessen Hilfe die PVs generiert werden, entscheidet über die „Brauchbarkeit“ der Ergebnisse. So werden beispielsweise Zusammenhänge der PVs mit Variablen, die nicht im Hintergrundmodell enthalten waren, unterschätzt. (2) Obwohl im Hintergrundmodell auch Merkmale auf Aggregatebene berücksichtigt werden können (indem etwa Klassenmittelwerte einbezogen werden), handelt es sich im Allgemeinen bei der Ziehung von PVs nicht um ein „echtes“ Mehrebenenmodell[20]. Insofern ist die Ebenenstruktur der Daten in den PVs nicht optimal repräsentiert. Insbesondere sogenannte *random slopes* – dabei handelt es sich um Mehrebenenmodelle mit beispielsweise über Schulklassen hinweg variierenden Regressionsgewichten – lassen sich derzeit wohl nur sehr bedingt bei der Generierung von PVs berücksichtigen. (3) Bei umfangreichen und rechenintensiven Analysen (beispielsweise Modelle mit latenten Variablen) stellt der höhere Zeitaufwand aufgrund der mehrfachen Schätzung eines Modells mit jeweils verschiedenen PVs oft einen erheblichen Nachteil gegenüber „einfachen“ *Maximum-Likelihood*-Schätzern dar.

Sowohl als Prädiktor wie auch als abhängige Variable (vgl. Kap. 2.2, Abbildung 1) wird das auf die Lehrkraft bezogene Globalurteil behandelt. Dieses wurde separat für jedes Unterrichtsfach mithilfe eines einzigen viertstufigen[21] Items erfasst: „Ich mag meinen Englischlehrer/ meine Englischlehrerin gern.“; „Ich mag meinen Deutschlehrer/ meine Deutschlehrerin gern.“

3.4 Methoden

Das folgende Kapitel ist in drei Unterkapitel unterteilt. Im ersten Teilkapitel wird der Frage der Übereinstimmung bzw. Reliabilität aus methodischer Sicht nachgegangen. Die Grundlagen mehrebenenanalytischer konfirmatorischer Faktorenanalysen sind Gegenstand des zweiten Unterkapitels. Die dort aufgeworfenen methodischen Fragen werden abschließend im Rahmen einer Simulationsstudie überprüft.

3.4.1 Absolute Übereinstimmung und Reliabilität von Urteilen

Zur Beurteilung der Güte von aggregierten Urteilen lassen sich zwei verschiedene Ansätze unterscheiden (vgl. Bliese, 2000; Lüdtke et al., 2006): Der erste Ansatz bezieht sich auf die *absolute Übereinstimmung* von Beurteilern bezüglich des jeweils zu beurteilenden Aggre-

[20] Solche Modelle dürften allein aufgrund des notwendigen Rechenaufwandes von Mehrebenen-IRT-Modellen, bei denen zusätzlich noch eine Vielzahl an Hintergrundvariablen auf jeder Ebene berücksichtigt werden müsste, derzeit gar nicht möglich sein.
[21] Identisches Antwortformat wie bei den Items zur Unterrichtswahrnehmung (s.o.).

gatmerkmals, der zweite auf die *relative Übereinstimmung* oder *Reliabilität* der aggregierten Urteile.

Zur Beurteilung der *absoluten Übereinstimmung* wurde eine ganze Reihe von Maßen entwickelt (vgl. z.B. Burnkrant, 2003), von denen Lüdtke et al. (2006) den *within-group interrater reliability*-Index (r_{wg}) von James, Demaree und Wolf (1984) und den *average deviation*-Index (AD) von Burke, Finkelstein und Dusig (1999) zur Beurteilung der Übereinstimmung der Unterrichtswahrnehmungen von Schülerinnen und Schülern vorschlagen. Der r_{wg}-Index steht dabei stellvertretend für eine Klasse von Übereinstimmungsmaßen, die auf dem Verhältnis einer *empirischen* Streuung (z.B. die Varianz der unterschiedlichen Wahrnehmungen bezüglich eines Unterrichtsmerkmals der Schülerinnen und Schüler einer Klasse; s_x^2) und einer *theoretischen* (Vergleichs-)Streuung (σ_E^2; der Subskript „E" steht dabei für *error*) – im Sinne eines theoretisch begründeten Erwartungswerts bezüglich der Fehlervarianz – basieren (vgl. dazu (2)). Im Falle des r_{wg} wird die Varianz einer Gleichverteilung[22] mit A Antwortkategorien als Referenzmaß verwendet (vgl. (3)). Dies entspricht der Annahme, dass (theoretisch) alle Antwortkategorien eines Items mit gleicher Wahrscheinlichkeit gewählt werden.

$$r_{wg} = 1 - \frac{s_x^{\,2}}{\sigma_E^{\,2}} \tag{2}$$

$$\sigma_E^{\,2} = \frac{A^2 - 1}{12} \tag{3}$$

Als theoretische Basis für diese Annahme kann beispielsweise das *range-frequency*-Modell herangezogen werden (Parducci, 1965; vgl. auch Tourangeau et al., 2000). Im Rahmen dieses Modells werden zwei verschiedene Prinzipien bei der Beantwortung von Rating-Skalen postuliert. Das *range*-Prinzip bezieht sich auf die Einteilung des gesamten Bereiches der Antwortalternativen auf der Basis des Minimums bzw. Maximums hinsichtlich des zu beurteilenden Merkmals. Das *frequency*-Prinzip postuliert, dass alle Antwortalternativen mit gleicher Häufigkeit gewählt werden.

Bereits James et al. (1984) weisen allerdings darauf hin, dass eine Gleichverteilung als Referenz nicht immer optimal erscheint. Wenn beispielsweise mit verschiedenen Verzerrungstendenzen bezüglich des Antwortverhaltens (z.B. soziale Erwünschtheit) gerechnet wer-

[22] Daraus ergibt sich, dass der r_{wg}-Index auch negative Werte annehmen kann, wenn „extremere" empirische Verteilungen als die Gleichverteilung vorliegen (z.B. bimodale Verteilungen mit Modalwerten bei Minimum und Maximum der Skala).

den muss, dann sollte eine andere Vergleichsverteilung (als sogenannte „Nullverteilung“) gewählt werden. Auch die Verankerung der Antwortalternativen sollte hier eine Rolle spielen: Wenn die Skala beispielsweise an beiden Enden (kaum zu erwartende) Extrema enthält (wie etwa: „niemals“, „immer“), dann könnte die Wahl einer Gleichverteilung als Referenz zu einer deutlichen Überschätzung der Beurteiler-Übereinstimmung führen.

Bei der Wahl einer solchen Nullverteilung sollten James et al. (ebd.) zufolge auch empirische Ergebnisse berücksichtigt werden, wenngleich die empirische Verteilung nicht grundsätzlich als Nullverteilung verwendet werden sollte. Die Frage nach einer geeigneten Nullverteilung verkompliziert sich zusätzlich, wenn mehrere Items gleichzeitig in die Analysen einbezogen werden sollen, da dann z.B. Abhängigkeiten zwischen den einzelnen Antworten bzw. Polungen berücksichtigt werden müssen (vgl. Lindell, 2001).

Daneben ist die maximal mögliche Varianz auch vom Mittelwert der Urteile abhängig. Liegt ein Mittelwert beispielsweise bei 4.5 auf einer 5-stufigen Likert-Skala, deren Antwortalternativen mit den Werten 1 bis 5 kodiert wurden, so ist mit einer deutlich geringeren Varianz zu rechnen als bei einem Mittelwert von 2.5. Brown und Hauenstein (2005) schlagen deshalb einen Index vor (a_{wg}), der diese Abhängigkeit der Varianz vom jeweiligen Mittelwert berücksichtigt, indem für jeden Mittelwert eine eigene Nullverteilung zugrunde gelegt wird.

Ursprünglich wurde der r_{wg}-Index entwickelt, um die Übereinstimmung von Beurteilern zu überprüfen, wenn nur *ein* Objekt beurteilt wird (vgl. James et al., 1984). Dementsprechend weisen James, Demaree und Wolf (1993) darauf hin, dass der r_{wg}-Index kein Maß der Reliabilität (die psychometrische Testtheorie setzt hierfür Varianz bezüglich des *true scores* – das heißt hier: zwischen *verschiedenen* Objekten – voraus), sondern ein Maß der Übereinstimmung darstellt.

Der von Burke et al. (1999) vorgeschlagene AD-Index (vgl. (4)) entspricht der mittleren absoluten Abweichung der einzelnen Urteile vom (arithmetischen) Mittelwert aller Urteile bzw. dem Median – bezogen auf *eine* Klasse. Dieses (Abweichungs-)Maß kann im Sinne der Ursprungsmetrik der Antwortskala des jeweiligen Items interpretiert werden. Anstatt einer Nullverteilung wird hier der erwartete Antwortbereich (*null response range*) als Referenz gewählt, so dass die zugrunde liegende Verteilung der erwarteten Antworten unter der Bedingung völlig zufälliger Urteile zumindest nicht explizit modelliert wird. Im Rahmen zweier empirischer Studien konnten Burke et al. (ebd.) einen erwartungsgemäß hohen (negativen) Zusammenhang[23] des AD- und des r_{wg}-Index nachweisen.

[23] Ein hoher Zusammenhang wurde deshalb erwartet, da sich die beiden Maße auf Streuungen innerhalb der Klasse beziehen. Der Zusammenhang ist negativ, weil sich der AD-Index auf Abweichungen, der r_{wg}-Index hingegen auf Übereinstimmungen bezieht.

$$\mathrm{AD} = \frac{\sum_{n=1}^{N} \left| x_k - \bar{x} \right|}{N} \tag{4}$$

Unklar bleibt, warum nicht die Standardabweichung (die sich ebenfalls auf die ursprüngliche Metrik der Antwortskala bezieht) anstelle der absoluten mittleren Abweichung verwendet wird. Die Standardabweichung hat den Vorteil, dass sie bezogen auf den arithmetischen Mittelwert als Erwartungswert definiert ist. Die absolute Abweichung hingegen wird minimiert, wenn als Erwartungswert der Median eingesetzt wird (vgl. Dwight, 1957; Diehl & Kohr, 1994). Da üblicherweise der Mittelwert (als im Gegensatz zum Median suffizientem Maß) als Aggregatmerkmal verwendet wird, wäre die Beurteilung der Übereinstimmung anhand der Standardabweichung wohl adäquater.

Im Gegensatz zur Beurteilung der absoluten Übereinstimmung setzt die Ermittlung der *Reliabilität* – wie bereits erwähnt – Urteile verschiedener Beurteiler bezüglich *unterschiedlicher Objekte* voraus, da Reliabilität als *Varianz* des *wahren Werts* (*true score*) im Verhältnis zur Gesamtvarianz definiert ist. Anders formuliert: Der relative Anteil der *true score*-Varianz bezogen auf die Gesamtvarianz entspricht der Reliabilität eines Maßes. Übertragen auf die Unterrichtswahrnehmung aus Schülersicht bedeutet dies, dass die Reliabilität eines *einzigen* Schülerurteils dem Verhältnis der Varianz dieses Merkmals zwischen Klassen und der Gesamtvarianz entspricht. Die Gesamtvarianz lässt sich wiederum aufteilen in Varianz zwischen Klassen und Varianz innerhalb von Klassen, wobei erstere den wahren Wert repräsentiert, letztere hingegen den Fehlerterm. Das entsprechende statistische Maß wird als ICC (*intraclass correlation*, ICC) bezeichnet (vgl. z.B. Snijders & Bosker, 1999; Kreft & de Leeuw, 1998; Hox, 2002).

Die ICC entspricht somit der Reliabilität des Urteils *einer* Schülerin bzw. *eines* Schülers. Bei der Aggregation der individuellen Urteile (z.B. Bildung von Klassenmittelwerten) erhöht sich die Reliabilität mit zunehmender Anzahl an Urteilen (also z.B. zunehmender Klassengröße), da die Aggregatvariable den wahren Wert zunehmend approximiert. Es lassen sich demnach zwei verschiedene Reliabilitätsmaße unterscheiden: Die ICC – auch ICC(1) genannt – bezieht sich auf die Reliabilität eines individuellen Urteils, die ICC(2) auf die der aggregierten Urteile (vgl. Bliese, 2000; Lüdtke et al., 2006).

Der Zusammenhang zwischen den beiden Maßen lässt sich anhand der mathematischen Definitionen (vgl. Snijders & Bosker, 1999) verdeutlichen (s. (5), (6)). τ^2 entspricht dabei der Varianz zwischen Schulklassen, σ^2 der Varianz innerhalb von Klassen, also dem Fehlerterm.

Dieser Fehlerterm wird bei der Aggregation (z.B. Bildung von Klassenmittelwerten) durch die (durchschnittliche) Anzahl an Schülerinnen und Schülern – mit *n* bezeichnet – pro Klasse geteilt.

$$\mathrm{ICC}(1) = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{5}$$

$$\mathrm{ICC}(2) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} = \frac{n \cdot \mathrm{ICC}(1)}{1 + (n-1) \cdot \mathrm{ICC}(1)} \tag{6}$$

Hier zeigt sich deutlich der Unterschied zwischen Übereinstimmungsmaßen und einem Reliabilitätsmaß: Eine sinnvolle Bestimmung der Reliabilität eines Instruments ist nur dann möglich, wenn eine hinreichend große Varianz bezüglich des zu erfassenden Merkmals auf Klassenebene vorliegt. Denn selbst bei hoher Übereinstimmung bezüglich eines Unterrichtsmerkmals ist die ICC gleich Null, wenn das Merkmal auf Klassenebene nicht variiert. Analog dazu ist die Reliabilität einer Skala gleich Null, wenn das zu messende Merkmal nicht variiert (z.B. zwischen Personen). In solchen Fällen bleibt unklar, ob bzw. wie reliabel das Instrument wäre, wenn man es z.B. in einer anderen Stichprobe mit großen Unterschieden bezüglich des zu erfassenden Merkmals einsetzen würde. Es lässt sich also letztlich nicht entscheiden, ob das zu messende Merkmal in der zurgunde liegenden Stichprobe nicht (oder zu wenig) variiert, oder ob das Messinstrument mangelhaft ist.

Bei der Beurteilung der Interrater-Reliabilität – aber auch bei der -Übereinstimmung (vgl. Lüdtke et al., 2006) – stellen sich zwei Fragen: (1) Ist die empirische Übereinstimmung statistisch signifikant? (2) Ist das Ausmaß der Übereinstimmung hoch genug, um eine reliable Messung eines Konstrukts auf Aggregatebene zu rechtfertigen? Die erste Frage lässt sich durch Überprüfung des Signifikanzniveaus der ICC überprüfen. Eine statistisch signifikante ICC[24] bedeutet, dass die Unterschiede auf Klassenebene nicht auf rein zufällige Unterschiede in der Zusammensetzung der Klassen zurückgeführt werden können. Das wäre etwa dann der Fall, wenn man die Daten von Schülerinnen und Schüler zufällig auf die verschiedenen Klas-

[24] Bei Varianzkomponenten sind Z-Tests allerdings häufig nicht zuverlässig. Bei kleineren Stichprobenumfängen ist ein Vergleich des r*andom-intercept*-Modells mit einem Nullmodell (ohne Variation auf Ebene 2) mittels eines *Loglikelihood Ratio Tests* angebracht. Ab einem Stichprobenumfang von etwa 100 Gruppen (hier Schulklassen) kann der Z-Test (vgl. Hox, 2002) als zuverlässig betrachtet werden. Da Varianzen definitionsgemäß nicht negativ sein können, wird in solchen Fällen einseitig getestet (vgl. Snijders & Bosker, 1999).

sen verteilen würde. Die zweite Frage muss auf der Basis theoretischer Überlegungen (und an empirischen Ergebnissen orientiert) beantwortet werden.

Da üblicherweise (latente, d.h. nicht direkt beobachtbare) Konstrukte auf der Basis mehrerer Items (also messfehlerbehafteter Indikatoren) gemessen werden, lassen sich im Zusammenhang mit der Unterrichtswahrnehmung aus Schülersicht drei verschiedene Reliabilitäten unterscheiden, nämlich die Reliabilitäten der Indikatoren einer latenten Dimension auf Individual- und auf Klassenebene sowie die Interrater-Reliabilität. Die ICCs von manifesten Variablen (insbesondere bei Einzelitems, aber auch – je nach Reliabilität – bei Skalen) stellen aufgrund des höheren Messfehlers auf Individualebene (verglichen mit dem Messfehler auf Klassenebene) eine Unterschätzung der tatsächlichen, d.h. der auf die latenten Variablen bezogenen ICC dar (vgl. B. O. Muthén, 1991).

Die Berechnung latenter ICCs setzt identische Messmodelle bezüglich der latenten Variablen auf beiden Ebenen voraus (also identische Faktorladungen sowohl auf Individual- als auch auf Klassenebene)[25]. Ansonsten sind die latenten Faktoren – und somit auch die zur Berechnung der ICCs erforderlichen Varianzen der latenten Variablen auf beiden Ebenen – nicht vergleichbar. Zeigen sich deutliche Abweichungen zwischen beiden Messmodellen, so deutet dies auf die Verschiedenheit der Konstrukte auf Individual- und Klassenebene hin. Wenn im Sinne des *fuzzy composition process*-Modells (vgl. Kap. 2.1.4.2) in solchen Fällen von einer gewissen Analogie zwischen beiden Konstrukten ausgegangen werden kann, so lassen sich zumindest die ICCs bezogen auf die *true score*-Varianzanteile der einzelnen Indikatoren berechnen.

Vergleicht man Maße der absoluten Übereinstimmung mit der *Interrater*-Reliabilität, so zeigen sich für beide Verfahren abhängig vom jeweiligen Kontext, in dem sie eingesetzt werden, verschiedene Vor- und Nachteile. Verglichen mit der Reliabilität ist ein Vorteil der Maße der absoluten Übereinstimmung darin zu sehen, dass für jede einzelne Beobachtungseinheit ein Kennwert berechnet werden kann[26]. Als Kriterium für hinreichend gute Übereinstimmung wird für den r_{wg}-Index beispielsweise meist ein Wert von mindestens .70 vorgeschlagen[27]. Auf dieser Basis könnten beispielsweise Klassen mit zu niedriger Übereinstimmung bezüglich eines Unterrichtsmerkmals von Analysen ausgeschlossen werden. Inwiefern sich ein solcher

[25] Von dieser Annahme wird – zumindest implizit – bei (Standard-)Mehrebenenanalysen mit manifesten Variablen ausgegangen.

[26] In Untersuchungen aus dem Bereich der Organisationspsychologie wird häufig nur eine einzige Aggregateinheit (z.B. eine Arbeitsgruppe, eine Organisation etc.) untersucht (vgl. Burke et al., 1999). In solchen Fällen kann nur die absolute Übereinstimmung (und nicht die Reliabilität) überprüft werden.

[27] Harvey und Hollander (2004) weisen jedoch darauf hin, dass dieses Kriterium häufig als zu niedrig beurteilt werden muss.

Ausschluss (theoretisch) rechtfertigen lässt, ist bisher allerdings eine offene Frage (vgl. Lüdtke et al., 2006).

Bei großen, repräsentativen Stichproben – wie im vorliegenden Fall – ist die Frage nach einzelnen Klassen mit geringer Übereinstimmung von geringerer Bedeutung. Hier ist die (globale) relative Übereinstimmung im Sinne einer Interrater-Reliabilität wesentlich aussagekräftiger als die mittlere absolute Übereinstimmung, da die Güte eines Messinstruments bezogen auf die Varianz des zu erfassenden Aggregatmerkmals (in der Population) beurteilt werden kann: Wenn in der Population ein Merkmal relativ geringe Variation aufweist, dann wird zu dessen (reliabler) Erfassung ein besonders präzises Messinstrument benötigt. Umgekehrt können Merkmale die eine hohe Variation aufweisen mit „gröberen" Instrumenten erhoben werden.

Ein weiterer Vorteil der Interrater-Reliabilität ist darin zu sehen, dass diese auch – im Gegensatz zu den oben dargestellten Übereinstimmungsindizes, die (implizit) metrische Daten voraussetzen – für ordinale Daten berechnet werden kann. Zusätzlich können ICCs – wie bereits erwähnt – auch für latente (also messfehlerbereinigte) Variablen ermittelt werden. Aus den genannten Gründen werden in der vorliegenden Arbeit *Interrater*-Reliabilitäten anstatt absoluter Übereinstimmungen berichtet (s. Kap. 4.2).

3.4.2 Mehrebenenanalytische konfirmatorische Faktorenanalysen

Betrachtet man die Unterrichtswahrnehmungen von Schülerinnen und Schülern, so sind – wie in Kapitel 2 deutlich wurde – zwei verschiedene Analyseebenen zu unterscheiden, nämlich die Individual- und die Klassenebene. Die Individualebene entspricht dabei der subjektiven oder nicht-geteilten, die Klassenebene der objektiven oder geteilten Wahrnehmung des Unterrichts. Solche Daten erfordern Analysen mithilfe sogenannter *Mehrebenenmodelle* (vgl. z.B. Snijders & Bosker, 1999; Kreft & de Leeuw, 1998; Hox, 2002; Lüdtke & Köller, 2006). In *multilevel*-Modellen wird der Tatsache Rechnung getragen, dass Personen innerhalb eines *clusters* (z.B. einer Schulklasse) keine unabhängigen „Ziehungen" darstellen, da sie sich meist in verschiedener Hinsicht ähnlich sind (z.B. Leistungsniveau, Sozialstatus etc.). Diese Ähnlichkeit drückt sich z.B. in einer von Null verschiedenen ICC aus. Da im Falle der Unterrichtswahrnehmungen aus Schülersicht ja theoretisch erwartet wird, dass sich die individuellen Wahrnehmungen innerhalb von Klassen ähneln (dies wurde bereits im Zusammenhang mit der *Interrater*-Reliabilität diskutiert), sind solche Urteile statistisch gesehen als abhängig von der jeweiligen Klassenzugehörigkeit zu betrachten und entsprechend als hierarchisch geschachtelte Daten zu behandeln. In Mehrebenenanalysen wird darüber hinaus der Tatsache

Rechnung getragen, dass unterschiedliche Stichprobenumfänge berücksichtigt werden müssen, je nachdem, auf welcher Ebene die Analyse betrachtet wird (z.B. Anzahl der Schüler auf Individualebene, Anzahl der Klassen auf Klassenebene).

Da theoretische Konstrukte üblicherweise auf der Basis mehrerer Items erhoben werden, bieten sich zwei verschiedene Analysestrategien an: (1) Die Items werden zu Skalen zusammengefasst (z.B. durch Mittelwertbildung). (2) Im Rahmen von Strukturgleichungsmodellen (*structural equation modeling*, SEM) können die Items als Indikatoren latenter (also nicht direkt beobachtbarer) Konstrukte interpretiert werden. Der Vorteil von Strukturgleichungsmodellen liegt darin, dass z.B. Zusammenhänge zwischen (messfehlerfreien) latenten Konstrukten berechnet werden können.

Dies wird dadurch ermöglicht, dass – im Gegensatz zu einfachen Regressionsanalysen, wo (messfehlerbehaftete) individuelle Skalenkennwerte durch verschiedene Prädiktoren erklärt werden – hier die Kovarianzmatrix auf Itemebene betrachtet wird (Bollen, 1989). Ein theoretisches Modell – bestehend aus dem sogenannten *Messmodell*, das die latenten Variablen beschreibt, und einem *Strukturmodell*, das die Zusammenhänge zwischen den latenten Variablen definiert – kann dabei die empirische Kovarianzmatrix mehr oder weniger gut replizieren. Ist die Übereinstimmung zwischen empirischer und vom Modell implizierter Kovarianzmatrix hinreichend groß – zur Beurteilung dieser Frage wurden verschiedene Gütekriterien entwickelt –, dann wird ein theoretisch fundiertes Modell als adäquate Beschreibung der Zusammenhänge in der Population akzeptiert (vgl. dazu Modellgütekriterien; s.u.).

Neben der Kovarianzmatrix (auch als *covariance structure* bezeichnet) können auch Interzepte[28] (*mean structure*) in die Modellierung einbezogen werden (vgl. z.B. B. O. Muthén, 2002a; Hox, 2002; Heck, 2001). Dadurch lassen sich beispielsweise – identische Messmodelle vorausgesetzt (s.u.) – Mittelwerte latenter Variablen miteinander vergleichen.

Dieses Grundprinzip der Strukturgleichungsmodelle – dazu gehören auch die sogenannten konfirmatorischen Faktorenanalysen (KFA bzw. CFA im angelsächsischen Sprachraum) – lässt sich auch auf Anwendungen im mehrebenenanalytischen Kontext übertragen (vgl. Longford & Muthén, 1992; B. O. Muthén, 2002a; Mehta & Neale, 2005). Dazu werden die Daten im Zwei-Ebenen-Fall in zwei verschiedene Kovarianzmatrizen zerlegt: eine sogenannte *within-cluster*- und eine *between-cluster*-Matrix (vgl. B. O. Muthén, 1994).

[28] Mit Interzept wird der Achsenabschnitt einer Regressionsgleichung bezeichnet. Betrachtet man ein Faktormodell als Regression einer latenten Variable (im Sinne eines Prädiktors) auf den jeweiligen Indikator, dann ergeben sich für jeden Indikator ein Regressionsgewicht (als Faktorladung bezeichnet) und ein Achsenabschnitt (der Interzept).

Während für die *within*-Struktur die Kovarianzmatrix auf der Basis gruppenzentrierter Daten (*pooled within*-Kovarianzmatrix) verwendet werden kann, ist die Berechnung der *between*-Matrix *nicht* auf der Basis der Gruppenmittelwerte möglich. Dies lässt sich wie folgt verdeutlichen: Die Gruppenmittelwerte besitzen eine Reliabilität, die der ICC(2) (s. (6)) entspricht. (Die ICC(2) ist im Falle identischer Gruppengrößen[29] analog zu Cronbachs α zu interpretieren, das üblicherweise zur Überprüfung der Reliabilität von Skalen – der Aggregation verschiedener Items anstatt wie hier verschiedener Messungen desselben Items – verwendet wird.) Der unreliable Varianzanteil der Klassenmittelwerte ist *within*-Varianz, die möglicherweise mit der Residualvarianz anderer aggregierter Items korreliert ist – diese Korrelationen können der *pooled within*-Matrix entnommen werden –, was einer Korrelation der Residualvarianzen entspricht. Daraus ergeben sich folgende Komplikationen: (1) Wenn die *within*-Korrelation zweier Variablen gleich Null ist, dann unterschätzt die Korrelation der Mittelwerte – absolut betrachtet – den tatsächlichen Zusammenhang. (2) Wenn die *between*-Korrelation zweier Variablen gleich Null ist, dann wird der Zusammenhang der Mittelwerte in Richtung der *within*-Korrelation verzerrt. (3) Liegen sowohl *within*- als auch *between*-Korrelationen zweier Variablen vor, dann hängt die Korrelation der Mittelwerte von deren Reliabilitäten ab: Mit zunehmender Reliabilität verliert der verzerrende Einfluss der *within*-Korrelation an Bedeutung. Dies lässt sich anhand der folgenden Formel zur Bestimmung der Korrelation von (Klassen-)Mittelwerten verdeutlichen (Snijders & Bosker, 1999, S. 32):

$$\rho(M(x), M(y)) = \sqrt{ICC(2)_x \, ICC(2)_y}\,\rho_{between} + \sqrt{(1 - ICC(2)_x)(1 - ICC(2)_y)}\,\rho_{within} \qquad (7)$$

Die Korrelation zweier Klassenmittelwerte ρ(M(x), M(y)) ergibt sich aus dem Produkt des geometrischen Mittelwerts der Reliabilitäten der beiden Klassenmittelwerte ($ICC(2)_x$, $ICC(2)_y$) und der *between*-Korrelation ($\rho_{between}$) – und: dem Produkt des geometrischen Mittelwerts der Unreliabilitäten und der *within*-Korrelation (ρ_{within}). Diese Gleichung lässt sich auch durch verschiedene Umformungen (s. Anhang A) in folgende Gleichung transformieren[30]:

$$\mathrm{cov}(M(x), M(y)) = \mathrm{cov}(x, y)_{between} + \frac{1}{n} \cdot \mathrm{cov}(x, y)_{within} \qquad (8)$$

[29] Sowohl zur Berechnung der ICC(2) als auch zur Bestimmung der *between*-Kovarianzmatrizen liegen Formulierungen für unterschiedliche Gruppengrößen vor (s. B. O. Muthén, 1994; Snijders & Bosker, 1999).

[30] Diese Formulierung ist analog zu Muthén (1994). Dort wird allerdings die um die Klassengröße gewichtete (also mit dem Faktor n multiplizierte) *between*-Kovarianzmatrix verwendet.

Es ist allerdings problematisch, wenn innerhalb der Gruppen Subgruppen existieren, deren Urteile in gewissem Umfang übereinstimmen (also nicht voneinander unabhängig sind). In diesem Fall sind die Residuen *innerhalb* einer Aggregatvariable korreliert, was zu Überschätzungen der Reliabilität des Aggregats (vgl. Komaroff, 1997) und somit der Varianzanteile auf *between*-Ebene führt (Hutchison & Healy, 2001).

Auf der Basis der beiden ebenenspezifischen Kovarianzmatrizen lassen sich prinzipiell völlig unterschiedliche Modelle auf der Individual- und der Klassenebene spezifizieren. In den vorliegenden Analysen wird allerdings theoretisch erwartet, dass sich die Modelle auf Individual- und Klassenebene zumindest ähneln, z.B. was die Zahl der Faktoren und der entsprechenden Zuordnung von Indikatoren anbelangt. Dies bezeichnet Meredith (1993) als *configural invariance*. Weiterhin unterscheidet Meredith zwischen *weak factorial invariance* bzw. *pattern invariance* (identische Faktorladungen), *strong factorial invariance* (identische Ladungen, Interzepte) und *strict factorial invariance* (identische Ladungen, Interzepte, Residualvarianzen).

Die Residualvarianz (also die Varianz eines Indikators, die nicht durch den jeweiligen Faktor – bzw. in einem Modell mit Mehrfachladungen: durch die jeweiligen Faktor*en* – erklärt wird) setzt sich theoretisch aus der Varianz eines (Indikator-)*spezifischen Faktors* und der *Messfehlervarianz* zusammen (vgl. z.B. Meredith & Teresi, 2006; Lubke, Dolan, Kelderman & Mellenbergh, 2003). Unterschiede bezüglich der spezifischen Faktoren sowie der Messfehleranteile bei verschiedenen Gruppen beschränken die Vergleichbarkeit der jeweiligen Faktorvarianzen bzw. -mittelwerte.

Weak factorial invariance wird vorausgesetzt, wenn Faktorvarianzen (bzw. -kovarianzen) miteinander verglichen werden sollen. Das heißt, nur wenn identische Ladungen auf Individual- und Klassenebene vorliegen, können die Varianzen der latenten Variablen zur Berechnung latenter ICCs herangezogen werden (vgl. dazu auch Kap. 3.4.1). Gleiches gilt für den ebenenspezifischen fachübergreifenden Vergleich der Faktorvarianzen. Hier lassen sich darüber hinaus auch die nur auf Klassenebene existierenden Mittelwerte korrespondierender Faktoren vergleichen, wenn die Annahme identischer Interzepte (und idealerweise identischer Residualvarianzen[31] im Sinne der *strict factorial invariance*) nicht verworfen werden muss.

[31] Little (1997) geht davon aus, dass die Annahme der *strict factorial invariance* zu einer größeren Verzerrung (*bias*) führen kann als die *strong factorial invariance*. Werden identische Residualvarianzen vorausgesetzt, so gehen die diesbezüglichen (geringfügigen) Differenzen zwischen Gruppen in die übrigen Modellparameter ein und verfälschen somit die Ergebnisse der Analysen. Lubke und Muthén (2005) weisen jedoch darauf hin, dass bei der Überprüfung der Ladungs- und Interzept-Invarianz – *ohne* Berücksichtigung der Residualvarianzen – an-

Auf der Ebene innerhalb von Klassen hingegen existieren keine Interzepte, da diese per Definition alle auf Null gesetzt werden. Deshalb sind dort nur Ladungen und Residualvarianzen zu überprüfen.

Gelegentlich wird auch darauf hingewiesen, dass bereits partielle Invarianz der Messmodelle ausreicht, um Faktoren (Varianzen, Mittelwerte) miteinander vergleichen zu können (Byrne, Shavelson & Muthén, 1989; Reise, Widaman & Pugh, 1993). Reise et al. (ebd.) zufolge sollten in solchen Fällen allerdings die Ladungen eines Faktors (über verschiedene Gruppen – im Sinne von Subpopulationen – hinweg) mehrheitlich konstant sein. Dies ließe sich auch auf die Messmodelle auf unterschiedlichen Analyseebenen übertragen. Da in den vorliegenden Analysen allerdings nur relativ wenige Indikatoren pro Faktor zur Verfügung stehen, erscheint ein solches Verfahren hier nicht angemessen.

Es sei darauf hingewiesen, dass sich die hier zitierten Publikationen zur Überprüfung der Faktorinvarianz auf den Vergleich von einer relativ geringen Zahl von Gruppen (z.B. Geschlecht als Gruppierungsvariable) mit jeweils relativ großem Stichprobenumfang (z.B. N = 1000) beziehen. Auf Analysen von Schülerdaten aus (vielen) unterschiedlichen Klassen mit jeweils relativ kleinem Umfang sind solche Analysestrategien nur bedingt übertragbar. Hier lassen sich die einzelnen Gruppen (Klassen) nicht sinnvoll im Rahmen sogenannter *multigroup confirmatory factor*-Modelle bezüglich invarianter Faktoren untersuchen (vgl. Ansari, Jedidi & Dube, 2002). Bei mehrebenenanalytischen (konfirmatorischen) Faktorenanalysen wird üblicherweise ein für alle Gruppen gemeinsames Faktormodell auf der Basis der *within*-Kovarianzmatrix formuliert. Das heißt, es wird angenommen, dass für alle Gruppen ein gemeinsames Messmodell (hier: identische Ladungen und Residualvarianzen) für alle spezifizierten Faktoren gilt. Auf der Aggregatebene wird ein für alle Gruppen gemeinsames Messmodell auf der Basis der *between*-Kovarianzmatrix der Indikatoren – die wiederum auf die *random intercepts*[32] der Indikatoren zurückgeht, also den zwischen Klassen *variierenden* Interzepten, die latente Variablen repräsentieren – formuliert. Hier können zusätzlich die mittle-

---

hand der üblichen *Loglikelihood-Ratio*-Tests Unterschiede bezüglich der Mittelwerte spezifischer Faktoren unentdeckt bleiben können. Sie empfehlen deshalb – sofern keine besonderen theoretischen Annahmen unterschiedliche Residualvarianzen nahe legen – die Überprüfung der Messeigenschaften eines Instruments im Sinne der *strict factorial invariance*.
Liegen Verletzungen der Modellannahmen bezüglich der Skalierung der Indikatoren vor (z.B. ordinale Daten auf der Basis von Likert-Skalen), so ist bei großen Mittelwert- oder Streuungsdifferenzen der latenten Variablen – und den damit einhergehenden unterschiedlichen Verteilungen der Antwortkategorien der Indikatoren (Stichwort: Boden- bzw. Deckeneffekte), die unterschiedliche *grouping error*-Anteile (s.u.) besitzen – auch mit unterschiedlichen Residualvarianzen zu rechnen. Insofern wäre hier von einer Überprüfung der Residualvarianzen abzusehen.

[32] Damit verbunden sind auch die für Mehrebenenmodelle bekannten Annahmen wie z.B. multivariat normalverteilte Effekte innerhalb und zwischen Klassen sowie Homoskedastizität der Residualvarianzen.

ren Interzepte der Indikatoren in die Analysen einbezogen werden – z.B. bei fachübergreifenden Vergleichen hinsichtlich eines bestimmten Unterrichtsmerkmals.

Bei identischen Faktorladungen auf beiden Analyseebenen – und somit identischen Messmodellen – kann nun davon ausgegangen werden, dass (Faktor-)Mittelwertdifferenzen zwischen Klassen lediglich auf unterschiedliche Ausprägungen eines gleichförmigen Faktors in verschiedenen Klassen zurückgeführt werden können und nicht auf andere (variierende) Einflüsse auf Klassenebene (wie z.B. Mädchenanteil, Leistungsniveau der Klasse etc.) (vgl. Lubke et al., 2003). Die Schülerinnen und Schüler können in einem solchen Fall also theoretisch als zufällig auf die Klassen verteilte Beurteiler betrachtet werden.

Wie oben bereits erwähnt werden in Mehrebenenmodellen Unterschiede zwischen Klassen nur bezüglich der *mean structure* (unterschiedliche Interzepte), nicht aber bezüglich der *covariance structure* betrachtet (vgl. Ansari et al., 2002; Jedidi & Ansari, 2001). Es wird ein identisches Messmodell der *within*-Faktoren für alle Gruppen angenommen. Ist diese Voraussetzung nicht erfüllt – und dies wird hier im Gegensatz zu den *multigroup confirmatory factor*-Modellen (üblicherweise) nicht überprüft[33] –, so wäre es theoretisch möglich, dass beispielsweise Interaktionen von unterschiedlichen Ladungen und unterschiedlichen Interzepten auf Klassenebene zu identischen Ladungen über die Ebenen hinweg führen. Inwiefern ein solches Szenario eine realistische Fehlerquelle darstellen könnte, wurde meines Wissens bisher noch nicht untersucht.

Unterschiedliche Faktorladungen zwischen analogen Konstrukten innerhalb von Klassen und zwischen Klassen hingegen weisen auf eine nicht-identische (sondern nur analoge) Bedeutung der Faktoren auf beiden Ebenen hin. Hier können die Schülerinnen und Schüler nicht als theoretisch unabhängige Beobachter betrachtet werden. Offensichtlich spielen in solchen Fällen bestimmte Merkmale der Klassenkomposition eine Rolle bei der Beurteilung des jeweiligen Unterrichtsmerkmals.

Ein besonderes Problem bei der Anwendung von Strukturgleichungsmodellen besteht darin, dass häufig die zugrunde liegenden Daten bestimmte Voraussetzungen nicht (oder nur unzureichend) erfüllen. Hier ist insbesondere die Voraussetzung intervallskalierter Indikatoren zu nennen (mittlerweile stehen auch Verfahren für nicht-intervallskalierte Indikatoren zur Verfügung, s.u.). Da in der Surveyforschung in Fragebögen meist Likert-Antwortskalen ein-

[33] Eine solche Überprüfung ist auch aufgrund der Vielzahl der Klassen und der geringen Klassengröße nicht trivial. Diese Annahmen ließen sich beispielsweise anhand von Mischverteilungsmodellen mit latenter Gruppierungsvariable auf Klassenebene (d.h. nur komplette Schulklassen werden einer Gruppe zugeordnet) mit unterschiedlichen *within*-Faktoren (Ladungen, Residualvarianzen) und -Faktorvarianzen für die jeweiligen Gruppen überprüfen.

gesetzt werden (so auch in DESI), sind solche Daten in der Regel nur ordinalskaliert. Dieses Problem wird oft dadurch umgangen, dass den einzelnen Antwortalternativen arbiträre Zahlen (z.B. von 1 bis 4 bei 4-stufigem Antwortformat) zugewiesen werden und diese anschließend als intervallskalierte Daten in Strukturgleichungsmodellen behandelt werden.

Sind bestimmte Voraussetzungen erfüllt, dann führt ein solches Verfahren in der Regel zu relativ robusten Ergebnissen. Zu diesen Voraussetzungen gehören insbesondere die univariate (und multivariate[34]) Normalverteilung der Indikatoren (vgl. z.B. West, Finch & Curran, 1995). In einer Simulationsstudie von Muthén und Kaplan (1985) zeigte sich, dass nur geringe Verzerrungen bezüglich der Parameter, Standardfehler und Chi^2-Statistiken (die z.B. zum Vergleich verschiedener Modelle herangezogen werden) zu erwarten sind, wenn die univariate Schiefe (*skewness*) und der Exzess (*kurtosis*) der meisten Variablen im Bereich von -1.0 bis +1.0 liegen. Dabei spielten in ihren Analysen die Zahl der Items und die Anzahl der Antwortalternativen kaum eine Rolle. In einer Folgestudie von Muthén und Kaplan (1992) ergaben sich unter solchen Bedingungen auch bei komplexeren Modellen keine Verzerrungen hinsichtlich der bereits genannten Kennwerte.

Außerdem stehen auch verschiedene Schätzverfahren zur Verfügung, die robust gegen Verletzungen der Normalverteilung sind, wie etwa die *SCALED* Chi^2-Statistik oder *Bootstrapping*-Verfahren (West et al., 1995). Ein solches robustes Schätzverfahren steht z.B. in *Mplus 4.1* (L. K. Muthén & Muthén, 2006) auch für Zwei-Ebenen-Modelle unter der Bezeichnung *MLR*[35] zur Verfügung. Zum Vergleich sogenannter geschachtelter Modelle – also Modelle, die durch Hinzufügung von Restriktionen aus weniger restriktiven Modellen hervorgehen (z.B. durch Gleichsetzung von Ladungen, Fixierung von Korrelationen o.ä.) – steht in Mplus (ab Version 4) neben dem *Likelihood Ratio* Test (bzw. Chi^2-Differenz-Test) auch der sogenannte Wald-Test (mithilfe des MODEL TEST-Kommandos) zur Verfügung. Der Wald-Test und der *Likelihood Ratio* Test (sowie der *Lagrangian Multiplier* Test) können als asymptotisch äquivalente Teststatistiken betrachtet werden (vgl. Bollen, 1989, S. 293ff.). Unter der Bedingung nicht-normalverteilter Variablen liefert Mplus auch einen Korrekturfaktor für die *Loglikelihood*- bzw. Chi^2-Werte mit dessen Hilfe z.B. der sogenannte *Satorra-Bentler scaled*

[34] Die Überprüfung der multivariaten Normalverteilungsannahme ist überaus schwierig (Kline, 2005, S. 48f.): Mardia (1970) hat dazu ein Verfahren vorgeschlagen, das auch in Mplus implementiert ist. Dieses Verfahren ist allerdings einerseits mit erheblichem Rechenaufwand verbunden, andererseits werden damit nur z-Werte bestimmt, die bei großen Stichproben fast immer signifikant sind – auch bei relativ trivialen Abweichungen. Im Regelfall werden solche Verletzungen der *multivariaten* Normalverteilungsannahme durch Inspektion der *univariaten* Verteilungen aufgedeckt. Whittaker und Stapleton (2006) weisen auch darauf hin, dass Verletzungen der Voraussetzung der *multivariaten* Normalverteilung wenig problematisch sind, da konsistente Schätzer im Rahmen von *Maximum Likelihood*-Schätzverfahren zu erwarten sind.

[35] "MLR– maximum likelihood parameter estimates with standard errors and a chi-square test statistic (when applicable) that are robust to non-normality and non-independence of observations when used with TYPE = COMPLEX" (L. K. Muthén & Muthén, 2006, S. 426).

chi-square difference Test sowie adjustierte Wald-Tests berechnet werden können (vgl. Satorra, 2000). Da die skalierten Chi^2- und Loglikelihood-Statistiken gelegentlich zu statistisch widersinnigen Ergebnissen führen (z.B. dass das restriktivere Modell die Daten besser erklärt als das weniger restriktive) empfiehlt B. Muthén die Verwendung des adjustierten Wald-Tests[36].

Theoretisch angemessenere Modelle für dichotome und polytome (ordinale und nominale) Indikatoren latenter Variablen wurden im Rahmen der sogenannten *Item-Response-Theory* (IRT; auch *Probabilistische Testtheorie* genannt) entwickelt (vgl. z.B. Rost, 2004; Steyer & Eid, 2001; Cavanagh & Romanoski, 2006). Im Gegensatz zur *Klassischen Testtheorie* (KTT), die einen linearen Zusammenhang zwischen intervallskalierten Indikatoren und latenten Variablen (z.B. Konstrukten) annimmt, wird im Rahmen der IRT mit zunehmender Ausprägung der latenten Variable eine zunehmende *Lösungswahrscheinlichkeit*[37] bezüglich eines Items postuliert. Dazu wird beispielsweise im Rasch-Modell das Verhältnis der bedingten Wahrscheinlichkeit, dass Item i gelöst bzw. nicht gelöst wird – jeweils unter der Bedingung einer bestimmten Fähigkeit einer Person (oder allgemeiner formuliert: einer bestimmten Position auf einer latenten Dimension) – betrachtet. Diese als Wettquotient (engl.: *odds*) bezeichnete Relation (Lösungswahrscheinlichkeit geteilt durch Gegenwahrscheinlichkeit) ist auf den Wertebereich 0 bis $+\infty$ beschränkt (wobei die „Mitte" – also Wahrscheinlichkeit gleich Gegenwahrscheinlichkeit – bei 1 liegt). Durch Transformation des Wettquotienten mithilfe des natürlichen Logarithmus erhält man den logarithmierten Wettquotienten, der im Wertebereich von $-\infty$ bis $+\infty$ definiert ist und eine symmetrische Projektion des Wettquotienten darstellt.

Für das ursprünglich nur für dichotome Items ausgelegte Rasch-Modell wurden verschiedene Erweiterungen für ordinale Items vorgeschlagen (vgl. dazu z.B. Ostini & Nering, 2006). Die bekanntesten darunter sind das *Rating-Scale-Modell* (RSM; Andrich, 1978), das *Partial-Credit-Modell* (PCM; Masters, 1982) sowie das *Graded-Response-Modell* (GRM; Samejima, 1969). Im RSM werden die *Abstände* zwischen den Schwellenparametern bzw. *thresholds* – damit werden die Stellen auf dem Kontinuum des latenten Merkmals bezeichnet, an denen die Wahrscheinlichkeit zweier benachbarter Antwortkategorien[38] gleich groß ist, also jeweils 50% – bei allen Items als identisch angenommen. Diese Annahme kann beispielsweise dann plausibel sein, wenn verschiedene Items einer Skala mithilfe desselben Likert-Antwortformats

[36] Vgl. dazu die Diskussion im Mplus-Forum unter http://www.statmodel.com/discussion/messages/9/189.html?1155247035.

[37] Da die IRT vor allem in der Testkonstruktion verbreitet ist, wird dieser Begriff häufig verwendet, obwohl er in anderen Kontexten – z.B. bei Einstellungsskalen – wenig angebracht erscheint.

[38] Bei einem vierstufigen Antwortformat (kodiert von 0 bis 3) gibt es drei Schwellenparameter: Der erste bezeichnet die Schwelle zwischen Kategorie 0 und Kategorie 1, der zweite die zwischen Kategorie 1 und 2 und der dritte die Schwelle zwischen den Kategorien 2 und 3.

erhoben wurden. Es kann jedoch auch in solchen Fällen nicht grundsätzlich davon ausgegangen werden, dass keine Interaktion zwischen den Items und den Abständen zwischen den Schwellenparametern (diese Abstände werden auch *step*-Parameter[39] genannt) besteht. Solche Item-*step*-Interaktionen werden im PCM berücksichtigt, indem itemspezifische *step*-Parameter geschätzt werden.

Sowohl im PCM als auch im RSM (das durch Gleichsetzung der *step*-Parameter aller Items aus dem PCM hervorgeht) sind sogenannte Schwellenvertauschungen möglich. Von Schwellenvertauschungen spricht man, wenn die einzelnen *thresholds* nicht in aufsteigender Reihenfolge geordnet sind – wenn also beispielsweise der Schwellenparameter für die Antwortkategorien 1 und 2 auf der latenten Dimension unterhalb des Schwellenparameters für die Kategorien 0 und 1 liegt. Solche Vertauschungen sind im GRM nicht möglich, da diese durch modellinhärente Restriktionen ausgeschlossen werden (vgl. dazu Tuerlinckx & Wang, 2004). Schwellenvertauschungen sind jedoch nur dann problematisch, wenn diese aufgrund theoretischer Annahmen unplausibel sind, was häufig nicht der Fall ist (ebd.). Insofern könnte man das GRM-Modell häufig als zu restriktiv betrachten.

Ein weiterer grundlegender Unterschied zwischen dem GRM und dem RSM bzw. dem PCM besteht darin, dass in den beiden letztgenannten Modellen nur die Wahrscheinlichkeiten der beiden jeweils benachbarten Antwortkategorien zur Schätzung der *thresholds* berücksichtigt werden (eine Antwort in Kategorie 3 ist also nur relevant für die Bestimmung der Schwellenparameter zwischen Kategorie 2 und 3 bzw. zwischen Kategorie 3 und 4), während im GRM jeweils alle Schwellenparameter von allen Kategorienwahrscheinlichkeiten abhängen (vgl. Ostini & Nering, 2006; Tuerlinckx & Wang, 2004). Daneben werden üblicherweise im GRM unterschiedliche Diskriminationsparameter (in der Terminologie der Faktorenanalyse: Ladungen) geschätzt (Ostini & Nering, 2006), während diese im PCM und im RSM für alle Items konstant gehalten werden (theoretisch lassen sich jedoch auch im PCM und im RSM unterschiedliche Diskriminationsparameter schätzen).

Solche IRT-Modelle lassen sich auch in den Kontext von Strukturgleichungsmodellen einbetten. Üblicherweise wird dazu eine, der beobachteten ordinalen Variable zugrunde liegende (latente), intervallskalierte Variable eingeführt. Beide Variablen sind über ein sogenanntes *threshold*-Modell miteinander verbunden (vgl. dazu De Boeck & Wilson, 2004; Grilli & Rampichini, 2007). Die intervallskalierte latente Variable wird dabei als ursächlich für die

[39] Bei einem vierstufigen Antwortformat (kodiert von 0 bis 3) gibt es zwei *step*-Paramter aus denen die drei Schwellenparameter berechnet werden können: Die ersten beiden Schwellenparameter entsprechen der Summe der Itemschwierigkeit und dem jeweiligen *step*-Parameter. Der letzte Schwellenparamter ergibt sich aus der Differenz der Itemschwierigkeit und der Summe der *step*-Parameter, da die Itemschwierigkeit als Mittelwert der Schwellenparameter definiert ist.

beobachtete ordinale Variable betrachtet. Das heißt, die Antwortwahrscheinlichkeiten in den einzelnen Kategorien sind bedingt durch die zugrunde liegende latente Variable. Das *threshold*-Modell besteht aus einer *link*-Funktion (also z.B. dem Probit- oder Logit-*link*[40]) und einem Modell für ordinale Antworten wie z.B. dem GRM oder PCM (vgl. z.B. Tuerlinckx & Wang, 2004).

Die so modellierte intervallskalierte latente Variable kann dann als Indikator eines latenten Merkmals in konfirmatorischen Faktorenanalysen behandelt werden. Dies gilt auch prinzipiell für mehrebenenanalytische Faktorenanalysen[41] (vgl. Grilli & Rampichini, 2007). Solche Modelle sind allerdings mit sehr großem Rechenaufwand verbunden, da sie (bisher) nur über numerische Integrationsverfahren berechnet werden können, so dass derzeit die Obergrenze für solche Modelle wohl bei zwei Faktoren pro Ebene in einem Zwei-Ebenen-Modell liegt. Und auch dies ist nur mit Einschränkungen möglich: Die Residualvarianzen der Indikatoren auf Ebene 2 müssen in der Regel auf Null fixiert werden[42]. In vielen Fällen ist diese Annahme möglicherweise unproblematisch, da – je nach (durchschnittlicher) Gruppengröße – die Indikatoren auf Ebene 2 oft sehr reliabel sind. Problematisch ist allerdings, dass diese Annahme kaum überprüft werden kann.

Die weitaus umfangreicheren Modelle in der vorliegenden Untersuchung sind (derzeit) nur unter der theoretisch suboptimalen Annahme metrischer (intervallskalierter) Indikatoren möglich. Da bisher meines Wissens keine Ergebnisse aus Simulationsstudien zu Zwei-Ebenen-Faktorenanalysen publiziert wurden (die oben genannten Ergebnisse zu nicht-normalverteilten Indikatoren beziehen sich sämtlich auf nicht-hierarchische Modelle), wird in Kapitel 3.4.3 eine eigens durchgeführte Simulationstudie kurz skizziert. Ziel dieser Simulationsstudie ist die Überprüfung des Ausmaßes der Verzerrung der Parameter sowie der Angemessenheit verschiedener Modellgütekriterien, die für intervallskalierte Indikatoren zur Verfügung stehen (nicht jedoch für IRT-Modelle!).

Was die Verzerrung der Populations-Schätzer anbelangt, so soll zusätzlich ein Verfahren zur Adjustierung überprüft werden. Dieses Verfahren basiert auf Überlegungen von O'Brien (1985) zum Zusammenhang von ordinalen Messungen (z.B. Likert-Skalen) mit theoretisch postulierten latenten Variablen, die den ordinalen Daten zugrunde liegen. O'Brien konnte für

[40] Die Wahl der *link*-Funktion hat einen Einfluss auf die Verteilung der itemspezifischen Fehlervarianz: Der Probit-*link* steht für die Annahme normalverteilter Residuen, der Logit-*link* für eine logistische Verteilung der Fehlerterme (vgl. Grilli & Rampichini, 2007).

[41] Ein anderes weit verbreitetes Verfahren zur Behandlung ordinal skalierter Variablen z.B. im Rahmen von Strukturgleichungsmodellen basiert auf der Berechnung polychorischer Korrelationskoeffizienten (vgl. z.B. B. O. Muthén, 1984; Jöreskog, 1994). Dieser Ansatz wurde aber meines Wissens bisher nicht auf mehrebenenanalytische Verfahren übertragen.

[42] Da jede Residualvarianz auf Ebene 2 eine eigene Integrationsdimension darstellt, steigt die Zahl der erforderlichen Integrationspunkte exponentiell.

zwei theoretische Verteilungen der latenten Variable (Normalverteilung bzw. Gleichverteilung) zeigen, dass der Zusammenhang zwischen latenter und beobachteter (ordinaler) Variable analytisch bestimmbar ist.

Bei der Umwandlung von intervall- in ordinalskalierte Variablen unterscheidet O'Brien zwischen zwei verschiedenen Fehlerarten: (1) Messfehler, die einzig auf die Unterteilung des latenten Kontinuums in eine begrenzte Anzahl von Abschnitten zurückführbar sind (*pure categorization errors*). (2) Messfehler, die aufgrund der Zuweisung arbiträrer Zahlen (z.B. die Zahlen eins bis vier bei einer vierstufigen Likert-Skala) für die einzelnen Kategorien entstehen (*pure transformation errors*). Die Kombination aus diesen beiden voneinander unabhängigen (also unkorrelierten) Messfehlern nennt O'Brien *grouping error*. Interessanterweise sind diese Messfehler wiederum unabhängig von der Zufallsfehlerkomponente der latenten Variable.

Mithilfe des von O'Brien vorgeschlagenen Verfahrens lassen sich die *pure categorization error*- und die *pure transformation error*-Komponenten – und somit die *grouping error*-Komponente – bestimmen. Im Folgenden wird dieses Vorgehen für normalverteilte latente Variablen kurz dargestellt.

Um die erste Fehlerkomponente zu bestimmen (*pure categorization error*) wird auf ein von Guilford (1954) vorgeschlagenes Verfahren zurückgegriffen, mit dessen Hilfe sich im Sinne einer Minimierung der Abweichungsquadrate optimale *scores* für die einzelnen Kategorien einer ordinalskalierten Variable (mit n Kategorien und zugrunde liegender normalverteilter Variable) bestimmen lassen (ein Beispiel ist in Tabelle 3 dargestellt). Auf der Basis der kumulierten Kategorienwahrscheinlichkeiten ($pcum_1$ bis $pcum_n$) – die sich aus den relativen Kategorienhäufigkeiten (Wahrscheinlichkeiten p_1 bis p_n der Kategorien) ergeben – werden dazu zunächst die Grenzen g_i (*boundaries*) zwischen den n Kategorien bestimmt. Diese entsprechen den Abschnitten der Normalverteilung, die jeweils $pcum_i$ Prozent der Fälle einschließen (bei der Standardnormalverteilung[43] entspricht eine kumulierte Wahrscheinlichkeit von $p = .975$ beispielsweise einem z-Wert von 1.96[44]).

Anhand der Grenzen lassen sich die jeweiligen Ordinaten[45] y_i (also die Dichte der Normalverteilung an der jeweiligen Stelle) bestimmen. Die optimalen *scores* ergeben sich aus der

[43] Hierbei werden implizit der Mittelwert und die Varianz der latenten intervallskalierten Variable definiert: M = 0, Var = 1.
[44] Mithilfe des Statistikpakets SAS (SAS Institute Inc., 2006) beispielsweise können solche z-Werte anhand der QUANTILE -Funktion bestimmt werden: z = QUANTILE('NORMAL', .975).
[45] Dazu kann beispielsweise in SAS die Funtion PDF verwendet werden: y = PDF('NORMAL', z)

Differenz jeweils zweier benachbarter Ordinaten geteilt durch die jeweilige Kategorienwahrscheinlichkeit[46]:

$$o_i = \frac{y_{i-1} - y_i}{p_i} \tag{9}$$

Die Standardabweichung[47] s_k dieser optimalen *scores* entspricht nun der Korrelation r_{ku} der kategorisierten Variable (optimale *scores*) mit der zugrunde liegenden unkategorisierten normalverteilten Variable (Peters & Van Voorhis, 1940; O'Brien, 1985):

$$s_k = r_{ku} = \sqrt{\sum_{i=1}^{n} p_i o_i^2} \tag{10}$$

Der relative Fehleranteil e_{ku}, den O'Brien als *pure categorization error* bezeichnet, ergibt sich wie folgt:

$$e_{ku} = 1 - r_{ku}^2 \tag{11}$$

Die Kovarianz[48] cov_{ka} zwischen den optimalen Kategorien*scores* o_i und den arbiträren *scores* a_i lässt sich folgendermaßen ermitteln (vgl. dazu auch Minium, 1970; O'Brien, 1985):

$$\text{cov}_{ka} = \sum_{i=1}^{n} p_i a_i o_i \tag{12}$$

---

[46] Bei der ersten Kategorie wird die untere Grenze $g_0 \rightarrow -\infty$ und die entsprechende Ordinate $y_0 = 0$ gesetzt. Die Grenze der letzten Kategorie approximiert $+\infty$ (kumulierte Kategorienwahrscheinlichkeit $pcum_n = 1$). Entsprechend wird auch die Ordinate $y_n = 0$ gesetzt.

[47] Da der Mittelwert der optimalen *scores* Null beträgt, lässt sich die Gleichung zur Berechnung der Varianz bzw. Standardabweichung hier vereinfacht darstellen. Beweis:
M(optimale *scores*) = $p_1[(y_0 - y_1) / p_1] + p_2[(y_1 - y_2) / p_2] \ldots p_n[(y_{n-1} - y_n) / p_n] =$
$(y_0 - y_1) + (y_1 - y_2) + \ldots + (y_{n-1} - y_n) = y_0 - y_n = 0$ (da $y_0 = y_n = 0$; vgl. Fußnote 46)

[48] Die vereinfachte Darstellung der Kovarianz hier lässt sich wie folgt begründen (vgl. Diehl & Kohr, 1994, S. 157):
$Cov(x, y) = \sum(x_i - M(x))(y_i - M(y)) = \sum(x_i y_i) - M(x)M(y)$
Da der Mittelwert der optimalen *scores* gleich Null ist, reduziert sich die Kovarianz hier um den letzten Term. Eine Anpassung an den Stichprobenumfang (n-1) ist hier nicht erforderlich, da dieser Term bei der Berechnung der Korrelation – die hier die entscheidende Größe darstellt – ohnehin wegfällt (vgl. Diehl & Kohr, 1994, S. 155).

Die entsprechende Korrelation r_{ka} ist definiert als die Kovarianz der optimalen und der arbiträren Kategorien*scores*, die anhand der Standardabweichungen der optimalen (s_k) und der arbiträren[49] (s_a; vgl. (13)) *scores* standardisiert wird (s. (14)):

$$s_a = \sqrt{\sum_{i=1}^{n} p_i (a_i - M(a))^2} \tag{13}$$

$$r_{ka} = \frac{\mathrm{cov}_{ka}}{s_k s_a} \tag{14}$$

Der von O'Brien (1985) als *pure transformation error* bezeichnete relative Varianzanteil[50] entspricht somit:

$$e_t = r_{ku}^2 (1 - r_{ka}^2) \tag{15}$$

Aus den beiden oben genannten voneinander unabhängigen Korrelationen r_{ku} und r_{ka} lässt sich die Korrelation der arbiträren *scores* mit der zugrunde liegenden normalverteilten Variable wie folgt berechnen:

$$r_g = r_{ku} r_{ka} \tag{16}$$

Der Gesamtanteil an Fehlervarianz, die auf die Umwandlung einer intervallskalierten, normalverteilten Variable in eine ordinale Variable mit n Kategorien, den Kategorienwahrscheinlichkeiten p_i und arbiträrem *scoring* zurückführbar ist, entspricht:

$$e_g = 1 - r_g^2 = e_{ku} + e_t \tag{17}$$

Anhang B enthält ein vom Autor dieser Untersuchung entwickeltes SAS-Macro, mit dessen Hilfe sich die oben genannten Korrektur-Faktoren auf der Basis der empirischen Häufig-

[49] M(a) in (13) entspricht dem arithmetischen Mittelwert der kategorisierten Variable mit arbiträren Kategorien*scores*.

[50] Da sich der relative Varianzanteil 1 - r_{ka}^2 auf die Varianz der optimalen *scores* bezieht – und nicht auf die Varianz der latenten intervallskalierten Variable – muss dieser Varianzanteil anhand des relativen Varianzanteils r_{ku}^2 „skaliert" werden. Somit addieren sich die Fehlervarianzanteile e_{ku} und e_t zum relativen Gesamtfehlervarianzanteil e_g.

keiten der Kategorien*scores* verschiedener Variablen berechnen lassen. Damit lässt sich der Aufwand solcher Korrekturen auf ein Minimum beschränken.

Für das Beispiel aus Tabelle 3 ergeben sich die in Tabelle 4 aufgeführten statistischen Kennwerte. Es zeigt sich, dass hier – verglichen mit der Transformation der Kategorien zu arbiträren Werten – ein deutlich größerer Fehlervarianzanteil (knapp 17%) aufgrund der Kategorisierung (Einteilung der normalverteilten Variable in vier Kategorien mit den entsprechenden Kategorienhäufigkeiten) entsteht. Das heißt, eine Transformation der arbiträren *scores* würde die Korrelation mit der zugrunde liegenden intervallskalierten Variable kaum erhöhen (von r = .910 auf maximal r = .913).

Tabelle 3: Beispiel zur Bestimmung der optimalen Scores für eine vierstufig ordinalskalierte Variable bei zugrunde liegender intervallskalierter, normalverteilter Variable

	Kategorie i				
	(0)	1	2	3	4
arbiträrer *score* a_i		1	2	3	4
Kategorienwahrscheinlichkeit p_i		0.10	0.20	0.30	0.40
kumulierte Kategorienwahrscheinlichkeit $pcum_i$		0.10	0.30	0.60	1.00
Grenze g_i	$-\infty$	-1.282	-0.524	0.253	$+\infty$
Ordinate y_i	0.000	0.175	0.348	0.386	0.000
optimaler *score* o_i		-1.755	-0.861	-0.129	0.966

Tabelle 4: Statistische Kennwerte der arbiträren bzw. optimalen *scores* einer vierstufig ordinalskalierten Variable (Kategorienwahrscheinlichkeiten: $p_1 = .10$, $p_2 = .20$, $p_3 = .30$, $p_4 = .40$) bei zugrunde liegender intervallskalierter, normalverteilter Variable

Kennwert	Wert
Mittelwert der arbiträren *scores* M(a)	3.0
Standardabweichung der arbiträren *scores* s_a	1.0
Standardabweichung der optimalen *scores* s_o (entspricht Korrelation r_{ku} der optimalen *scores* mit zugrunde liegender intervallskalierter Variable)	0.913
Korrelation optimaler und arbiträrer *scores* r_{ka}	0.996
Korrelation der arbiträren *scores* mit zugrunde liegender intervallskalierter Variable r_g	0.910
Anteil Fehlervarianzen	
categorization error (e_{ku})	0.166
transformation error (e_t)	0.007
grouping error (e_g)	0.173

Übertragen auf konfirmatorische Faktorenanalysen für intervallskalierte Indikatoren bedeuten diese Ergebnisse Folgendes: Wenn der Anteil der Varianz eines Indikators, der lediglich auf die Umwandlung einer latenten intervallskalierten Variable in eine ordinale Variable mit arbiträren Kategorien*scores* (*grouping error*), analytisch bestimmbar ist, dann lässt sich auch (mit gewissen Einschränkungen, s.u.) der tatsächliche relative Anteil der *true-score*-

($\lambda^2_{adj,std}$; vgl. (18)) und *error*-Varianz[51] ($e_{res,adj,std}$; vgl. (19)) eines Indikators ermitteln, also die (Un-)Reliabilität, indem die auf der Basis von kategorisierten Indikatoren ermittelten (Un-) Reliabilitäten nur auf den um die *grouping error*-Komponente bereinigten relativen Varianzanteil des Indikators bezogen werden.

$$\lambda^2_{adj,std} = \frac{\lambda^2_{std}}{r_g^2} = \frac{1 - e_{res,adj,std}}{r_g^2} \tag{18}$$

$$e_{res,adj,std} = 1 - \lambda^2_{adj,std} \tag{19}$$

Da die *grouping error*-Komponente als *random error* betrachtet werden kann, der bei der Aggregation von Daten eliminiert wird, sollten bei mehrebenenanalytischen KFA-Modellen die Varianzkomponenten auf Aggregatebene – zumindest weitgehend – unverzerrt sein[52]. Das heißt, die gesamte *grouping error*-Varianz verbleibt auf Ebene 1. Entsprechend muss der relative Anteil der *grouping error*-Komponente auf Ebene 1, e_{gw}, wie folgt auf die *within*-Varianz adjustiert werden:

$$e_{gw} = \frac{1 - r_g^2}{1 - ICC} \tag{20}$$

Der relative, um die *grouping error*-Komponente bereinigte Varianzanteil der kategorisierten Variable auf Ebene 1, r^2_{gw}, entspricht:

$$r^2_{gw} = 1 - e_{gw} = 1 - \frac{1 - r_g^2}{1 - ICC} \tag{21}$$

[51] Dies ist deshalb möglich, weil die Fehlerkomponente der zugrunde liegenden intervallskalierten Variable unkorreliert mit der *grouping-error*-Komponente ist.

[52] Da bei mehrebenenanalytischen Verfahren mit intervallskalierten Variablen (meist) normalverteilte Residuen (auf beiden bzw. allen Ebenen) vorausgesetzt werden (vgl. z.B. Snijders & Bosker, 1999), ist hier jedoch – auch bei der Verwendung robuster Schätzverfahren, die lediglich zu anderen Standardfehlerschätzungen führen – mit einem gewissen *bias* bezüglich der Schätzer auf Aggregatebene zu rechnen. Auf der Aggregatebene ist die Normalverteilungsannahme bezüglich der Residuen wohl weniger problematisch, da aufgrund des zentralen Grenzwertsatzes auch bei von der Normalverteilung abweichenden Werteverteilungen Mittelwerte zur Normalform streben (vgl. z.B. Diehl & Arbinger, 1990). Allerdings ist diese Tendenz abhängig von der jeweiligen Gruppengröße und der Art der Verteilung.

Die ICC, die dem (auf die Gesamtvarianz bezogenen) relativen Varianzanteil auf Aggregatebene entspricht, kann anhand der um die *grouping error*-Komponente verminderten Gesamtvarianz adjustiert werden[53]:

$$ICC_{adj} = \frac{ICC}{1 - e_g} = \frac{ICC}{r_g^2} \tag{22}$$

Es sei hier noch auf zwei Einschränkungen dieses Adjustierungsverfahrens hingewiesen: (1) Obwohl der Zusammenhang zwischen einer kategorisierten Variable und ihrer zugrunde liegenden intervallskalierten, normalverteilten Variable analytisch bestimmt werden kann, ist der wahre Zusammenhang zwischen zwei (oder mehreren) kategorisierten Variablen *nicht* mithilfe des von O'Brien vorgeschlagenen Verfahrens bestimmbar. Dies lässt sich wie folgt verdeutlichen: Wenn zwei intervallskalierte Variablen perfekt korreliert sind (r = 1), dann sind auch die identisch kategorisierten Variablen – trotz *grouping error* – perfekt korreliert. Anders formuliert: In diesem Fall sind auch die Fehlerkomponenten des *grouping errors* perfekt korreliert (anstatt unabhängig voneinander). Mit steigenden Korrelationen ist also damit zu rechnen, dass auch die *grouping error*-Komponenten zunehmend miteinander korreliert sind. Dies führt dann bei den oben vorgeschlagenen Korrekturen zu einer Überadjustierung im Sinne einer Überschätzung der Koeffizienten[54]. (2) Bei der Schätzung der Koeffizienten auf Aggregatebene ist bei (starken) Verletzungen der Normalverteilungsannahme mit einem gewissen *bias* zu rechnen, was ebenfalls tendenziell zu einer Überschätzung der tatsächlichen Koeffizienten führen sollte. Insofern ist eine Überprüfung des Verfahrens auf der Basis einer Simulationsstudie angebracht.

Die beiden Einschränkungen gelten allerdings nicht nur für dieses Adjustierungsverfahren, sondern generell für die Verwendung kategorisierter Variablen in Modellen, die für intervallskalierte Variablen ausgelegt sind. Insbesondere das erstgenannte Problem bezüglich der Bestimmung der wahren Korrelation zweier kategorisierter Variablen stellt eine generelle Einschränkung auch für die Verwendung von – v.a. aus wenigen Items mit geringer Anzahl

[53] Aus (22) wird auch ersichtlich, dass ein völlig anderer Ansatz zur Ermittlung eines Schätzers für die Fehlervarianzanteile, die durch „Inkompatibilität" der Messmodelle (ordinal vs. metrisch) entstehen, denkbar ist, und zwar über die Bestimmung des Verhältnisses der metrischen und der ordinalen ICCs:

$$\frac{ICC_{metrisch}}{ICC_{ordinal}} = \frac{\frac{\tau^2}{\tau^2 + \sigma^2 + e_g}}{\frac{\tau^2}{\tau^2 + \sigma^2}} = \frac{\tau^2 + \sigma^2}{\tau^2 + \sigma^2 + e_g} = \frac{s^2_{true,total}}{s^2_{total}}$$

[54] Hier sind auch Überschreitungen des theoretischen Bereichs der Koeffizienten möglich.

an Antwortkategorien bestehenden – Skalen dar, unabhängig davon, welche Skalierungsmethode verwendet wurde (Summen- bzw. Mittelwertbildung, Faktor*score*, Rasch-Skalierung etc.), da bei allen diesen Verfahren einer Skala nur eine bestimmte Anzahl unterschiedlicher Werte zugeordnet wird. Das heißt, im Prinzip stellen auch (kurze) Skalen kategorisierte „Versionen" der zugrunde liegenden latenten Variable (inklusive Messfehler) dar. Dies kann – besonders bei hohen Zusammenhängen zwischen den latenten Variablen – teilweise dazu führen, dass Korrelationen zwischen manifesten *scores* die analogen latenten Korrelationen absolut betrachtet überschreiten, obwohl die manifesten *scores* messfehlerbehaftet sind und daher eher „nach oben" korrigiert werden sollten (Stichwort: *correcting for attenuation*).

Wie bereits oben erwähnt, sollte in der Simulationsstudie – neben den (adjustierten) Parameterschätzungen – auch die Angemessenheit verschiedener Modellgüte-Indizes überprüft werden. Diese Modell-Fit-Indizes lassen sich grob in zwei Klassen unterteilen (Bollen, 1989; Hu & Bentler, 1995): (1) Absolute Fit-Indizes geben Auskunft darüber, wie gut das apriorische Modell die empirischen Daten aus der Stichprobe reproduziert. Im Gegensatz dazu vergleichen (2) inkrementelle Fit-Indizes das theoretische Modell mit einem restriktiveren *baseline*-Modell.

Eines der bekanntesten absoluten Modellgütekriterien ist die Chi^2-Teststatistik. Dabei wird getestet, ob die vom Modell implizierte und die empirische Kovarianzmatrix *identisch* sind. Mit zunehmendem Stichprobenumfang reichen dabei allerdings bereits triviale Abweichungen aus, um die Null-Hypothese zurückzuweisen (vgl. z.B. Hu & Bentler, 1995).

Im Gegensatz zur Chi^2-Teststatistik vergleichen sogenannte Komparative Fit-Indizes die „Passung" des Modells mit der eines Nullmodells[55] (*null* bzw. *independent model*), bei dem völlige Unkorreliertheit aller Variablen angenommen wird. Mplus berechnet hierzu den *Tucker-Lewis-Index* (TLI) sowie den *Comparative Fit Index* (CFI) (vgl. Yu, 2002).

Ein anderer Ansatz liegt den *Error-of-Approximation*-Indizes zugrunde. Hier wird die modellimplizierte Kovarianzmatrix nicht mit der Stichproben-, sondern mit der (geschätzten) Populations-Kovarianzmatrix verglichen. Entsprechend sollten solche Fit-Indizes, wie z.B. der *Root-mean-square Error of Approximation* (RMSEA), auch nicht (oder nur minimal) stichprobenabhängig sein.

Residuums-basierte Fit-Indizes, wie z.B. das *Standardized Root-mean-square Residual* (SRMR), basieren auf der Differenz zwischen der Stichproben- und der modellimplizierten

[55] Bei der Chi^2-Teststatistik wird dagegen das hypothetische Modell mit einem saturierten Modell verglichen.

Kovarianzmatrix. Je geringer die Abweichungen, desto niedrigere Werte nimmt das SRMR an.

Aufgrund umfangreicher Simulationsstudien empfiehlt Yu (2002) für konfirmatorische (nicht-hierarchische) Faktorenanalysen folgende *cutoff*-Kriterien: Der TLI sowie der CFI sollten mindestens bei .95 liegen, der RMSEA sollte maximal .05 betragen und das SRMR sollte .07 nicht überschreiten.

Neben Modellgütekriterien, die sich auf die Passung des Gesamtmodells beziehen, wurden auch Indizes entwickelt, die den Vergleich (auch nicht ineinander geschachtelter) Alternativmodelle erlauben. Zu diesen sogenannten Informationskriterien gehören beispielsweise *Akaike's information criterion* (AIC) und das *Bayesian information criterion* (BIC). Im Rahmen einer Simulationsstudie konnten Whittaker und Stapleton (2006) zeigen, dass insbesondere das BIC auch bei Verletzung der Normalverteilungsannahme ein zuverlässiges Kriterium zur Identifikation jeweils angemessenerer Modelle darstellt.

3.4.3 Simulationsstudie: Spezifikation ordinaler Indikatoren als intervallskaliert in einer konfirmatorischen Zwei-Ebenen-Faktorenanalyse mit je zwei Faktoren auf jeder Ebene

Ausgangspunkt der Simulationsstudie ist eine konfirmatorische Faktorenanalyse zweier in dieser Arbeit verwendeter Skalen[56], nämlich *Thematische Motivierung* und *Strukturiertheit* im Fach Englisch, wobei die Indikatoren im Sinne des GRM[57] (mit Logit-*link*) mit itemspezifischen Diskriminationsparametern und Schwellenparametern geschätzt wurden[58]. Die Residualvarianzen der Indikatoren auf Ebene 2 wurden auf Null gesetzt[59]. Die Parameter-Schätzungen wurden anschließend in einem Populationsmodell verwendet, auf dessen Basis 1000 Datensätze generiert wurden[60], die bezüglich des Stichprobenumfangs sowie der verschiedenen Klassengrößen exakt der ursprünglichen Stichprobe entsprachen[61]. Diese Datensätze wurden daraufhin auf der Basis verschiedener Modelle mit metrischen Indikatoren[62] analysiert.

---

[56] Diese beiden Skalen wurden aufgrund der Zwei-Ebenen-Faktorenanalysen in DESI (A. Helmke et al., in Druck-a) exemplarisch ausgewählt, da sie aus wenigen Items bestehen und zudem auf Klassenebene nicht auf einen übergeordneten Faktor laden.

[57] In Mplus steht derzeit nur das GRM zur Verfügung.

[58] In diesem Fall wurde auf eine Gewichtung der Daten, sowie auf Angaben zu Strata verzichtet. Es wurde also hier die Stichprobe, nicht die Population beschrieben.

[59] Dies ist auch Voreinstellung in Mplus bei Zwei-Ebenen-Modellen mit ordinalen Daten.

[60] Die Daten wurden mithilfe des MODEL POPULATION-Kommandos in Mplus generiert. Auf die Erzeugung fehlender Werte wurde verzichtet.

[61] Der Stichprobenumfang beträgt 7420 Fälle aus 330 Klassen mit 28 verschiedenen Klassengrößen (3 bis 33 Schüler pro Klasse). Die Angaben zur „Klassengröße" beziehen sich hier auf die Schülerinnen und Schüler, für die Angaben zu den entsprechenden Items vorliegen, nicht auf die tatsächliche Klassengröße.

[62] Die Analysen erfolgten ebenfalls in Mplus.

In Tabelle 5 sind die Schwellenparameter des IRT-Modells für die insgesamt fünf Items der beiden Faktoren *Thematische Motivierung* und *Strukturiertheit* dargestellt[63]. Die restlichen Parameter dieses Modells können Tabelle 6 entnommen werden, da im Populationsmodell die Schätzungen – also auch die *threshold*-Parameter aus Tabelle 5 – aus dem IRT-Modell auf der Basis der Stichprobe übernommen wurden.

Tabelle 5: Schwellenparameter (*thresholds*) des Zwei-Ebenen-IRT-Modells mit den Faktoren *Thematische Motivierung* und *Strukturiertheit* im Fach Englisch

	thresholds		
Item	**1**	**2**	**3**
mote1	-3.210	0.115	3.737
mote2	-3.441	0.619	5.083
strukte1	-1.515	1.150	4.285
strukte2	-2.272	1.330	5.303
strukte3	-2.800	-0.025	3.750

Tabelle 6 enthält außerdem die gemittelten Ergebnisse aus den 1000 Replikationen anhand eines analogen metrischen Modells – also ebenfalls mit den beiden Faktoren auf beiden Ebenen, erste Ladung jeweils fixiert auf 1, Fixierung der Residualvarianzen auf Ebene 2 auf Null. Die unstandardisierten Koeffizienten können jeweils nur innerhalb des jeweiligen Modells interpretiert werden, da ihnen im IRT- und im metrischen Modell unterschiedliche Item- und Faktorvarianzen zugrunde liegen. Deshalb wurden auf der Basis der Mittelwerte (Spalte „M“) der unstandardisierten Koeffizienten[64] und der jeweiligen Faktorvarianz standardisierte Koeffizienten berechnet (also standardisierte Ladungen und Korrelationen). Daneben wurden sowohl für das Populationsmodell als auch für das Replikationsmodell unstandardisierte sowie standardisierte Kommunalitäten[65] – also Varianzanteile der Items, die auf den Einfluss des Faktors zurückgeführt werden – berechnet. Auch hier sind die unstandardisierten Koeffizienten nur innerhalb des jeweiligen Modells interpretierbar. Deshalb wurde eine Standardisierung auf eine Gesamtvarianz von 1 vorgenommen, d.h. die Summe der Kommunalitäten jedes Items auf beiden Ebenen ergibt eine Varianz von 1. Dementsprechend kann die standardisierte Kommunalität auf Ebene 2 im Sinne einer ICC interpretiert werden.

[63] Im *threshold*-Modell wird nicht zwischen Ebenen unterschieden, weshalb hier für jedes Item nur drei Schwellenparameter angegeben werden (vgl. Grilli & Rampichini, 2007). Die Indikatoren des Faktors *Thematische Motivierung* im Fach Englisch sind mit mote1-mote2 bezeichnet, die Indikatoren des Faktors *Strukturiertheit* im Fach Englisch mit strukte1-strukte3.

[64] Die Spalte „SD“ enthält die Standardabweichung der Koeffizienten aus den 1000 Replikationen.

[65] Die Kommunalitäten entsprechen hier der quadrierten Faktorladung multipliziert mit der Faktorvarianz auf der jeweiligen Ebene (vgl. Grilli & Rampichini, 2007).

Tabelle 6: Simulation I: IRT-Populationsmodell und identisches Analysemodell, jedoch mit intervallskalierten Indikatoren

	Populationsmodell					Replikationen (intervallskalierte Indikatoren)						*bias* in Prozent		
				Kommunalität		unstandardisierte Koeffizienten				Kommunalität				
	unstandardisierte Koeffizienten	standardisierte Koeffizienten	Reliabilität	unstandardisiert	standardisiert	M	SD	standardisierte Koeffizienten	Reliabilität	unstandardisiert	standardisiert	standardisierte Koeffizienten	Reliabilität	Kommunalität (standardisiert)
Ebene 1 (innerhalb)														
Ladungen (mote)														
mote1	1.000	0.805	0.648	6.053	0.767	1.000	0.000	0.758	0.574	0.412	0.770	-5.9	-11.4	0.4
mote2	1.221	0.856	0.733	9.024	0.726	1.009	0.025	0.801	0.642	0.420	0.730	-6.4	-12.4	0.7
Ladungen (strukte)														
strukte1	1.000	0.711	0.506	3.356	0.765	1.000	0.000	0.668	0.446	0.312	0.770	-6.1	-11.8	0.6
strukte2	1.391	0.815	0.664	6.493	0.855	1.109	0.024	0.763	0.582	0.384	0.858	-6.4	-12.4	0.3
strukte3	1.132	0.753	0.567	4.300	0.838	1.055	0.023	0.704	0.496	0.347	0.838	-6.5	-12.5	0.1
Korrelation														
mote, strukte	2.803	0.622				0.221	0.008	0.615				-1.1		
Varianzen der latenten Faktoren														
mote	6.053	1.000				0.412	0.014							
strukte	3.356	1.000				0.312	0.011							
Ebene 2 (zwischen)														
Ladungen (mote)														
mote1	1.000	1.000		1.834	0.233	1.000	0.000	1.000		0.123	0.230			-1.2
mote2	1.364	1.000		3.412	0.274	1.123	0.028	1.000		0.155	0.270			-1.7
Ladungen (strukte)														
strukte1	1.000	1.000		1.033	0.235	1.000	0.000	1.000		0.093	0.230			-2.1
strukte2	1.034	1.000		1.104	0.145	0.826	0.031	1.000		0.064	0.142			-2.0
strukte3	0.898	1.000		0.833	0.162	0.847	0.034	1.000		0.067	0.162			-0.3
Korrelation														
mote, strukte	1.045	0.759				0.082	0.009	0.761				0.3		
Varianzen der latenten Faktoren														
mote	1.834	1.000				0.123	0.011							
strukte	1.033	1.000				0.093	0.010							

Die letzten drei Spalten in Tabelle 6 enthalten Angaben zur Verzerrung (*bias*) aufgrund der Behandlung der ordinalen Daten aus dem IRT-Populationsmodell (kodiert mit den Werten 0-3) im Sinne intervallskalierter Daten. Die standardisierten Faktorladungen auf Ebene 1[66] weisen einen *bias* im Bereich von -6.5% bis -5.9% auf, der *bias* bezüglich der Reliabilitäten liegt im Bereich von -12.5% bis -11.4%. Das heißt, die Ladungen bzw. Reliabilitäten werden systematisch – jedoch über verschiedene Items hinweg jeweils in vergleichbarer Größenordnung – unterschätzt. Betrachtet man allerdings die standardisierten Kommunalitäten, so zeigen sich nur geringfügige Überschätzungen auf Ebene 1 (0.1% bis 0.7%) und minimale Unterschätzungen (-2.1% bis -0.3%) auf Ebene 2. Auch die Korrelationen der Faktoren auf beiden Ebenen sind kaum verzerrt: Auf Ebene 1 wird die Korrelation geringfügig unterschätzt (-1.1%), auf Klassenebene überschätzt (0.3%).

Wie bereits in Kapitel 3.4.2 erwähnt, sollte im Rahmen der Simulationsstudie auch das auf der Basis der Überlegungen von O'Brien vorgeschlagene Adjustierungsverfahren überprüft werden. Dazu wurden zunächst die Kategorienwahrscheinlichkeiten auf der Basis der 1000 Replikationen des oben beschriebenen Modells ermittelt (vgl. Tabelle 7). Anhand dieser Wahrscheinlichkeiten wurden der Mittelwert der arbiträren *scores* (M(a)) sowie die Korrelationen der optimalen (r_{ku}) und der arbiträren *scores* mit der hypothetischen intervallskalierten, normalverteilten Variable ermittelt (vgl. Tabelle 7). Die beiden hier dargestellten Korrelationen liegen jeweils dicht beieinander. Das heißt, der Einfluss des arbiträren *scorings* ist hier relativ gering. Die Korrelationen der arbiträren mit den hypothetischen *scores* liegen im Bereich von $.927 \leq r_g \leq .938$, was 86% bis 88% gemeinsamer Varianz entspricht (bzw. 12% bis 14% *grouping error*-Varianz bezogen auf die Gesamtvarianz).

Tabelle 7: Kategorienwahrscheinlichkeiten, Mittelwerte der arbiträren *scores* und Korrelationen der optimalen (r_{ku}) sowie der arbiträren *scores* (r_g) mit den hypothetisch zugrunde liegenden intervallskalierten, normalverteilten Variablen[67]

	Kategorie (*score*) Wahrscheinlichkeit in Prozent						
Indikator	1(0)	2(1)	3(2)	4(3)	M(a)	r_{ku}	r_g
mote1	16.7	34.7	35.6	13.0	1.45	.939	.938
mote2	19.2	37.1	33.8	9.9	1.34	.937	.936
strukte1	28.9	37.6	27.6	6.0	1.11	.928	.927
strukte2	24.4	41.5	28.8	5.4	1.15	.929	.928
strukte3	16.4	33.3	40.7	9.6	1.44	.934	.933

[66] Da die Residualvarianz auf Ebene 2 auf Null fixiert wurde, sind dort in beiden Modellen alle standardisierten Ladungen gleich 1.

[67] Die Angaben beziehen sich auf die Daten der gesamten 1000 Replikationen des in Tabelle 6 dargestellten Modells.

Tabelle 8 zeigt die zur Adjustierung benötigten ICCs sowie die jeweils mithilfe der ICC ermittelte Korrelation der kategorisierten Variable mit der (hypothetisch) zugrunde liegenden intervallskalierten, normalverteilten Variable auf Ebene 1. Der *bias* der entsprechend adjustierten Koeffizienten liegt bei den standardisierten Ladungen zwischen 1.0% und 2.5%, bei den Reliabilitäten zwischen 2.0% und 5.0%[68]. Die adjustierten Koeffizienten liegen hier somit deutlich dichter an den im Populationsmodell vorgegebenen Werten als die unadjustierten (vgl. Tabelle 6). Besonders deutliche Unterschiede finden sich bei den ICCs: Die unadjustierten Koeffizienten (vgl. Tabelle 8) unterschätzen hier die Populationsparameter um 10.9% bis 13.2%, während der *bias* der adjustierten ICCs zwischen -2.0% und -0.7% liegt.

Tabelle 8: Unadjustierte ICC, gemeinsame Varianz der kategorisierten Variable mit arbiträren *scores* und der hypothetischen intervallskalierten Variable (innerhalb), adjustierte Koeffizienten, *bias* in Prozent

			adjustierte Koeffizienten			*bias* der adjustierten Koeffizienten in Prozent			
Indikator	ICC (unadjustiert)	r_{gw}^2	standardisierte Ladung (innerhalb)	Reliabilität (innerhalb)	ICC	*bias* der unadjustierten ICC in Prozent	standardisierte Ladung (innerhalb)	Reliabilität (innerhalb)	ICC
mote1	.146	.859	.817	.668	.166	-10.9	1.5	3.1	-1.3
mote2	.192	.847	.870	.758	.219	-11.7	1.7	3.4	-0.7
strukte1	.118	.840	.729	.531	.137		2.5	5.0	-2.0
strukte2	.088	.849	.828	.685	.102	-12.4	1.6	3.2	-0.7
strukte3	.087	.858	.760	.578	.100	-13.2	1.0	2.0	-1.5

Ein weiteres Ziel der Simulationsstudie war die Überprüfung der Angemessenheit verschiedener Modellgüte-Indizes. Dazu wurden insgesamt fünf verschiedene metrische KFA-Modelle getestet, die in unterschiedlicher Weise vom Populationsmodell abweichen:

- Modell I: Dieses Modell entspricht dem Populationsmodell. Der einzige Unterschied besteht darin, dass die ordinalen Daten des Populationsmodells als intervallskalierte Indikatoren interpretiert werden.
- Modell II: Hier wurden die fünf Indikatoren auf Ebene 1 einem einzigen Faktor zugeordnet. Das Modell auf Ebene 2 hingegen ist identisch mit dem in Modell 1.

[68] Die in Tabelle 6 aufgeführten und hier zugrunde gelegten standardisierten Ladungen und Reliabilitäten des ordinalen Modells wurden aus Mplus übernommen. Da Mplus die Standardisierung bei Logit-*link*-Modellen anhand der Residualvarianz von Var(Residual, logit) = $\pi^2/3 \approx 3.290$ vornimmt (vgl. B. O. Muthén, 2004), die durch eine bessere Approximation an die Standardnormalverteilung ersetzt werden kann – Var(Residual, logit) = $[(15/16)(\pi/\sqrt{3})]^2 \approx 2.891$ (vgl. dazu Kap. 4.2) – wurden die adjustierten Koeffizienten zusätzlich mit der auf der Basis der letztgenannten Approximation standardisierten Koeffizienten verglichen. Der *bias* der adjustierten standardisierten Ladungen liegt hier im Bereich zwischen -1.6% und 0.1%, der *bias* der adjustierten Reliabilitäten zwischen -3.2% und 0.1%. Die adjustierten ICCs weisen Überschätzungen im Bereich von 1.8% bis 3.3% auf.

- Modell III: Das Modell auf Ebene 1 ist identisch mit dem in Modell 1. Die fünf Indikatoren auf Ebene 2 wurden einem einzigen Faktor zugeordnet.
- Modell IV: Hier wurde auf Ebene 1 das dritte Item des Faktors *Strukturiertheit* fälschlicherweise dem Faktor *Thematische Motiviertheit* zugeordnet (anstatt dem Faktor *Strukturiertheit*; also keine Doppelladung). Das Modell auf Ebene 2 ist hingegen korrekt spezifiziert.
- Modell V: Analog zu Modell IV wurde hier auf Ebene 2 (statt auf Ebene 1) das dritte Item des Faktors *Strukturiertheit* fälschlicherweise dem Faktor *Thematische Motiviertheit* zugeordnet, während das Modell auf Ebene 1 korrekt spezifiziert wurde.

In Tabelle 9 sind die Ergebnisse der Simulation zu den verschiedenen Modell-Fit-Indizes dargestellt. Legt man die von Yu (2002) für KFA-Modelle vorgeschlagenen *cutoff*-Kriterien zugrunde (vgl. Kap. 3.4.2), so zeigt sich, dass das korrekt spezifizierte Modell allen Gütekriterien entspricht. Zudem sind die Streuungen über die verschiedenen Replikationen hinweg sehr gering. Bei den Modellen mit Fehlspezifikationen auf Ebene 1 (Modell II und Model IV) sind alle Indizes, die sich auf das Gesamtmodell (also beide Ebenen gleichzeitig) beziehen – CFI, TLI, RMSEA –, sowie der SRMR für Ebene 1 oberhalb des jeweiligen Kriteriums. Umgekehrt zeigt sich, dass die sich auf das Gesamtmodell beziehenden Indizes bei den Modellen mit Fehlspezifikationen auf Ebene 2 (Modell III und Modell V) keinen adäquaten „Schutz" vor Fehlspezifikationen darstellen. Dies dürfte auf die relativ geringen Varianzanteile der Variablen auf Ebene 2 zurückzuführen sein[69]. Hier sind die ebenenspezifischen SRMR-Indizes deutlich effektiver: Beide Modelle mit Ebene-2-Fehlspezifikationen werden auf der Basis der *cutoff*-Kriterien zurückgewiesen.

[69] Im umgekehrten Fall – also bei deutlich größeren Varianzanteilen auf Ebene 2 verglichen mit denen auf Ebene 1 (was in der Praxis allerdings vermutlich selten vorkommen dürfte) – wäre wohl ein gegensätzlicher Effekt zu erwarten.

Tabelle 9: Modellgüte-Indizes für verschiedene KFA-Modelle (Mittelwerte und Standardabweichungen bezüglich der 1000 Replikationen)

	Modellbeschreibung	CFI M / SD	TLI M / SD	RMSEA M / SD	SRMR (Ebene 2) M / SD	SRMR (Ebene 1) M / SD
Modell I	„korrektes" Modell: jeweils 2 Faktoren auf beiden Ebenen	1.000 / 0.000	1.000 / 0.001	0.001 / 0.002	0.013 / 0.005	0.003 / 0.001
Modell II	Fehlspezifikation auf Ebene 1: nur 1 Faktor auf Ebene 1	0.860 / 0.007	0.800 / 0.010	0.127 / 0.003	0.024 / 0.003	0.079 / 0.002
Modell III	Fehlspezifikation auf Ebene 2: nur 1 Faktor auf Ebene 2	0.979 / 0.004	0.970 / 0.005	0.049 / 0.004	0.154 / 0.020	0.013 / 0.002
Modell IV	Fehlspezifikation auf Ebene 1: strukte3 wird Faktor mote zugeordnet	0.879 / 0.006	0.814 / 0.010	0.122 / 0.003	0.025 / 0.003	0.081 / 0.004
Modell V	Fehlspezifikation auf Ebene 2: strukte3 wird Faktor mote zugeordnet	0.984 / 0.003	0.976 / 0.005	0.043 / 0.004	0.121 / 0.016	0.008 / 0.001

Die insgesamt sehr gute Schätzung der Koeffizienten sowie die Angemessenheit der Modellgüte-Indizes ist vermutlich zu einem großen Teil auf die relativ geringen Verletzungen der Normalverteilungsannahmen der generierten Daten zurückzuführen. Tabelle 10 zeigt die Schiefe und den Exzess der generierten Daten (zusammengefasst für alle 1000 Replikationen) auf Itemebene. Wie bereits oben erwähnt (vgl. Kap. 3.4.2), werden üblicherweise Schiefe und Exzess in einem Bereich von -1 bis +1 als minimale Verletzung der Normalverteilungsannahme betrachtet.

Tabelle 10: Schiefe und Exzess[70] der auf der Basis des Zwei-Ebenen-IRT-Modells generierten Daten (Basis: Daten aus allen 1000 Replikationen)

Item	Schiefe (skewness)	Exzess (kurtosis)
mote1	0.007	-0.831
mote2	0.088	-0.796
strukte1	0.299	-0.806
strukte2	0.230	-0.697
strukte3	-0.106	-0.742

Da im Rahmen der vorliegenden Arbeit auch die Annahme der Invarianz des Messmodells eines Faktors auf zwei Ebenen überprüft werden soll – dies ist wie in Kapitel 3.4.1 ausgeführt Voraussetzung für die Berechnung latenter ICCs – wurden auf der Basis eines zweiten, leicht abgewandelten Populationsmodells erneut 1000 Datensätze generiert. In diesem zweiten Populationsmodell wurden folgende Parameter geändert:

- Um ein invariantes Messmodell für den Faktor *Strukturiertheit* auf beiden Ebenen in der Population zu formulieren, wurden die Ladungen des Faktors auf Ebene 2 den ent-

[70] Die Berechnung erfolgte mit der SAS-Prozedur UNIVARIATE.

sprechenden Ladungen auf Ebene 1 (aus dem ursprünglichen Populationsmodell) angepasst.

- Um sicherzustellen, dass sich das Messmodell für den ersten Faktor (*Thematische Motivierung*) auf beiden Ebenen hinreichend unterscheidet, wurde die zweite Ladung (mote2) auf Ebene 2 auf 1.6 gesetzt (während die entsprechende Ladung auf Ebene 1 – $\lambda = 1.221$ – beibehalten wurde).

Die so generierten Daten[71] wurden anschließend anhand verschiedener metrischer Modelle analysiert. Folgende metrischen Modelle wurden getestet:

- Modell A entspricht dem analogen metrischen Modell zum IRT-Populationsmodell: jeweils zwei Faktoren pro Ebene, erste Ladung jeweils auf 1 fixiert, Residualvarianz auf Ebene 2 auf Null fixiert, itemspezifisch identische Ladungen für den Faktor *Strukturiertheit* auf beiden Ebenen.
- Modell B entspricht im Wesentlichen Modell A. Im Unterschied zu Modell A werden hier jedoch alle Ladungen (bis auf die erste, die jeweils zur Identifikation auf 1 fixiert wurde) frei geschätzt. Das heißt, die Ladungen für den Faktor *Strukturiertheit* auf Ebene 1 und Ebene 2 wurden nicht gleichgesetzt.
- Modell C entspricht Modell A mit der zusätzlichen (falschen) Restriktion paralleler Ladungen für den Faktor *Thematische Motivierung*.

Modell A führte zu vergleichbar minimalen Verzerrungen[72] bezüglich der standardisierten Ladungen bzw. Reliabilitäten und Kommunalitäten wie Modell I in der ersten Simulation. Auf eine detaillierte Ergebnisdarstellung wird deshalb verzichtet. Da in dieser Simulation Messinvarianz bezüglich des Faktors *Strukturiertheit* auf beiden Ebenen angenommen wird, lassen sich entsprechend auch latente ICCs berechnen: Die ICC im Populationsmodell beträgt 0.235, die mittlere Schätzung aus den metrischen Modellen liegt bei 0.232. Das Ergebnis der Intervallskalenniveau voraussetzenden Analyse ist also – analog zu den standardisierten Kommunalitäten – nahezu identisch mit den Vorgaben des IRT-Populationsmodells (*bias*: -1.6%).

In Tabelle 11 sind die Ergebnisse bezüglich der Modellgüte-Indizes dargestellt. Es zeigt sich, dass diese Indizes für solche eher minimalen Fehlanpassungen – hier wurden ja lediglich unterschiedliche Messmodelle miteinander verglichen, während die Zuordnung von Indikato-

[71] Auch hier ergaben sich nur minimale Verletzungen der Normalverteilungsannahme: Die Schiefe lag im Bereich von -0.094 bis 0.299, die Kurtosis im Bereich von -0.852 bis -0.738 (bezogen auf die zusammengefassten Daten aus den 1000 Replikationen). Die Berechnung erfolgte ebenfalls mit SAS PROC UNIVARIATE.

[72] *bias* auf Ebene 1: -6.7% bis -5.9% bei den standardisierten Ladungen, -12.9% bis -11.4% bei den Reliabilitäten, 0.4% bis 0.9% bei den Kommunalitäten; *bias* auf Ebene 2: -1.8% bis -1.2% bei den Kommunalitäten

ren zu den entsprechenden Faktoren jeweils dem Populationsmodell entsprach – wenig sensibel sind. Die Indizes für Modell A (das „korrekt“ spezifizierte Modell) und Modell B (frei geschätzte Ladungen) sind nahezu identisch, während Modell C insgesamt eine etwas schlechtere Anpassung aufweist, die aber keineswegs zu einer Ablehnung dieses Modells führen würde.

Tabelle 11: Modellgüte-Indizes für verschiedene KFA-Modelle (Mittelwerte und Standardabweichungen bezüglich der 1000 Replikationen des zweiten Populationsmodells)

		CFI	TLI	RMSEA	SRMR (Ebene 2)	SRMR (Ebene 1)
	Modellbeschreibung	**M / SD**	**M / SD**	**M / SD**	**M / SD**	**M / SD**
Modell A	„korrektes“ Modell: identische Ladungen für *Strukturiertheit* auf beiden Ebenen	1.000 / 0.000	1.000 / 0.001	0.001 / 0.003	0.011 / 0.004	0.004 / 0.001
Modell B	freie Ladungen für beide Faktoren auf beiden Ebenen	1.000 / 0.000	1.000 / 0.001	0.001 / 0.003	0.011 / 0.004	0.003 / 0.001
Modell C	identische Ladungen für *Strukturiertheit* und *Thematische Motivierung* (Fehlspezifikation) auf beiden Ebenen	0.996 / 0.001	0.995 / 0.002	0.021 / 0.003	0.011 / 0.004	0.014 / 0.002

In Modell B (frei geschätzte Ladungen) wurde zusätzlich mittels Wald-Test die Restriktion itemspezifisch identischer Ladungen für den Faktor *Strukturiertheit* auf beiden Ebenen getestet. Im Mittel ergab sich ein Wert von 2.057 mit einer Standardabweichung von 2.028 für die Chi^2-verteilte Teststatistik mit zwei Freiheitsgraden (da das erste Item jeweils auf den Wert 1 fixiert wurde, wurden parallele Ladungen für die verbleibenden zwei Items getestet). Bei zwei Freiheitsgraden sind Chi^2-Werte ab etwa 6 signifikant (5%-Niveau), d.h. der Mittelwert von 2.028 wäre nicht signifikant. Dies bedeutet, dass die Restriktion die Modellpassung nicht signifikant „verschlechtert“, weshalb die Restriktion (im Allgemeinen) im Sinne des Sparsamkeitskriteriums (einfache Modelle mit wenigen Parametern sind komplexeren Modellen vorzuziehen) akzeptiert würde.

Etwas genauer aufgeschlüsselt ist dieses Ergebnis in Tabelle 12 dargestellt. Die Tabelle ist folgendermaßen zu interpretieren: Die Spalten mit der Bezeichnung „Erwartet“ stellen korrespondierende Wahrscheinlichkeiten (unter „Proportionen in Prozent“) und Chi^2-Werte (bei zwei Freiheitsgraden; unter „Perzentile (Chi^2)“) dar. Der Chi^2-Wert beim üblichen 5%-Signifikanzniveau liegt also bei 5.991. Das fünfte Perzentil der Chi^2-Werte in der Replikationsstudie überschreitet diesen Wert minimal: Dort liegt der Wert bei 6.246 (Spalte: „Beobachtet“ bei „Perzentile (Chi^2)“). Analog dazu zeigt sich in der Spalte „Beobachtet“ bei den „Proportionen“, dass 5.5% der Replikationen den kritischen Wert für das 5%-Niveau (also 5.991) überschritten haben. Das heißt, der Wald-Test testet hier sehr zuverlässig auf dem 5%-Niveau.

Darüber hinaus zeigt sich, dass die Teststatistik über den gesamten Bereich – also von 1% bis 99% – relativ nahe bei den erwarteten Werten liegt.

Tabelle 12: Wald-Test für itemspezifisch identische Ladungen (Basis: Modell B) für den Faktor *Strukturiertheit* auf beiden Ebenen (Erwartungswerte und beobachtete Werte auf der Basis von 1000 Replikationen)

Proportionen in Prozent		Perzentile (Chi^2)	
Erwartet	Beobachtet	Erwartet	Beobachtet
99	99.1	0.020	0.029
98	98.2	0.040	0.052
95	95.6	0.103	0.119
90	91.8	0.211	0.236
80	81.8	0.446	0.495
70	71.4	0.713	0.778
50	49.9	1.386	1.384
30	30.0	2.408	2.405
20	21.9	3.219	3.369
10	10.2	4.605	4.668
5	5.5	5.991	6.246
2	2.3	7.824	8.009
1	1.0	9.210	8.982

Auch Modell C (Fehlspezifikation: itemspezifisch identische Ladungen für beide Faktoren auf beiden Ebenen) wurde mittels Wald-Test überprüft, indem in Modell A (das „korrekte“ Modell) die Restriktion identischer Ladungen für das zweite Item des Faktors *Thematische Motivierung* getestet wurde. Der Mittelwert der Chi^2-Werte bezogen auf die 1000 Replikationen liegt bei 55.280, die Standardabweichung bei 15.247. Verglichen mit dem Chi^2-Wert für das 5%-Niveau bei einem Freiheitsgrad – dieser liegt bei etwa 3.8 – zeigt sich, dass die Ergebnisse im Durchschnitt deutlich über diesem Wert liegen. Dies hätte zur Folge, dass die fehlspezifizierte Restriktion abgelehnt würde. Und dies wohl zuverlässig in (nahezu) allen Fällen: Das 99. Perzentil der beobachteten Chi^2-Werte liegt noch bei 25.378, d.h. nur 1% der Chi^2-Werte unterschreiten diesen Wert.

Zusätzlich zum Wald-Test wurden auch Informationskriterien zur Identifikation des jeweils adäquateren Modells herangezogen. Dazu wurden für jeden einzelnen Replikationsdatensatz zwei konkurrierende Modelle anhand des AIC, des BIC sowie des anhand des Stichprobenumfangs adjustierten BIC (BIC-adj.) miteinander verglichen. Der jeweils kleinere Wert identifiziert dabei das unter den Gesichtspunkten Sparsamkeit und Modellpassung angemessenere Modell. Beim Vergleich der Modelle A und B (also „korrektes“ vs. zu wenig restriktives Modell) ergaben sich folgende „Trefferquoten“: Das AIC identifizierte in 85.9%, das BIC in 100% (!) und das BIC-adj. in 99.7% der Fälle das „korrekte“, sparsamere Modell A[73].

[73] Zur hervorragenden Performanz des BIC (auch im Falle nicht-normalverteilter Indikatoren) vergleiche auch Whittaker und Stapleton (2006).

Beim Vergleich der Modelle A und C (also dem „korrekten“ und dem fehlspezifizierten Modell mit identischen Ladungen für das zweite Item des Faktors *Thematische Motivierung*) lag die „Trefferquote“ bei allen drei Indizes bei 100%: Das fehlspezifizierte Modell C wurde immer abgelehnt.

Als Ergebnis der Simulationsstudie lassen sich folgende Befunde zusammenfassen:

Erwartungsgemäß stellen die standardisierten Faktorladungen – und somit auch die Reliabilitäten der Indikatoren – der metrischen Modelle auf Individualebene eine (geringfügige) Unterschätzung der tatsächlichen Koeffizienten des IRT-Populationsmodells dar, da gewissermaßen bei metrischen Modellen die *threshold*-Informationen nicht genutzt werden[74] bzw. die *grouping error*-Komponente nicht berücksichtigt wird. Dies ist gleichbedeutend mit größeren Fehlervarianzkomponenten der Indikatoren, die aber offensichtlich – das zeigen die sehr günstig ausfallenden Modellgüte-Indizes an – auch tatsächlich im Sinne unkorrelierter Fehler betrachtet werden können. Die standardisierten Ladungen und Reliabilitäten – sowie die ICCs – lassen sich jedoch mithilfe des in Kapitel 3.4.2 vorgeschlagenen Verfahrens hier relativ gut adjustieren.

Betrachtet man hingegen die standardisierten Kommunalitäten des IRT-Populationsmodells und des metrischen Modells, so zeigen sich nur minimale Unterschiede (der maximale *bias* betrug 0.9% auf Ebene 1 und -2.1% auf Ebene 2). Da sich die unstandardisierten Ladungen auf diese Varianzkomponente der Indikatoren beziehen (und nicht auf die Fehlervarianz), ist dieser Befund im Zusammenhang mit der Überprüfung der Messinvarianz eines Faktors (im Sinne identischer *unstandardisierter* Ladungen) über die Ebenen hinweg wesentlich bedeutsamer als die Unterschiede bezüglich der *standardisierten* Ladungen.

Entsprechend zeigten sich auch die auf die Messinvarianz bezogenen Tests im Rahmen des metrischen Modells – der Wald-Test sowie die verschiedenen Informationskriterien (hier vor allem das BIC) – als äußerst zuverlässig. Beim Wald-Test zeigte sich zudem, dass die beobachtete Verteilung der Koeffizienten der erwarteten Verteilung sehr gut entspricht. Das heißt, es kann davon ausgegangen werden, dass ein angenommenes 5%-Signifikanzniveau auch (nur) in etwa 5% aller Fälle zu einer Fehlentscheidung führt. Dies legt die Verwendung des Wald-Tests für die vorliegende Untersuchung nahe, da hier – im Gegensatz zu den Informationskriterien – ein Signifikanzniveau angegeben werden kann. Ein weiterer Vorteil des Wald-Tests besteht darin, dass nur ein Modell (mit frei geschätzten Ladungen auf beiden Ebe-

[74] Das zeigt sich auch bei den ICCs auf Indikatorebene (vgl. dazu auch Kap. 4.2): Die metrischen ICCs stellen Unterschätzungen der ordinalen ICCs dar, da dort der relative Varianzanteil auf Ebene 1 durch diese „Fehlanpassung“ überschätzt wird.

nen) geschätzt werden muss, bei dem die Restriktionen itemspezifisch identischer (unstandardisierter) Ladungen in Mplus mittels des MODEL TEST-Kommandos getestet werden können.

4. Ergebnisse

Alle im Folgenden dargestellten Ergebnisse beziehen sich auf die in Kapitel 3.2 beschriebene Population. Bei allen Analysen wurden dazu die entsprechenden DESI-Populationsgewichte verwendet. Zusätzlich wurden bei den Mplus-Analysen Stratuminformationen berücksichtigt[75].

Zunächst werden deskriptive Befunde berichtet (Kap. 4.1). Es folgen Überprüfungen der *Interrater*-Reliabilitäten (Kap. 4.2) hinsichtlich der Unterrichtswahrnehmungen aus Schülersicht. Dort wird auch auf Unterschiede in den Wahrnehmungen zwischen den weitgehend als getrennt zu betrachtenden sozialen Netzwerken der Mädchen und Jungen eingegangen. In Kapitel 4.3 wird der Frage der Differenziertheit der Unterrichtswahrnehmung aus Schülersicht nachgegangen. Dort werden auch intraindividuelle Unterschiede der Wahrnehmungen in den Fächern Deutsch und Englisch untersucht. Kapitel 4.4 befasst sich mit der Frage der Isomorphie der parallel in beiden Fächern erhobenen Unterrichtswahrnehmungen. Es folgen Untersuchungen zum Einfluss der Itemformulierung (Ich- vs. Klassen-Bezug; Kap. 4.5). Abschließend (Kap. 4.6) geht es um Fragen der möglichen Beeinflussung der Unterrichtswahrnehmungen durch die „Prädiktoren" Geschlecht, Fachleistung bzw. -noten sowie das auf die Lehrkraft bezogene Globalurteil. Letzteres wird gleichfalls als „abhängige Variable" betrachtet. Dabei wird untersucht, ob die jeweiligen Unterrichtswahrnehmungen in konstanter Weise zur Bildung des Globalurteils beitragen oder ob die Gewichtungen über die Lehrkräfte hinweg variieren.

4.1 Deskriptive Befunde

Betrachtet man die Verteilungen der Indikatoren der hier untersuchten Unterrichtsmerkmale aus Schülersicht (vgl. Tabelle 13), dann zeigt sich, dass die relativen Häufigkeiten der Antwortkategorien im Bereich zwischen 5.5% (vierte Kategorie des zweiten Items der Skala *Strukturiertheit* im Fach Englisch) und 53.4% (dritte Kategorie des ersten Items der Skala *Verständlichkeit* im Fach Deutsch) liegen. Werden für die Antwortkategorien die Kategorien*scores* verwendet, dann sind eher geringe Abweichungen von der Normalverteilung (vgl. dazu auch Tabelle 14) zu beobachten: Die Kurtosis liegt im Bereich von -0.893 (erstes Item der Skala *Strukturiertheit* im Fach Deutsch) bis -0.113 (erstes Item der Skala *Verständlichkeit* im Fach Deutsch), die Schiefe (*skewness*) im Bereich von -0.555 (drittes Item der Skala *Ver-*

[75] Dies wurde mithilfe der Mplus-Optionen STRATIFICATION und COMPLEX TWOLEVEL realisiert. In DESI wurden jeweils zwei Schulen (also zwei bis vier Klassen) zu einer sogenannten *sampling zone* zusammengefasst.

ständlichkeit im Fach Deutsch) bis 0.248 (erstes Item der Skala *Strukturiertheit* im Fach Englisch). Die auf der Basis der Kategorien*scores* ermittelten (gewichteten) Mittelwerte weichen nur mäßig von dem theoretischen Mittelwert (M = 2.5) der vierstufigen Skala ab: Das Minimum (M = 2.13) findet sich beim ersten Item der Skala *Strukturiertheit* im Fach Englisch, das Maximum (M = 2.89) beim dritten Item der Skala *Verständlichkeit* im Fach Deutsch. Die Streuungen variieren zwischen 0.81 und 0.93 Standardabweichungen. Valide Angaben liegen von 7170 bis 7503 Schülerinnen und Schülern vor, was einem Anteil von 10.9% bis 14.9% fehlender Werte entspricht.

In Tabelle 15 sind die Reliabilitäten der Skalen bzw. Faktor*scores* der Unterrichtswahrnehmungen aus Sicht der Schülerinnen und Schüler angegeben. Dabei bleibt hier – aus Gründen der Vergleichbarkeit mit anderen Studien – die Mehrebenenstruktur zunächst unberücksichtigt. Aus denselben Gründen wird auch das ungewichtete Cronbachs α (mit „Stichprobe" überschriebene Spalten) berichtet, das sich hier erwartungsgemäß kaum von dem gewichteten unterscheidet. Die *factor determinacy* gibt die Korrelation der Faktor*scores* mit dem latenten Faktor an (vgl. B. O. Muthén, 2004). Um die Vergleichbarkeit mit Cronbachs α zu erleichtern, das sich auf den relativen *true score*-Varianzanteil einer Skala bezieht, sind auch die quadrierten Korrelationen in der Tabelle angegeben (Spalte „quadrierte *factor determinacy*"). Die Reliabilitäten der Faktor*scores* liegen erwartungsgemäß über den Reliabilitäten, die auf der Basis von Cronbachs α berechnet wurden. Dies ist in zweierlei Hinsicht erwartungsgemäß: (1) Wenn – wie hier – von kongenerischen Messungen (d.h.: eine gemeinsame latente Variable weist verschiedene Ladungen auf verschiedene Indikatoren auf, die wiederum unterschiedliche Messfehleranteile besitzen) ausgegangen wird, dann stellt Cronbachs α eine untere Grenze der Reliabilität dar, d.h. die wahre Reliabilität beträgt *minimal* α (vgl. z.B. Bollen, 1989; Steyer & Eid, 2001). (2) Da sich die in Tabelle 15 dargestellten Faktor*score*-Reliabilitäten auf ein Gesamtmodell mit allen fünf Unterrichtsmerkmalen in beiden Fächern (also insgesamt zehn Faktoren) beziehen, die teilweise erhebliche Interkorrelationen aufweisen, erhöhen sich die Reliabilitäten, weil zusätzlich Informationen der Indikatoren der jeweils kovariierenden Faktoren genutzt werden können.

Tabelle 13: Verteilungen der Kategorien, Mittelwerte, Standardabweichungen, Anzahl gültiger Werte, Anteil fehlender Werte, Kurtosis und Schiefe der Indikatoren[76]

		prozentuale Kategorienhäufigkeit (gewichtet)[77]							fehlende Angaben		
	Item	1	2	3	4	M[78]	SD	N	(Prozent)	Kurtosis	Schiefe
Fach: Englisch											
Thematische Motivierung	1	17.2 (0.938)	34.3 (0.949)	36.6 (1.039)	11.9 (0.674)	2.43 (0.026)	0.91	7303	13.3	-0.818	-0.002
	2	18.9 (1.177)	35.5 (0.847)	36.1 (1.210)	9.6 (0.574)	2.36 (0.029)	0.89	7278	13.6	-0.798	0.029
Verständlichkeit	1	7.8 (0.505)	24.3 (0.734)	50.4 (0.889)	17.5 (0.522)	2.77 (0.016)	0.82	7298	13.4	-0.296	-0.378
	2	8.0 (0.522)	25.7 (0.798)	51.2 (0.901)	15.0 (0.682)	2.73 (0.019)	0.81	7289	13.5	-0.271	-0.359
	3	9.1 (0.704)	19.7 (0.792)	45.1 (0.830)	26.1 (1.073)	2.88 (0.027)	0.90	7290	13.5	-0.483	-0.488
	4	10.8 (0.834)	20.0 (0.908)	44.8 (0.963)	24.5 (1.119)	2.83 (0.031)	0.92	7399	12.2	-0.575	-0.462
Schülerorientierung	1	11.2 (0.804)	30.7 (0.797)	49.1 (1.062)	9.0 (0.559)	2.56 (0.023)	0.81	7237	14.1	-0.413	-0.303
	2	9.9 (0.683)	24.1 (0.817)	50.4 (0.935)	15.6 (0.633)	2.72 (0.021)	0.84	7226	14.2	-0.373	-0.387
	3	10.1 (0.663)	24.5 (0.785)	50.8 (0.989)	14.7 (0.679)	2.70 (0.020)	0.84	7229	14.2	-0.347	-0.394
	4	8.8 (0.654)	22.1 (0.807)	51.9 (0.957)	17.2 (0.691)	2.77 (0.022)	0.83	7226	14.2	-0.243	-0.441
	5	10.5 (0.754)	24.9 (0.797)	49.4 (0.952)	15.3 (0.704)	2.69 (0.022)	0.85	7193	14.6	-0.429	-0.360
Strukturiertheit	1	28.2 (1.045)	36.9 (0.873)	29.1 (1.039)	5.8 (0.605)	2.13 (0.027)	0.89	7364	12.6	-0.852	0.248
	2	23.3 (0.946)	40.2 (0.793)	31.1 (0.914)	5.5 (0.461)	2.19 (0.022)	0.85	7360	12.6	-0.751	0.157
	3	16.3 (0.863)	30.8 (0.757)	43.2 (1.013)	9.7 (0.517)	2.46 (0.025)	0.88	7325	13.0	-0.734	-0.182
Klassenführung	1	13.8 (0.899)	24.4 (0.991)	44.7 (1.021)	17.2 (0.870)	2.65 (0.029)	0.92	7177	14.8	-0.716	-0.298
	2	11.6 (0.808)	31.4 (0.851)	42.2 (1.015)	14.8 (0.809)	2.60 (0.024)	0.88	7170	14.9	-0.653	-0.161

[76] Zur Berechnung des Mittelwerts, der Standardabweichung, der Kurtosis und der Schiefe wurden folgende Kategorien*scores* zugrunde gelegt: (1) stimmt gar nicht, (2) stimmt eher nicht, (3) stimmt eher, (4) stimmt ganz genau. Die Anzahl gültiger Werte sowie der relative Anteil fehlender Werte (hier wurden die „*missing value*"-Kategorien „nicht bearbeitet", „nicht bearbeitbar" und „nicht valide" zusammengefasst) sind ungewichtete Angaben und beziehen sich entsprechend auf die Stichprobe. Die gewichtete Kurtosis und Skewness wurde in Mplus mithilfe der OUTPUT-Option „TECH12" für Mischverteilungsmodelle realisiert. Dazu wurde ein „pseudo"-Mischverteilungsmodell mit einer Klasse spezifiziert. Da alle Variablen gleichzeitig in einem Modell (alle Interkorrelationen wurden spezifiziert) mithilfe der MISSING-Option geschätzt wurden, beziehen sich die Angaben auf N = 7930 Schülerinnen und Schüler.

[77] Die in Klammern angegebenen Standardfehler wurden in WesVar (Westat, 2002) mittels Fay-*replicates* (k = 0.3) ermittelt (Fay, 1989). Dieses Verfahren wurde im Projekt DESI als Standard verwendet.

[78] Standardfehler in Klammern (zur Berechnung der Standardfehler vgl. Fußnote 77).

Tabelle 13 (fortgesetzt)

		prozentuale Kategorienhäufigkeit (gewichtet)[77]							fehlende Angaben		
	Item	1	2	3	4	M[78]	SD	N	(Prozent)	Kurtosis	Schiefe
Fach: Deutsch											
Thematische Motivierung	1	18.1 (1.011)	33.2 (0.835)	36.4 (0.934)	12.2 (0.645)	2.43 (0.026)	0.92	7426	11.8	-0.872	-0.009
	2	19.9 (1.198)	36.0 (0.946)	35.2 (1.185)	9.0 (0.632)	2.33 (0.031)	0.89	7423	11.9	-0.814	0.050
Verständlichkeit	1	8.2 (0.604)	19.7 (0.760)	53.4 (0.720)	18.8 (0.706)	2.83 (0.020)	0.83	7409	12.0	-0.113	-0.522
	2	8.9 (0.632)	27.5 (0.939)	49.3 (0.939)	14.2 (0.724)	2.69 (0.023)	0.82	7413	12.0	-0.368	-0.320
	3	9.4 (0.836)	17.8 (0.818)	47.8 (0.856)	25.0 (1.061)	2.89 (0.030)	0.89	7396	12.2	-0.345	-0.555
	4	11.4 (0.902)	19.1 (0.807)	44.9 (1.000)	24.5 (1.174)	2.83 (0.031)	0.93	7503	10.9	-0.562	-0.494
Schülerorientierung	1	10.8 (0.837)	25.9 (0.889)	51.9 (1.085)	11.4 (0.622)	2.64 (0.024)	0.82	7363	12.6	-0.326	-0.403
	2	9.1 (0.748)	21.2 (1.015)	49.7 (1.097)	20.0 (0.778)	2.81 (0.025)	0.86	7379	12.4	-0.328	-0.462
	3	9.1 (0.747)	21.7 (0.772)	49.9 (0.968)	19.3 (0.754)	2.79 (0.025)	0.86	7338	12.9	-0.334	-0.448
	4	9.4 (0.647)	20.3 (0.797)	49.2 (0.751)	21.2 (0.857)	2.82 (0.023)	0.87	7382	12.4	-0.328	-0.489
	5	10.5 (0.760)	22.2 (0.791)	49.4 (0.932)	18.0 (0.725)	2.75 (0.023)	0.87	7330	13.0	-0.388	-0.437
Strukturiertheit	1	25.4 (0.783)	35.9 (0.727)	31.9 (0.878)	6.8 (0.505)	2.20 (0.020)	0.90	7449	11.6	-0.893	0.156
	2	22.8 (0.933)	38.5 (0.755)	31.7 (0.877)	7.0 (0.453)	2.23 (0.020)	0.88	7437	11.7	-0.794	0.152
	3	17.8 (0.912)	29.6 (0.738)	42.8 (1.034)	9.8 (0.450)	2.45 (0.023)	0.89	7410	12.0	-0.817	-0.169
Klassenführung	1	14.6 (1.016)	24.2 (0.944)	44.5 (1.051)	16.6 (0.896)	2.63 (0.028)	0.93	7349	12.8	-0.743	-0.307
	2	13.3 (1.011)	30.1 (0.935)	40.9 (1.132)	15.6 (0.671)	2.59 (0.027)	0.91	7336	12.9	-0.753	-0.168

Tabelle 14: Verteilungen der Kategorien der Indikatoren (inkl. Normalverteilungskurven)

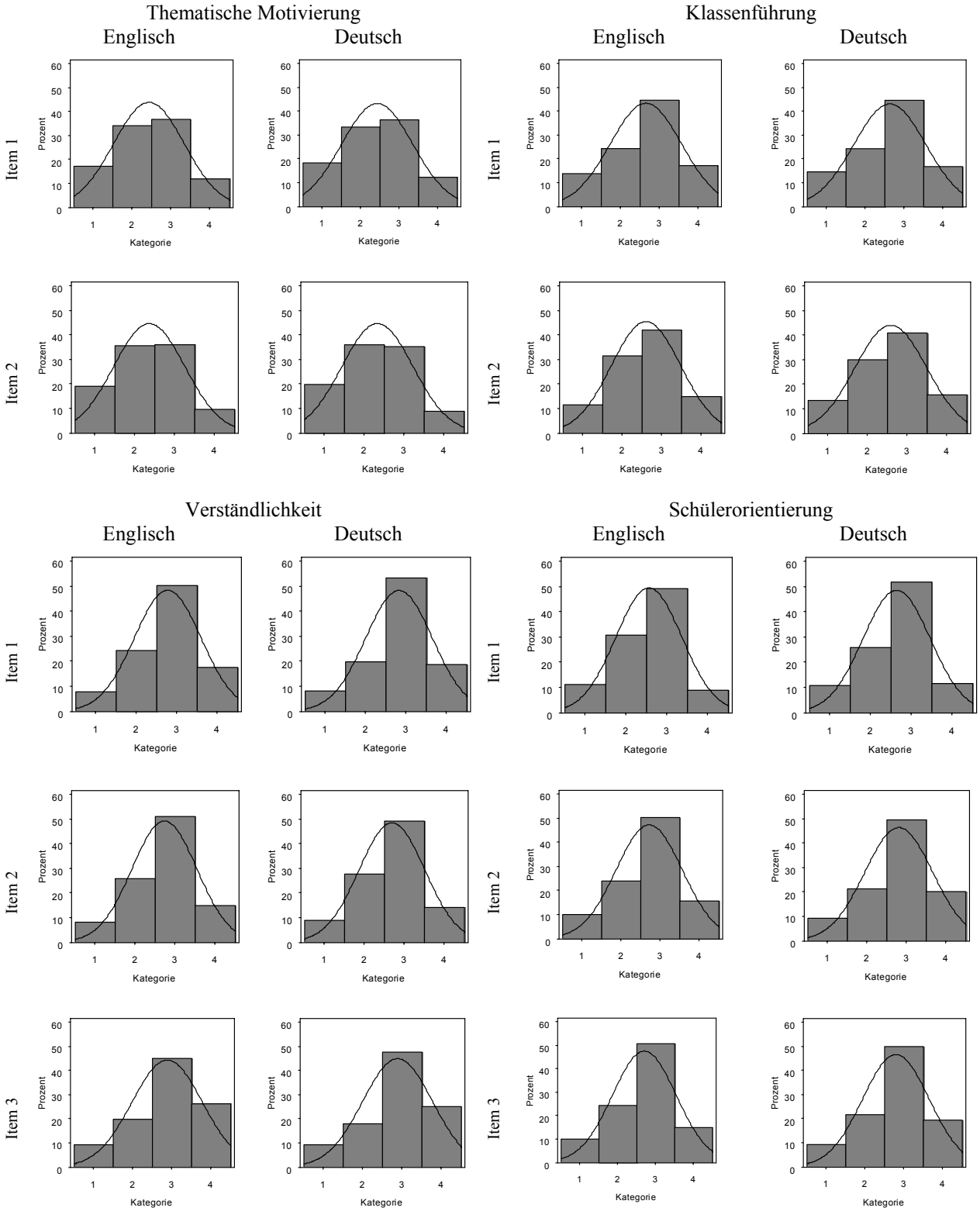

Tabelle 14 (fortgesetzt)

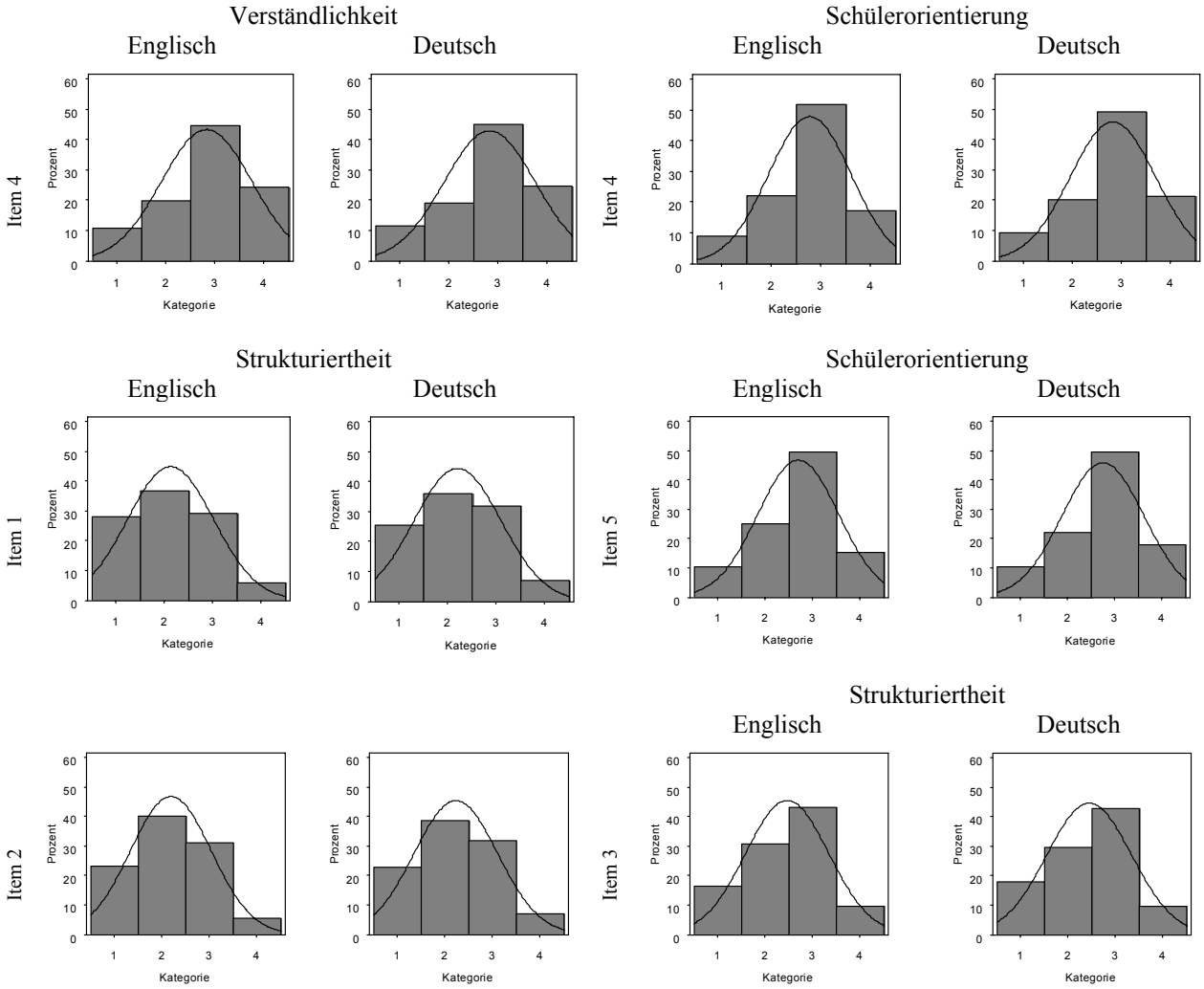

Tabelle 15: Reliabilität der Skalen bzw. Faktor*scores* zur Unterrichtswahrnehmung aus Schülersicht (Ein-Ebenen-Analyse)

		Cronbachs α [79]					*factor determinacy*[80]		quadrierte *factor determinacy*	
		Englisch			Deutsch					
Unterrichtsmerkmal	N	Stichprobe	Population	N	Stichprobe	Population	Englisch	Deutsch	Englisch	Deutsch
Thematische Motivierung	7225	.80	.79	7383	.80	.80	.92	.93	.85	.86
Verständlichkeit	6974	.80	.80	7108	.82	.81	.93	.94	.87	.89
Schülerorientierung	6974	.88	.88	7100	.88	.88	.95	.95	.89	.90
Strukturiertheit	7265	.77	.77	7341	.79	.79	.91	.91	.82	.84
Klassenführung	7127	.78	.77	7295	.78	.77	.90	.90	.81	.81

[79] Die Angaben beziehen sich auf die Rohwerte der Items.

[80] Die *factor determinacy* wurde anhand eines Ein-Ebenen-Modells mit allen Unterrichtsmerkmalen in beiden Fächern in Mplus (freie Faktorladungen, Faktorvarianzen auf 1 fixiert) berechnet und bezieht sich auf die (gewichtete) Teilstichprobe (N = 5620) der Schülerinnen und Schüler, bei denen zu allen Items der hier untersuchten Unterrichtsmerkmale in den Fächern Deutsch und Englisch gültige Angaben vorlagen. Im Falle fehlender Werte können sich jeweils alle Koeffizienten verändern (d.h. sie können niedriger ausfallen), so dass im Prinzip für jedes *missing value pattern* (insgesamt gab es in der untersuchten Stichprobe 739 solcher *patterns*) angegeben werden können. Auf die Berücksichtigung der *cluster*-Struktur der Daten (TYPE = COMPLEX) musste hier verzichtet werden, da Mplus sonst keine *factor determinacy* berechnet.
Modellgüte-Indizes: $Chi^2 = 2346.215$, Scaling Correction Factor = 1.844; DF = 419; CFI = .968; TLI = .962; RMSEA = .024; SRMR = .022

In Tabelle 16 sind die Verteilungen, Mittelwerte, Streuungen und die Zahl der vorliegenden Schülerangaben bezüglich der in kategorialer Form vorliegenden Prädiktoren angegeben. Bei den Schulnoten zeigt sich, dass die beiden Extrema – „sehr gut“ bzw. „ungenügend“ – relativ selten vergeben werden: Nur 2.4% (Deutschnote, Ende achte Klassenstufe) bis maximal 4.1% (erwartete Englischnote, Ende neunte Klassenstufe) der Schülerinnen und Schüler erhalten die Note „sehr gut“. Die Note „ungenügend“ wird mit 0.2% (Deutschnote, Ende achte Klassenstufe) bis 0.5% (Englisch- bzw. Deutschnote, Ende der neunten Klassenstufe) sogar noch seltener vergeben. Der Notendurchschnitt liegt in allen Fällen relativ nahe bei drei, also unterhalb des theoretischen Mittelwerts der Skala von M = 3.5, die Streuung knapp unterhalb einer Notenstufe.

Tabelle 16: Verteilungen, Mittelwerte, Streuungen und Anzahl gültiger Werte der kategorialen Prädiktoren[81]

	prozentuale Kategorienhäufigkeit[82]								
Prädiktor	1	2	3	4	5	6	M	SD	N
Englischnote	3.5	22.4	39.7	28.9	5.3	0.3	3.11	0.94	7595
(Ende 8. Klassenstufe)	(0.261)	(0.849)	(0.802)	(0.822)	(0.385)	(0.094)	(0.023)		
Englischnote	4.1	22.7	44.0	24.9	3.8	0.5	3.03	0.91	7347
(Ende 9. Klassenstufe)	(0.391)	(0.758)	(0.922)	(0.874)	(0.291)	(0.083)	(0.021)		
Deutschnote	2.4	24.0	46.4	24.5	2.5	0.2	3.02	0.84	7610
(Ende 8. Klassenstufe)	(0.285)	(0.859)	(0.772)	(0.757)	(0.311)	(0.078)	(0.019)		
Deutschnote	3.3	25.9	49.3	19.6	1.5	0.5	2.91	0.83	7352
(Ende 9. Klassenstufe)	(0.357)	(0.898)	(0.950)	(0.963)	(0.172)	(0.086)	(0.021)		
Globalurteil	15.9	23.3	46.2	14.6			2.60	0.92	7070
(Englischlehrkraft)	(1.054)	(0.801)	(1.181)	(0.729)	–	–	(0.029)		
Globalurteil	16.8	24.1	43.8	15.3			2.58	0.94	7231
(Deutschlehrkraft)	(1.140)	(0.907)	(1.123)	(0.874)	–	–	(0.034)		

Bei den auf die Lehrkraft bezogenen Globalurteilen halten sich völlige Ablehnung bzw. Zustimmung in etwa die Waage. Die Mittelwerte liegen mit M ≈ 2.6 knapp oberhalb des theoretischen Mittelwerts von M = 2.5. Die Angaben verteilen sich um den Mittelwert mit einer Streuung von etwas weniger als einer „Antwortalternative“.

Tabelle 17 können Angaben zu den hier verwendeten DESI-Testleistungen entnommen werden. Die *scores* sind in der Gesamtstichprobe auf die Metrik M = 500, SD = 100 für den jeweils zweiten Messzeitpunkt normiert. Die Leistungen am Ende der neunten Klassenstufe in der hier verwendeten Teilstichprobe liegen demnach etwa eine siebtel Standardabweichung über der Populationsschätzung aus der Gesamtstichprobe. Auch die Streuungen sind hier minimal eingeschränkt.

[81] Die Mittelwerte und Standardabweichungen basieren auf den Kategorien*scores* (vgl. dazu Fußnote 82). Standardfehler sind in Klammern angegeben (zur Berechnung der Standardfehler vgl. Fußnote 77).

[82] Die auf die Noten bezogenen Kategorien entsprechen den Schulnoten: (1) sehr gut, (2) gut, (3) befriedigend, (4) ausreichend, (5) mangelhaft, (6) ungenügend. Die auf das Globalurteil bezogenen Kategorien sind wie folgt zugeordnet: (1) stimmt gar nicht, (2) stimmt eher nicht, (3) stimmt eher, (4) stimmt ganz genau. Standardfehler sind jeweils in Klammern angegeben (zur Berechnung der Standardfehler vgl. Fußnote 77).

Tabelle 17: Mittelwerte und Streuungen der testbasierten Leistungsindikatoren[83]

Leistungstest	M	SD
Englisch (Anfang 9. Klassenstufe)	490.78	90.43
	(3.233)	
Englisch (Ende 9. Klassenstufe)	514.05	95.81
	(3.307)	
Englisch (Zuwachs)	23.27	43.35
	(1.081)	
Deutsch (Anfang 9. Klassenstufe)	479.02	91.48
	(3.473)	
Deutsch (Ende 9. Klassenstufe)	514.77	94.50
	(3.449)	
Deutsch (Zuwachs)	35.75	52.94
	(1.690)	

4.2 Übereinstimmungen von Schülerwahrnehmungen des Unterrichts

Wie in Kapitel 3.4.1 bereits erwähnt, lassen sich die relativen Übereinstimmungen der Urteile von Schülerinnen und Schülern innerhalb von Klassen anhand von ICCs bestimmen. Die im Folgenden dargestellten Analysen beziehen sich dabei zunächst auf einzelne Items. Die Analysen erfolgten in Mplus (Version 4.2, L. K. Muthén & Muthén, 2006) mithilfe des MODEL CONSTRAINT-Kommandos, das zur Definition der ICC verwendet wurde. Es wurden sowohl Analysen für metrische (intervallskalierte) als auch für ordinale Daten berechnet. Als *link*-Funktion für die Berechnung der ordinalen ICCs wurde – aufgrund ihrer weiten Verbreitung in der praktischen Anwendung der IRT – die Logit-*link*-Funktion verwendet[84]. Bei ordinalen Daten ist die Varianz innerhalb von Klassen eine Konstante. Diese Varianz entspricht der Varianz der logistischen Verteilung $\pi^2/3 \approx 3.290$ (vgl. Hox, 2002; Grilli & Rampichini, 2007). De Boeck und Wilson (2004) weisen jedoch darauf hin, dass die Varianz einer Standardnormalverteilung durch $[(15/16)(\pi/\sqrt{3})]^2 \approx 2.891$ besser approximiert wird (vgl. auch Snijders & Bosker, 1999). Entsprechend wurde die Varianz innerhalb von Klassen auf 2.891 festgelegt.

[83] Die Angaben basieren jeweils auf fünf *plausible values*. Zur Berechnung der in Klammern angegebenen Standardfehler vgl. Fußnote 77.

[84] Im Allgemeinen sind die Unterschiede in den Ergebnissen bezüglich verschiedener *link*-Funktionen allerdings vernachlässigbar (vgl. z.B. Grilli & Rampichini, 2007).

Für die so geschätzten ICCs gibt Mplus t-Werte an, die zur Bestimmung der Signifikanz herangezogen werden können. Da die ICC per Definition im Wertebereich[85] zwischen $0 \leq ICC \leq 1$ liegt, wurde einseitig getestet. Alle in Tabelle 18 dargestellten ICCs sind hoch signifikant ($p < .001$).

Tabelle 18: Interrater-Reliabilität: ICCs (metrisch und ordinal) auf Itemebene (in Prozent)

	Itemnummer	ICC in Prozent			
	(I) = Ich-Bezug	Englisch		Deutsch	
Unterrichtsmerkmal	(K) = Klassen-Bezug	metrisch	ordinal	metrisch	ordinal
Thematische Motivierung	1 (I)	16.2	18.9	15.6	18.3
	2 (K)	19.7	23.2	23.4	27.7
Verständlichkeit	1 (I)	12.6	15.0	12.1	14.5
	2 (K)	12.0	14.8	15.7	19.2
	3 (K)	19.2	21.9	23.3	26.9
	4 (I)	21.7	25.0	25.4	29.7
Schülerorientierung	1 (K)	18.5	22.6	19.2	23.7
	2 (K)	14.9	18.0	18.1	21.5
	3 (K)	13.3	16.1	16.9	19.8
	4 (I)	14.4	17.0	15.2	18.1
	5 (I)	12.9	15.6	16.5	19.6
Strukturiertheit	1 (K)	13.2	15.7	12.6	15.1
	2 (K)	10.2	12.3	14.8	18.3
	3 (K)	9.9	11.7	13.3	16.0
Klassenführung	1 (K)	20.9	25.0	26.0	30.9
	2 (K)	17.4	21.1	22.2	27.2
Mittelwert		15.4	18.4	18.1	21.7

Die ICCs auf Itemebene liegen im Bereich zwischen etwa 10% bis 26%, wenn man die Itemantworten – wie in diesem Zusammenhang üblich – als metrische Daten betrachtet, die hier die ordinalen ICCs um bis zu etwa 19% (*Strukturiertheit*, Item 2, Fach Deutsch) unterschätzen. Verglichen mit anderen Studien liegen die ICCs hier in einem mittleren bis sehr hohen Bereich: Lüdtke et al. (2006) berichten ICCs im Bereich von 6% bis 20% für sechs verschiedene *Skalen* zur Unterrichtswahrnehmung aus Schülersicht. Die Interrater-Reliabilitäten von Klimaskalen im Bereich der Organisationspsychologie liegen – so Bliese (2000) – üblicherweise zwischen 5% und 20% mit einer Obergrenze von 30%.

Insgesamt ist die Interrater-Reliabilität im Fach Deutsch etwas höher als im Fach Englisch (mittlere metrische ICC etwa 18% im Fach Deutsch, 15% im Fach Englisch). Dieses Ergebnis könnte darauf hinweisen, dass entweder die Items im Fach Deutsch mit größerer absoluter Übereinstimmung beurteilt werden, oder aber, dass es im Deutschunterricht größere Un-

[85] Da bei der Berechnung der ICC für ordinale Variablen die Varianz innerhalb von Klassen eine konstante Größe ist, approximiert die ICC hier nur den Wert 1, d.h. es gilt $0 \leq ICC < 1$.

terschiede bezüglich der hier erfassten Unterrichtsmerkmale gibt. Diese Frage wird in Kapitel 4.4 vertieft.

Interessanterweise sind die ICCs der Items mit Ich-Bezug in vergleichbarer Größenordnung wie die der auf die Klasse bezogenen Items. Dies kann als Hinweis auf eine eher untergeordnete Rolle des Adressaten hinsichtlich der Itemformulierung interpretiert werden.

Auf der Basis der Kategorienwahrscheinlichkeiten der Indikatoren (diese entsprechen den in Kapitel 4.1, Tabelle 13 dargestellten relativen Kategorienhäufigkeiten[86]) lassen sich anhand des von O'Brien vorgeschlagenen Verfahrens (vgl. Kap. 3.4.2) die Korrelationen der arbiträren Kategorien*scores* mit den hypothetisch zugrunde liegenden intervallskalierten und normalverteilten Variablen berechnen (vgl. Tabelle 19). Es zeigen sich auch hier – analog zu den Ergebnissen der Simulationsstudie (s. Kap. 3.4.3) – relativ geringe Differenzen zwischen den Korrelationen der optimalen bzw. der arbiträren *scores* mit der latenten Variable. Das heißt, auch hier trägt das *scoring* (arbiträre Werte von eins bis vier) nur unwesentlich zur *grouping error*-Varianz bei.

Tabelle 19: Korrelationen der optimal bzw. arbiträr kategorisierten Variablen mit den (hypothetisch) zugrunde liegenden intervallskalierten, normalverteilten Variablen

		Englisch		Deutsch	
Unterrichtsmerkmal	Item	r_{ku}	r_g	r_{ku}	r_g
Thematische Motivierung	1	.938	.937	.938	.937
	2	.936	.935	.935	.935
Verständlichkeit	1	.926	.924	.921	.916
	2	.925	.923	.928	.926
	3	.927	.923	.925	.919
	4	.929	.925	.929	.924
Schülerorientierung	1	.924	.922	.922	.919
	2	.927	.924	.927	.923
	3	.926	.923	.927	.923
	4	.924	.920	.927	.922
	5	.928	.926	.928	.924
Strukturiertheit	1	.928	.927	.931	.930
	2	.930	.929	.933	.932
	3	.932	.931	.932	.930
Klassenführung	1	.934	.931	.934	.931
	2	.936	.935	.937	.936

Anhand der in Tabelle 19 dargestellten Korrelationen der arbiträren Kategorien*scores* mit den jeweiligen latenten Variablen lassen sich (1) die ICCs der als intervallskaliert behandelten kategorisierten Variablen adjustieren (vgl. Kap. 3.4.2, (22)) sowie (2) die zur Adjustierung der Indikator-Reliabilitäten auf Ebene 1 benötigten quadrierten Korrelationen der kategori-

[86] Da sich diese Kategorienhäufigkeiten nur auf vorliegende valide Angaben beziehen, wird hier implizit davon ausgegangen, dass die fehlenden Daten nicht systematisch zustande kommen (*missing completely at random*, MCAR).

sierten Variablen mit den zugrunde liegenden intervallskalierten, normalverteilten Variablen *innerhalb* von Klassen bestimmen (vgl. Kap. 3.4.2, (21)). Die Ergebnisse sind in Tabelle 20 dargestellt. Bezüglich der adjustierten metrischen ICCs zeigen sich nur noch relativ geringe Unterschiede zu den in Tabelle 18 berichteten ordinalen ICCs: Betrachtet man die ordinalen ICCs als die wahren Werte, dann liegt der *bias* ihrer adjustierten metrischen Pendants im Bereich von Unterschätzungen bis maximal 7.1% (zweites Item der Skala *Strukturiertheit* im Fach Deutsch) und Überschätzungen bis zu 2.6% (drittes Item der Skala *Verständlichkeit* im Fach Deutsch) mit einer mittleren Unterschätzung von 2.4%.

Tabelle 20: Adjustierte metrische ICCs und gemeinsame Varianzanteile der kategorisierten Variablen mit den zugrunde liegenden intervallskalierten, normalverteilten Variablen innerhalb von Klassen

		ICC (adjustiert)		r_{gw}^2	
Unterrichtsmerkmal	Item	Englisch	Deutsch	Englisch	Deutsch
Thematische Motivierung	1	18.4	17.8	.855	.856
	2	22.5	26.8	.844	.835
Verständlichkeit	1	14.8	14.4	.832	.817
	2	14.1	18.3	.831	.831
	3	22.5	27.6	.818	.798
	4	25.4	29.8	.816	.804
Schülerorientierung	1	21.8	22.7	.816	.808
	2	17.5	21.3	.828	.819
	3	15.6	19.8	.829	.822
	4	17.0	17.9	.821	.824
	5	15.1	19.3	.835	.825
Strukturiertheit	1	15.4	14.6	.839	.846
	2	11.8	17.0	.848	.846
	3	11.4	15.4	.851	.845
Klassenführung	1	24.1	30.0	.831	.819
	2	19.9	25.3	.848	.841

Die auf die Varianz innerhalb von Klassen adjustierte gemeinsame Varianz der kategorisierten und der zugrunde liegenden intervallskalierten, normalverteilten Variable (r_{gw}^2) liegt im Bereich zwischen 79.8% (Item 3 der Skala *Verständlichkeit* im Fach Deutsch) und 85.6% (Item 1 der Skala *Thematische Motivierung* im Fach Deutsch). Der Fehleranteil, der auf die Kategorisierung der Variablen zurückgeht, variiert also relativ wenig über die Items hinweg.

Im Folgenden geht es um die Frage, ob die Annahme isomorpher Konstrukte auf Individual- und Klassenebene gerechtfertigt ist, oder ob sich die Konstrukte auf beiden Ebenen nur „ähnlich" sind (vgl. dazu Kap. 2.1.4.2). Dazu wurde in einer Reihe von Analysen mittels Wald-Test die Messinvarianz[87] (im Sinne itemspezifisch identischer Faktorladungen) der jeweiligen Faktoren auf beiden Ebenen überprüft. Da pro Fach jeweils zwei Faktoren durch nur

[87] Streng genommen lässt sich Isomorphie hier lediglich falsifizieren, nicht aber verifizieren durch Nachweis der Messinvarianz.

zwei Indikatoren repräsentiert sind (*Thematische Motivierung* und *Klassenführung*), ist in diesen Fällen eine solche Überprüfung nicht unmittelbar möglich, da zur Modellidentifikation hier beide Ladungen fixiert werden müssen[88] (z.B. auf den Wert 1). Deshalb wurden hier jeweils fachübergreifend alle fünf Unterrichtsmerkmale in die Analysen einbezogen. Dabei wurden alle Faktorladungen frei geschätzt (mit Ausnahme der ersten Ladung, die jeweils zur Skalierung der latenten Variable auf den Wert 1 fixiert wurde) und die Restriktion itemspezifisch identischer Ladungen auf beiden Ebenen bei jeweils einem bestimmten Unterrichtsmerkmal per Wald-Test überprüft[89].

Die Ergebnisse dieser Analysen sind in Tabelle 21 dargestellt. Die beiden Merkmale *Strukturiertheit* und *Klassenführung* können in beiden Fächern über die Ebenen hinweg als messinvariant betrachtet werden (die p-Werte für die Chi^2-verteilte Teststatistik des Wald-Tests mit den entsprechenden Freiheitsgraden liegen jeweils oberhalb des 5%-Niveaus).

Da die ersten drei Unterrichtsmerkmale aus – bezüglich des Adressaten – „gemischten" Items (Ich-Bezug sowie Klassen-Bezug) bestehen, wurde zusätzlich überprüft, ob die jeweiligen Subskalen isomorphe Konstrukte darstellen. Im Falle der Thematischen Motivierung ist dies jedoch nicht möglich, da das Merkmal nur durch zwei Items erfasst wurde (jeweils ein Item pro Adressat des Lehrerverhaltens). Bei den beiden übrigen Unterrichtsmerkmalen wurde lediglich für *Schülerorientierung* mit Ich-Bezug im Fach Englisch sowie für *Schülerorientierung* mit Klassen-Bezug im Fach Deutsch ein identisches Messmodell auf beiden Ebenen bestätigt. Das heißt, der Adressatbezug der Items spielt bezogen auf die Messinvarianz dieser Merkmale keine entscheidende – oder zumindest keine eindeutige – Rolle. Offensichtlich handelt es sich bei diesen Unterrichtsmerkmalen um eher analoge Konstrukte (im Sinne des *fuzzy composition process*- Modells; vgl. Kap. 2.1.4.2) auf beiden Ebenen.

[88] Alternativ dazu können auch die Faktorvarianz auf einen beliebigen Wert fixiert und die beiden Faktorladungen auf jeder Ebene jeweils gleichgesetzt werden. Im Prinzip werden hier zwar zwei verschiedene Faktoren spezifiziert, jeweils ein Faktor pro Ebene. Da aber alle Variablen auf beiden Faktoren per Definition unkorreliert sind, gelten die üblichen Identifikationsregeln (vgl. Bollen, 1989) hier ebenenspezifisch.

[89] Modellgüte-Indizes: $Chi^2 = 4512.159$, Scaling Correction Factor = 1.028; DF = 841 ; CFI = .961; TLI = .953; RMSEA = .023; SRMR (zwischen) = .054; SRMR (innerhalb) = .022. Insgesamt wurden drei negative Residualvarianzen auf Klassenebene auf Null fixiert (jeweils auf das Fach Deutsch bezogene Indikatoren): Item 1 der Skala *Thematische Motivierung* ($p < .05$), Item 3 der Skala *Schülerorientierung* (nicht signifikant) und Item 3 der Skala *Strukturiertheit* (nicht signifikant). Die Fixierung solcher *Heywood-Cases* ist bei KFAs auf Klassenebene aufgrund der hohen Reliabilitäten der Indikatoren häufig erforderlich. Im Allgemeinen ist eine solche Restriktion unbedenklich, insbesondere wenn es sich um Restriktionen nicht signifikanter Residualvarianzen handelt (vgl. Dillon, Kumar & Mulani, 1987). Signifikante Heywood-Cases können u.U. auf eine Überfaktorisierung (*overfactoring*) des Modells hinweisen. Dann sollten sparsamere Modelle (mit weniger Faktoren) getestet werden (vgl. dazu Kap. 4.3).

Tabelle 21: Überprüfung der Messinvarianz der Faktoren (identische itemspezifische unstandardisierte Ladungen) auf beiden Ebenen: Wald-Tests

	Unterrichtsfach				
	Englisch		Deutsch		
Faktor	Koeffizient	p[90]	Koeffizient	p	DF[91]
Thematische Motivierung	6.398	0.011	66.477	0.000	1
Verständlichkeit	76.373	0.000	54.940	0.000	3
Schülerorientierung	42.056	0.000	39.132	0.000	4
Strukturiertheit	1.933	**0.380**	4.848	**0.089**	2
Klassenführung	1.649	**0.199**	1.440	**0.230**	1
Subskalen: Items mit Ich-Bezug[92]					
Verständlichkeit	12.141	0.001	27.874	0.000	1
Schülerorientierung	2.709	**0.100**	4.425	0.035	1
Subskalen: Items mit Klassen-Bezug[93]					
Verständlichkeit	15.392	0.000	24.346	0.000	1
Schülerorientierung	27.034	0.000	3.442	**0.179**	2

Da mit zunehmendem Stichprobenumfang auch relativ trivial fehlspezifizierte Modelle auf der Basis von Gütekriterien zum Vergleich zweier Modelle verworfen werden (vgl. La Du & Tanaka, 1995; Widaman & Reise, 1997; B. O. Muthén, 1994), wurden zusätzlich die Unterschiede zwischen den unstandardisierten Ladungen auf Klassenebene und den entsprechenden Ladungen innerhalb von Klassen aus dem frei geschätzten Modell (jeweils nur erste Ladung auf 1 fixiert) inspiziert (Tabelle 22). Bei den Unterrichtsmerkmalen, die als ebenenübergreifend messinvariant identifiziert wurden (vgl. Tabelle 21) – also *Strukturiertheit* und *Klassenführung* in beiden Fächern – liegen die prozentualen Abweichungen der klassenspezifischen Faktorladungen im Bereich von -9.1% (*Strukturiertheit* Englisch, Item 2) bis +12.2% (*Strukturiertheit* Deutsch, Item 3). Bei den übrigen Unterrichtsmerkmalen liegen die prozentualen Abweichungen – absolut betrachtet – im Mittel deutlich höher. Eine Ausnahme stellt die Skala *Thematische Motivierung* Englisch dar, bei der ebenfalls nur relativ geringe Abweichungen vorliegen (das zeigt sich auch an dem relativ geringen Koeffizienten des Wald-Tests bei dieser Skala; s. Tabelle 21). Bei allen anderen Skalen mit signifikanten Unterschieden be-

[90] Nicht signifikante p-Werte (5%-Niveau) fett gedruckt.
[91] Degrees of freedom (Freiheitsgrade)
[92] Die Items mit Klassen-Bezug des jeweils untersuchten Merkmals wurden – aufgrund ihres hohen Zusammenhangs mit den ichbezogenen Items des jeweiligen Faktors – in diesen Analysen nicht berücksichtigt. Bei zwei Modellen mussten zusätzlich negative Residualvarianzen auf Null fixiert werden (über die bereits im Gesamtmodell vorgenommenen Fixierungen hinaus): *Verständlichkeit* im Fach Deutsch (Item 4 der Skala *Verständlichkeit* Deutsch); *Schülerorientierung* im Fach Englisch (Item 5 der Skala *Schülerorientierung* Englisch)
[93] Die Items mit Ich-Bezug des jeweils untersuchten Merkmals wurden – aufgrund ihres hohen Zusammenhangs mit den klassenbezogenen Items des jeweiligen Faktors – in diesen Analysen nicht berücksichtigt. Bei dem Modell zur *Schülerorientierung* im Fach Englisch musste zusätzlich die negative Residualvarianz von Item 3 auf Null fixiert werden.

züglich der ebenenübergreifenden itemspezifischen Ladungen zeigen sich Differenzen in einer Größenordnung, die über die statistische Signifikanz hinaus auch als praktisch relevant betrachtet werden können.

Tabelle 22: Prozentuale Abweichung der unstandardisierten Ladungen: innerhalb von Klassen vs. Klassenebene

	Englisch			Deutsch		
	unstandardisierte Ladung[94]		prozentuale Abweichung[95]	unstandardisierte Ladung		prozentuale Abweichung
Unterrichtsmerkmal[96]	innerhalb (Ebene 1)	zwischen (Ebene 2)	(Ebene 2 vs. Ebene 1)	innerhalb (Ebene 1)	zwischen (Ebene 2)	(Ebene 2 vs. Ebene 1)
Thematische Motivierung						
Item 2	0.942	1.058	11.6	0.897	1.163	25.8
	(0.014)	(0.039)		(0.016)	(0.027)	
Verständlichkeit						
Item 2	1.138	1.250	9.4	1.079	1.387	25.0
	(0.032)	(0.088)		(0.026)	(0.107)	
Item 3	1.318	1.770	29.3	1.171	1.821	43.4
	(0.036)	(0.133)		(0.029)	(0.142)	
Item 4	1.113	1.977	55.9	1.036	1.985	62.8
	(0.036)	(0.137)		(0.032)	(0.171)	
Schülerorientierung						
Item 2	1.156	0.958	-18.7	1.136	1.038	-9.0
	(0.022)	(0.040)		(0.023)	(0.030)	
Item 3	1.128	0.918	-20.5	1.101	1.007	-8.9
	(0.020)	(0.030)		(0.021)	(0.026)	
Item 4	1.146	0.936	-20.2	1.204	0.926	-26.1
	(0.024)	(0.042)		(0.026)	(0.036)	
Item 5	1.177	0.918	-24.7	1.195	1.005	-17.3
	(0.024)	(0.038)		(0.022)	(0.034)	
Strukturiertheit						
Item 2	1.065	0.972	-9.1	1.052	1.120	6.3
	(0.022)	(0.063)		(0.021)	(0.053)	
Item 3	1.070	1.036	-3.2	1.035	1.170	12.2
	(0.027)	(0.059)		(0.025)	(0.055)	
Klassenführung						
Item 2	0.880	0.942	6.8	0.899	0.954	5.9
	(0.024)	(0.040)		(0.024)	(0.034)	

Analog zu Tabelle 22 sind in Tabelle 23 die, auf der Basis der in Tabelle 20 enthaltenen r_{gw}^2-Koeffizienten adjustierten, standardisierten Ladungen und Reliabilitäten der Indikatoren dargestellt. Die Reliabilitäten der Indikatoren – also die quadrierten standardisierten Ladungen – auf der Ebene innerhalb von Klassen liegen im Bereich zwischen 34.8% bis 62.9% bei

[94] Standardfehler sind jeweils in Klammern angegeben.

[95] Die Abweichung bezieht sich auf die Differenz der Ladungen auf Klassenebene und auf der Ebene innerhalb von Klassen im Verhältnis zum Mittelwert der beiden Ladungen. Der Mittelwert der Ladungen wurde deshalb als Referenz verwendet, da hier kein theoretisch plausibles Referenzmodell vorliegt: Es könnte sowohl das Modell innerhalb von Klassen als auch das Modell auf Klassenebene als Referenz verwendet werden. Da sich die prozentualen Abweichungen je nach Referenzmodell aber – wenn auch minimal – unterscheiden, wurden die mittleren Ladungen als Referenz verwendet.

[96] Da in den zugrunde liegenden Modellen jeweils die erste Ladung – also die Ladung des ersten Items – auf beiden Ebenen auf 1 fixiert wurde, sind diese jeweils identisch.

den unadjustierten Koeffizienten bzw. zwischen 41.8% und 75.7% *true score*-Varianzanteil bei den adjustierten. Dies zeigt, dass es sich hier um systematische Nicht-Übereinstimmungen bezüglich des Unterrichts handelt, denen eine analoge Faktorstruktur zu der auf Klassenebene zugrunde gelegt werden kann (und nicht um Abweichungen im Sinne von „Zufallsfehlern"). Auf der Klassenebene sind im Durchschnitt deutlich höhere Reliabilitäten als innerhalb von Klassen im Bereich von 55.8% bis 100% (die allerdings auch artifiziell aufgrund der Fixierung der Residualvarianzen auf Null zustande kommen) zu verzeichnen.

Besonders auffällig sind die durchweg niedrigen – unadjustierten sowie adjustierten, im Fach Englisch wie auch im Fach Deutsch, innerhalb von Klassen sowie auf Klassenebene – Reliabilitäten des ersten Items des Faktors *Verständlichkeit*. Dies ist möglicherweise auf die Referenz des Itemtextes auf „Aufgabenstellungen" im jeweiligen Fach zurückzuführen („Die Aufgabenstellungen im Englisch(Deutsch-)unterricht sind für mich klar und verständlich."). Es ist denkbar, dass in sprachlichen Fächern wie Englisch und Deutsch der Begriff „Aufgabenstellungen" weniger klar verständlich ist als beispielsweise in naturwissenschaftlichen Unterrichtsfächern oder in Mathematik. Daneben könnte hier auch die „diagnostische Kompetenz" der Schülerinnen und Schüler bezüglich des tatsächlichen Verstehens der Aufgaben bzw. Unterrichtsinhalte von Bedeutung sein. Dafür sprechen die ebenfalls vergleichsweise niedrigen Reliabilitäten auf der Ebene innerhalb von Klassen des vierten Items der Skala *Verständlichkeit* („Mein Englisch(Deutsch)lehrer/ meine Englisch(Deutsch)lehrerin erklärt so, dass ich es verstehen kann."), die jeweils nur knapp oberhalb des oben genannten Items liegen. Die Reliabilitäten auf Klassenebene sind hier jedoch deutlich größer als bei dem auf die Aufgabenstellungen bezogenen Item. Dies legt eine Interpretation im Sinne einer größeren Bedeutung der auf das *Verstehen* der von der Lehrkraft intendierten Unterrichtsinhalte bezogenen diagnostischen Kompetenz für die Itemspezifität auf Ebene 1 nahe, während der Verwendung des Begriffs „*Aufgabenstellungen*" insbesondere auf Klassenebene eine spezifische Bedeutung zukommt.

Tabelle 23: (Adjustierte) standardisierte Ladungen (λ_{std}) und Indikatorreliabilitäten (Rel.) innerhalb von Klassen und auf Klassenebene

	Englisch						Deutsch					
	innerhalb (Ebene 1)		innerhalb (Ebene 1), adjustiert		zwischen (Ebene 2)		innerhalb (Ebene 1)		innerhalb (Ebene 1), adjustiert		zwischen (Ebene 2)	
Unterrichtsmerkmal	λ_{std}	Rel.	λ_{std}	Rel.	λ_{std}	Rel.	λ_{std}	Rel.	λ_{std}	Rel.	λ_{std}	Rel.
Thematische Motivierung												
Item 1	.781	.610	.845	.714	.979	.958	.787	.619	.851	.724	1.000	1.000
Item 2	.771	.594	.839	.704	.947	.897	.767	.588	.839	.704	.968	.937
Verständlichkeit												
Item 1	.590	.348	.647	.418	.747	.558	.629	.396	.696	.484	.831	.691
Item 2	.682	.465	.748	.560	.958	.918	.690	.476	.757	.573	.982	.964
Item 3	.743	.552	.822	.675	.968	.937	.723	.523	.809	.655	.963	.927
Item 4	.623	.388	.690	.476	.988	.976	.630	.397	.703	.494	.982	.964
Schülerorientierung												
Item 1	.681	.464	.754	.568	.944	.891	.678	.460	.754	.569	.959	.920
Item 2	.739	.546	.812	.660	.977	.955	.729	.531	.806	.649	.989	.978
Item 3	.724	.524	.795	.632	.993	.986	.711	.506	.784	.615	1.000	1.000
Item 4	.743	.552	.820	.672	.991	.982	.757	.573	.834	.696	.975	.951
Item 5	.739	.546	.809	.654	.987	.974	.754	.569	.830	.689	.990	.980
Strukturiertheit												
Item 1	.665	.442	.726	.527	.819	.671	.683	.466	.742	.551	.889	.790
Item 2	.733	.537	.796	.634	.936	.876	.740	.548	.805	.647	.927	.859
Item 3	.707	.500	.766	.587	.983	.966	.710	.504	.773	.597	1.000	1.000
Klassenführung												
Item 1	.793	.629	.870	.757	.950	.903	.777	.604	.858	.737	.955	.912
Item 2	.706	.498	.767	.588	.989	.978	.696	.484	.759	.576	.987	.974

Auf der Basis des Gesamtmodells[97] (also mit allen fünf Unterrichtsmerkmalen pro Fach auf beiden Ebenen) wurden die auf die Kommunalitäten bezogenen ICCs berechnet. Diese sind im Sinne des *true score*-Varianzanteils auf Klassenebene bezogen auf die *true-score*-Gesamtvarianz des jeweiligen Items zu interpretieren. Die Ergebnisse sind Tabelle 24 zu entnehmen.

Bezogen auf das Fach Englisch variieren die ICCs hier zwischen 16.1% (zweites Item des Faktors *Strukturiertheit*) und 40.6% (viertes Item des Faktors *Verständlichkeit*). Im Fach Deutsch liegen die ICCs im Bereich von 18.0% (*Verständlichkeit*, erstes Item) und 44.7% (*Verständlichkeit*, viertes Item). Die auf die Kommunalitäten bezogenen ICCs liegen durchweg über den auf die Gesamtvarianz (also *true-score*- und Fehlervarianz) der Items bezogenen metrischen ICCs (vgl. Tabelle 18). Im Fach Englisch liegt der Mittelwert der kommunalitätsbezogenen ICCs um etwa 8, im Fach Deutsch um etwa 10 Prozentpunkte über den jeweiligen auf die Gesamtvarianz bezogenen ICCs. Es sei noch einmal darauf hingewiesen, dass

[97] Die Ladungen wurden frei geschätzt (nur die erste Ladung wurde jeweils auf 1 fixiert). Modellgüte-Indizes: vgl. Fußnote 89.

die ICCs der Kommunalitäten – auf der Basis metrischer Modelle – hier als sehr zuverlässige Approximationen der auf IRT-Modellen basierenden Schätzungen betrachtet werden können (vgl. dazu die Simulationsstudie in Kap. 3.4.3).

Tabelle 24: Interrater-Reliabilität bezüglich der messfehlerbereinigten Unterrichtswahrnehmungen auf Itemebene: ICCs der Kommunalitäten

Unterrichtsmerkmal	Itemnummer (I) = Ich-Bezug (K) = Klassen-Bezug	ICC (Kommunalität) in Prozent[98] Englisch	Deutsch
Thematische Motivierung	1 (I)	22.7	22.7
	2 (K)	27.0	33.1
Verständlichkeit	1 (I)	17.8	18.0
	2 (K)	20.7	26.6
	3 (K)	28.1	34.7
	4 (I)	40.6	44.7
Schülerorientierung	1 (K)	29.4	32.0
	2 (K)	22.2	28.2
	3 (K)	21.6	28.2
	4 (I)	21.7	21.7
	5 (I)	20.2	24.9
Strukturiertheit	1 (K)	18.7	19.6
	2 (K)	16.1	21.7
	3 (K)	17.8	23.8
Klassenführung	1 (K)	26.5	34.1
	2 (K)	29.2	36.9
Mittelwert		23.8	28.2

Für die als ebenenbezogen messinvariant zu betrachtenden Unterrichtsmerkmale lassen sich, wie in Kap. 3.4.1 erwähnt, latente ICCs berechnen, da die Konstrukte auf beiden Ebenen als isomorph betrachtet werden können, und die Faktorvarianzen – aufgrund identischer Metrik der Faktoren – miteinander vergleichbar sind. Die in Tabelle 25 dargestellten ICCs liegen im Bereich zwischen 17.1% (*Strukturiertheit* Englisch) und 35.4% (*Klassenführung* Deutsch).

[98] Die Angaben beziehen sich auf das Modell mit frei geschätzten Ladungen auf beiden Ebenen im jeweiligen Fach.

Tabelle 25: Interrater-Reliabilität bezüglich messfehlerbereinigter Unterrichtswahrnehmungen auf Konstruktebene: Latente ICCs der ebenenbezogen messinvarianten Faktoren[99]

	ICC der Faktoren in Prozent	
Unterrichtsmerkmal	Englisch	Deutsch
Strukturiertheit	17.1	22.1
Klassenführung	28.0	35.4
Schülerorientierung (Ich-Bezug)	20.1	-
Schülerorientierung (Klassen-Bezug)	-	27.9

Zur Überprüfung der Hypothese, ob Kommunikation innerhalb von Klassen, die im Wesentlichen geschlechtsspezifisch stattfinden sollte (vgl. dazu Kap. 2.1.4.2), zu einer Erhöhung der Übereinstimmung führt, wurden zunächst Zwei-Ebenen-Analysen mit über Klassen hinweg variierendem Regressionsgewicht für die Geschlechtszugehörigkeit (ein sogenanntes *random-slope*-Modell) berechnet. Eine erhöhte Übereinstimmung innerhalb der geschlechtshomogenen Gruppen sollte zu unterschiedlichen Gruppenmittelwerten (der Gruppe der Schülerinnen vs. der Gruppe der Schüler) innerhalb von Klassen führen, die sich in klassenspezifischen Regressionsgewichten für das Geschlecht ausdrücken sollten.

Die Ergebnisse in Tabelle 26 belegen, dass das Geschlecht tatsächlich eine bedeutsame Rolle bezogen auf die Unterrichtsbeurteilung spielt. Schülerinnen bewerten die hier erfassten Unterrichtsmerkmale meist deutlich positiver (Spalte: Rohwerte – Regressionsgewicht, Mittelwert; aufgrund der *dummy*-Kodierung für das Geschlecht ist der Wert hier als „mädchenspezifische Differenz" zu den Jungen zu interpretieren). Wenn man – wie hier – die Antwortskalierung als metrisch betrachtet, dann liegen die Angaben der Schülerinnen um bis zu etwa einem viertel „Antwortkästchen" (Item 1 der Skala *Thematische Motivierung* im Fach Deutsch), was fast einem Zehntel des theoretischen Wertebereichs (Skalierung der Antwortalternativen: 1–4) entspricht, über ihren männlichen Mitschülern. Dies entspricht einer viertel Standardabweichung (Spalte: z-standardisierte Werte – Regressionsgewicht, Mittelwert).

Kontrolliert wurde zusätzlich die Geschlechtszusammensetzung der Klassen anhand des Mädchenanteils. Dieser Zusammenhang (in Tabelle 26 nicht dargestellt) ist – mit Ausnahme der Analyse bezüglich des zweiten Items der Skala *Strukturiertheit* im Fach Deutsch (Regressionsgewicht bezogen auf Rohwerte = -.39; $p < .05$) – bei keinem Item signifikant. Das heißt,

[99] Dazu wurde ein Modell mit allen fünf Unterrichtsmerkmalen in beiden Fächern spezifiziert, bei dem die Ladungen der ebenenübergreifend als invariant zu betrachtenden Messmodelle gleich gesetzt wurden. Bei den Merkmalen mit Ich-Bezug bzw. Klassen-Bezug wurden die jeweiligen komplementären Items aus der Analyse ausgeschlossen. Die negative Residualvarianz von Item 5 der Skala *Schülerorientierung* (Ich-Bezug) Englisch wurde auf Null fixiert. Modellgüte-Indizes: $Chi^2 = 3154.797$, Scaling Correction Factor = 1.040; DF = 571; CFI = 0.963; TLI = 0.955; RMSEA = 0.024; SRMR (zwischen) = 0.055; SRMR (innerhalb) = 0.021

die geteilte Wahrnehmung des Unterrichts hängt – von einer Ausnahme abgesehen – nicht mit dem Mädchenanteil in der Klasse zusammen.

Die Standardabweichung der klassenspezifischen Regressionsgewichte für das Geschlecht ist bei allen Items signifikant. Sie beträgt bis zu 0.32 Einheiten (Item 4 der Skala *Verständlichkeit* im Fach Deutsch, mittleres Regressionsgewicht = 0.14), d.h. die Unterschiede zwischen Mädchen und Jungen innerhalb von Klassen liegen dort in 95% aller Klassen demnach im Bereich von $0.14 \pm 1.96 \cdot 0.32$, also zwischen -0.49 und 0.77 (d.h. die Einschätzungen der Mädchen liegen im Bereich zwischen etwa einem halben „Antwortkästchen" unterhalb bis etwa einem dreiviertel „Antwortkästchen" überhalb der Jungen).

Interessanterweise liegen die Streuungen der auf die z-standardisierten Variablen bezogenen Regressionsgewichte bei den Items mit Ich-Bezug in derselben Größenordnung wie bei den Items mit Klassen-Bezug. Wenn man davon ausgeht, dass bei Items mit Klassen-Bezug die Kommunikation mit Mitschülern bedeutsamer sein sollte, dann müsste sich dies hier in stärker geschlechtsspezifischen Wahrnehmungen innerhalb von Klassen ausdrücken. Die Gruppen der Mädchen bzw. der Jungen sollten derart formulierte Items jeweils homogener einschätzen, was zu größerer Variation bezüglich der Unterschiede zwischen den Gruppen führen sollte – und somit zu größeren Streuungen der Regressionsgewichte zwischen Klassen.

In den Spalten mit der Bezeichnung „Korrelation Interzept, Regressionsgewicht" sind die Zusammenhänge zwischen klassenspezifischem Interzept (aufgrund der *dummy*-Kodierung des Geschlechts entspricht der Interzept – d.h. der Achsenabschnitt – hier der Einschätzung der Jungen in der jeweiligen Klasse) und dem Regressionsgewicht (die „mädchenspezifische Differenz" zu den Jungen) dargestellt. Ein positiver Zusammenhang bedeutet hier also: Je höher die Einschätzung der Jungen (bezüglich des jeweiligen Unterrichtsindikators) umso größer ist auch der Differenzwert (nicht der absolute Unterschied!) zwischen Mädchen und Jungen innerhalb von Klassen.

Tabelle 26: Geschlechtsspezifische Interrater-Übereinstimmung innerhalb von Klassen: Zwei-Ebenen-Modell[100] mit klassenspezifisch variierendem Regressionsgewicht für die Geschlechtszugehörigkeit[101]

		Englisch					Deutsch				
		Rohwerte[102]			z-standardisierte Werte[103]		Rohwerte			z-standardisierte Werte	
		Regressionsgewicht			Regressionsgewicht		Regressionsgewicht			Regressionsgewicht	
Unterrichtsmerkmal	Itemnummer (I) = Ich-Bezug (K) = Klassen-Bezug	Mittelwert	Variation zwischen Klassen (SD[104])	Korrelation[105] Interzept, Regressionsgewicht	Mittelwert	Variation zwischen Klassen (SD)	Mittelwert	Variation zwischen Klassen (SD)	Korrelation Interzept, Regressionsgewicht	Mittelwert	Variation zwischen Klassen (SD)
Thematische Motivierung	1 (I)	0.14***	0.24***	-.18	0.15	0.26	0.23***	0.27***	-.25	0.25	0.30
	2 (K)	0.03 n.s.	0.21***	.03	0.03	0.23	0.10**	0.25***	.07	0.11	0.27
Verständlichkeit	1 (I)	0.10***	0.17**	-.71	0.12	0.21	0.21***	0.20***	-.45	0.25	0.25
	2 (K)	0.11***	0.22***	-.37	0.14	0.27	0.12***	0.21***	-.02	0.14	0.25
	3 (K)	0.09**	0.21***	-.14	0.10	0.24	0.12**	0.31***	-.20	0.13	0.34
	4 (I)	0.08*	0.26***	-.33	0.09	0.28	0.14***	0.32***	-.16	0.15	0.34
Schülerorientierung	1 (K)	0.08**	0.22***	-.26	0.10	0.28	0.11***	0.17***	.03	0.14	0.21
	2 (K)	0.16***	0.20***	-.20	0.19	0.23	0.16***	0.19***	.09	0.19	0.22
	3 (K)	0.10**	0.24***	-.45	0.12	0.29	0.11**	0.25***	-.25	0.13	0.30
	4 (I)	0.14***	0.25***	-.32	0.17	0.30	0.19***	0.24***	-.33	0.21	0.27
	5 (I)	0.12***	0.18***	-.26	0.14	0.21	0.18***	0.24***	-.34	0.21	0.27
Strukturiertheit	1 (K)	0.01 n.s.	0.13**	.68	0.01	0.14	0.06 n.s.	0.17**	.04	0.06	0.19
	2 (K)	-0.03 n.s.	0.12**	.34	-0.04	0.14	0.10**	0.18***	.25	0.11	0.20
	3 (K)	0.04 n.s.	0.24***	-.39	0.05	0.27	0.07*	0.15**	-.01	0.08	0.17
Klassenführung	1 (K)	0.13***	0.26***	-.23	0.14	0.29	0.17***	0.27***	.03	0.19	0.29
	2 (K)	0.06*	0.24***	-.02	0.07	0.27	0.09**	0.20***	.01	0.09	0.22

[100] Die Analysen erfolgten mithilfe der Software HLM 6.04 (Raudenbush, Bryk, Cheong & Congdon, 2004). Ausgewählt wurden Klassen, bei denen von jeweils mindestens vier Schülerinnen und vier Schülern Angaben zu den hier analysierten Items vorlagen. Die Zahl der Klassen reduzierte sich hierdurch auf 294.

[101] *dummy*-kodiert: 0 = Jungen, 1 = Mädchen.

[102] Skalierung: 1 *stimmt gar nicht*, 2 *stimmt eher nicht*, 3 *stimmt eher*, 4 *stimmt ganz genau*

[103] Die in diesen Spalten angegeben Werte beziehen sich auf die z-standardisierten Werte der Items. Die Signifikanzniveaus entsprechen den Angaben in den jeweiligen Spalten unter Rohwerte.

[104] Der besseren Interpretierbarkeit halber in Standardabweichungen statt Varianzen angegeben.

[105] Für diese Korrelation gibt HLM keine Informationen bezüglich des Signifikanzniveaus aus.

Abbildung 2 veranschaulicht einen solchen Zusammenhang anhand einer Auswahl an Klassen aus dem gesamten Spektrum (von minimaler bis maximaler Ausprägung des Interzepts) grafisch – hier der absolut betrachtet größte Zusammenhang von r = -.71 beim ersten Item der Skala *Verständlichkeit* im Fach Englisch. Es zeigen sich im linken Bereich der Abbildung, wo die mittleren Einschätzungen der Jungen relativ niedrig sind, tendenziell höhere Ausprägungen bei den Mädchen verglichen mit den Jungen. Umgekehrt sind bei hohen Einschätzungen des Items durch die Jungen innerhalb der Klassen (rechte Seite der Abbildung) die Ausprägungen bei den Mädchen tendenziell niedriger, wenngleich die absoluten Differenzen hier – aufgrund der im Mittel höheren Ausprägungen bei Mädchen (mittleres Regressionsgewicht = 0.10) – niedriger sind als im linken Bereich der Abbildung.

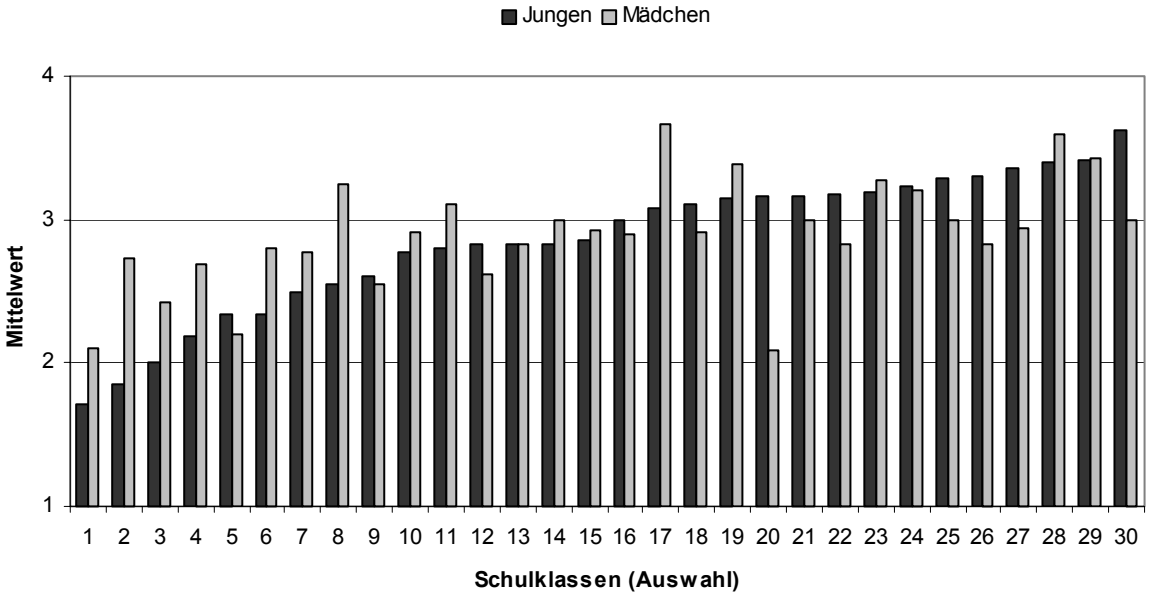

Abbildung 2: Zusammenhang zwischen Interzept und Regressionsgewicht am Beispiel des ersten Items der Skala *Verständlichkeit* im Fach Englisch. Sortierung aufsteigend nach mittlerer Einschätzung der Jungen in Klassen[106]

Auf der Basis dieser Analysen lässt sich jedoch nicht beurteilen, ob sich die Streuungen der Urteile innerhalb bzw. zwischen den geschlechtsspezifischen Subgruppen unterscheiden, ob sich also die Interrater-Reliabilität geschlechtstypisch unterscheidet. Wenn man davon ausgeht, dass lediglich die Kommunikation zwischen Schülerinnen und Schülern innerhalb von Klassen minimiert ist, so folgt daraus nur, dass die Übereinstimmungen sowohl *zwischen Schülerinnen* als auch *zwischen Schülern* erhöht sein sollten. Das heißt, auf dieser Grundlage

[106] Die Auswahl der Klassen erfolgte nach diesem Prinzip: Zunächst wurden alle Klassen aufsteigend nach dem Mittelwert der Jungen sortiert. Anschließend wurde die erste (also die Klasse mit dem niedrigsten Mittelwert für die Jungen), die zehnte, die zwanzigste Klasse usw. ausgewählt.

werden keine Unterschiede bezüglich der Übereinstimmungen von Schülerinnen bzw. Schülern innerhalb von Klassen erwartet.

Zur Beantwortung dieser Frage wurden zusätzlich die geschlechtsspezifischen Interrater-Reliabilitäten überprüft. Die Ergebnisse in Tabelle 27 zeigen, dass die Übereinstimmungen bezüglich der Unterrichtswahrnehmungen aus Sicht der Schüler*innen* – bis auf wenige nicht signifikante Unterschiede – durchweg *über* denjenigen ihrer (männlichen) Mitschüler liegen, und zwar bis zu 14.7 Prozentpunkte (Item 1 der Skala *Klassenführung* im Fach Deutsch). Die Interrater-Reliabilität beträgt hier bei den Schülerinnen 34.9%, was gleichzeitig die maximale ICC darstellt.

Tabelle 27: Geschlechtsunterschiede bezüglich der Interrater-Reliabilität von Unterrichtswahrnehmungen auf Itemebene: ICCs (metrisch, in Prozent) auf Itemebene nach Geschlecht

		ICC in Prozent[107]					
		Englisch			Deutsch		
		Geschlecht			Geschlecht		
Unterrichtsmerkmal	Itemnummer (I) = Ich-Bezug (K) = Klassen-Bezug	weiblich (w)	männlich (m)	Differenz (w-m)	weiblich (w)	männlich (m)	Differenz (w-m)
Thematische Motivierung	1 (I)	22.0	15.4	6.6**	20.9	15.0	5.9**
	2 (K)	27.8	16.8	11.0***	30.6	20.3	10.3***
Verständlichkeit	1 (I)	12.8	13.5	-0.7 n.s.	16.7	11.4	5.3 n.s.
	2 (K)	16.5	12.7	3.8 n.s.	22.2	12.8	9.4***
	3 (K)	27.4	16.8	10.6***	31.7	20.5	11.2***
	4 (I)	27.5	22.7	4.8 n.s.	32.6	23.7	8.9***
Schülerorientierung	1 (K)	25.4	17.2	8.2**	25.4	16.3	9.1***
	2 (K)	21.1	13.1	8.0***	24.0	14.8	9.2***
	3 (K)	19.3	12.3	7.0**	23.5	15.1	8.4***
	4 (I)	20.2	13.9	6.3**	21.0	14.0	7.0**
	5 (I)	18.4	11.3	7.1**	20.5	15.9	4.6*
Strukturiertheit	1 (K)	19.0	9.7	9.3***	15.8	9.8	6.0**
	2 (K)	13.8	8.4	5.4**	19.9	11.5	8.4***
	3 (K)	14.3	10.3	4.0*	16.0	11.7	4.3*
Klassenführung	1 (K)	26.6	17.9	8.7***	34.9	20.2	14.7***
	2 (K)	23.6	14.4	9.2***	27.0	18.0	9.0***
Mittelwert		21.0	14.2	6.8	23.9	15.7	8.2

Zur Erklärung dieser beachtlichen Unterschiede sind mindestens drei – miteinander kombinierbare – Hypothesen denkbar:

[107] Die Berechnungen erfolgten als Zwei-Ebenen-Modelle in Mplus 4.1. Im Prinzip würde sich hier die Verwendung sogenannter Mischverteilungsmodelle (TYPE IS MIXTURE) anbieten. In Mplus können derzeit jedoch in solchen Modellen keine unterschiedlichen Varianzen auf Ebene 2 modelliert werden. Da hier nicht zwingend davon ausgegangen werden kann, dass sich die Varianzen nur innerhalb von Klassen geschlechtstypisch unterscheiden (sondern möglicherweise auch auf Aggregatebene), wurde eine andere Modellierung gewählt. Dazu wurden jeweils Modelle mit je einer Variable für Schülerinnen und Schüler (die nicht zutreffende Variable enthielt jeweils ein missing value) berechnet. Die über das MODEL CONSTRAINT-Kommando berechneten ICCs wurden mithilfe des MODEL TEST-Kommandos (Wald-Test) auf identische Ausprägungen (ICC(Schülerinnen) = ICC(Schüler)) überprüft. Die Ergebnisse des Wald-Tests sind in den Spalten mit der Bezeichnung „Differenz (w-m)" dargestellt (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$; n.s. für $p \geq .05$).

1. Jungen kommunizieren *insgesamt* weniger miteinander als Mädchen.
2. Jungen sprechen weniger über *unterrichtsbezogene* Inhalte als Mädchen.
3. Jungen gewichten relevante unterrichtsbezogene Informationen ihrer Mitschüler bei der Urteilsbildung weniger stark als Mädchen (z.B. weil sie sich ihrer Urteile sicherer sind als Mädchen).

Aus methodischer Sicht können die unterschiedlichen ICCs und die über Klassen hinweg variierenden Regressionsgewichte für den Prädiktor „Geschlecht" im Sinne einer Varianzheterogenität der Ebene 1-Residuen interpretiert werden (vgl. z.B. Snijders & Bosker, 1999), was eine Verletzung der Modellannahmen in „Standard"-Analysen darstellt. In verschiedenen Multilevel-Softwarepaketen – wie z.B. in HLM[108], SAS MIXED-Prozedur (SAS Institute Inc., 2004) etc. – lassen sich solche Heterogenitäten modellieren. In Mplus wäre eine Modellierung über Mischverteilungsmodelle denkbar. Da solche Modelle mithilfe (speicher- und rechenintensiver) numerischer Integrationsverfahren geschätzt werden, sind hier der Modellkomplexität allerdings enge Grenzen gesetzt. Für die relativ komplexen Analysen in der vorliegenden Untersuchung bedeutet dies, dass die Varianzheterogenität auf Ebene 1 nicht modelliert werden kann.

Gleichzeitig sollte die Existenz von Subgruppen innerhalb von Klassen, deren Urteile in gewissem Maße übereinstimmen, und die insofern eine Ebene zwischen den beiden hier modellierten Ebenen darstellt, zu einer Überschätzung der (Zwei-Ebenen-)ICCs führen (vgl. Kap. 3.4.2). Es ist davon auszugehen, dass diese Verletzung der Annahmen eines Zwei-Ebenen-KFA-Modells auch zu einem gewissen *bias* bei den übrigen Modellparametern (wie Faktorladungen, Faktorkorrelationen etc.) führt. Dazu gibt es meines Wissens allerdings bisher keine (Monte-Carlo-) Studien.

4.3 Differenziertheit und intraindividuelle Unterschiede (fachspezifische Unterrichtswahrnehmung)

Im Folgenden wird zunächst der Frage nachgegangen, ob bzw. in welchem Ausmaß sich die theoretischen Unterrichtsmerkmale in den Wahrnehmungen der Schülerinnen und Schüler abbilden. Wenn sich die Indikatoren (Items) der jeweiligen theoretischen Konstrukte zu jeweils einem Faktor zusammenfassen lassen (ohne Doppelladungen mit anderen Faktoren und bei akzeptabler Modellgüte), dann ist damit die Differenzierung der Konstrukte statistisch ab-

[108] Dort allerdings nur dann, wenn auf die Verwendung von (Populations- bzw. Stichproben-)Gewichten verzichtet wird.

gesichert. Bei relativ hohen Korrelationen zwischen den verschiedenen Konstrukten ist jedoch die praktische Bedeutung einer solchen Unterscheidung in Frage gestellt.

Hier wird zusätzlich erwartet, dass bei allen untersuchten Unterrichtsmerkmalen zumindest eine Analogie zwischen den Konstrukten auf beiden Ebenen (innerhalb von Klassen und zwischen Klassen) vorliegt. Aus methodischer Sicht bedeutet dies, dass sich die jeweiligen Indikatoren eines Unterrichtsmerkmals *auf beiden Ebenen* jeweils einem Faktor zuordnen lassen. Die drei Items des Merkmals *Strukturiertheit* z.B. sollten also sowohl auf Individual- als auch auf Klassenebene einen Faktor repräsentieren. Dies ist das Minimalkriterium, um von analogen Konstrukten auf beiden Ebenen sprechen zu können. Wenn sich hingegen die Strukturen auf Individual- und Klassenebene vollkommen unterscheiden (z.B. unterschiedliche Anzahl an Faktoren, Faktoren laden ebenenspezifisch auf unterschiedliche Indikatoren), dann kann m.E. nicht mehr von analogen Konstrukten gesprochen werden.

Zur Überprüfung dieser Fragen wurde eine zweiebenenanalytische konfirmatorische Faktorenanalyse durchgeführt, bei der die Indikatoren der fünf Konstrukte in beiden Fächern auf beiden Ebenen jeweils einem Faktor zugeordnet wurden. In Tabelle 28 sind die latenten Interkorrelationen der Unterrichtswahrnehmungen auf beiden Ebenen dargestellt. Die Modellgüte-Indizes (vgl. Kap. 4.2, Fußnote 89) zeigen eine akzeptable Passung des Modells an, d.h. die theoretisch erwarteten Faktoren lassen sich auf beiden Ebenen darstellen.

Tabelle 28: Latente Interkorrelationen[109] der Unterrichtsmerkmale innerhalb von Klassen und zwischen Klassen (Korrelationen innerhalb[110] von Klassen sind unterhalb, Korrelationen zwischen Klassen oberhalb der Hauptdiagonale dargestellt; Interkorrelationen innerhalb eines Fachs sind fett gedruckt; Korrelationen analoger Merkmale in beiden Fächern sind grau hinterlegt)

		Englisch					Deutsch				
	Unterrichtsmerkmal	1	2	3	4	5	6	7	8	9	10
Englisch	1. Themat. Motivierung		**.94*****	**.88*****	**.79*****	**.77*****	-.05	-.04	-.01	.06	.06
	2. Verständlichkeit	**.73**		**.87*****	**.74*****	**.79*****	-.04	.02	.07	.03	.09
	3. Schülerorientierung	**.67**	**.74**		**.64*****	**.57*****	-.02	.02	.07	-.03	.06
	4. Strukturiertheit	**.60**	**.55**	**.54**		**.60*****	.15	.15*	.16*	.36***	.21**
	5. Klassenführung	**.50**	**.60**	**.54**	**.41**		-.11	-.09	-.04	-.07	-.04
Deutsch	6. Themat. Motivierung	.23	.15	.16	.20	.16		**.95*****	**.90*****	**.83*****	**.79*****
	7. Verständlichkeit	.14	.29	.21	.16	.24	**.75**		**.92*****	**.84*****	**.81*****
	8. Schülerorientierung	.15	.22	.26	.17	.22	**.68**	**.81**		**.76*****	**.71*****
	9. Strukturiertheit	.15	.16	.18	.47	.15	**.58**	**.57**	**.53**		**.75*****
	10. Klassenführung	.15	.22	.18	.17	.32	**.53**	**.61**	**.56**	**.44**	

Insbesondere auf *Klassenebene* zeigen sich jedoch z.T. sehr hohe Zusammenhänge zwischen verschiedenen Konstrukten: Die höchste Korrelation liegt bei .95 (*Thematische Moti-*

[109] Modellgüte-Indizes: vgl. Kap. 4.2, Fußnote 89

[110] Da alle Korrelationen innerhalb von Klassen hoch signifikant sind ($p < .001$; zweiseitige Tests) wurde auf eine Kennzeichnung verzichtet. Bei den Korrelationen auf Klassenebene wurde die übliche Kennzeichnung verwendet (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$; zweiseitige Tests).

vierung und *Verständlichkeit* im Fach Deutsch). Korrelationen dieser Größenordnung bezeichnet Rindskopf (1984) als nahezu perfekt (also nahe 1). Eine perfekte Korrelation zwischen zwei Faktoren ist eine äquivalente Formulierung eines einfaktoriellen Modells (d.h. die beiden Faktoren lassen sich nicht unterscheiden). Eine Überprüfung dieses Zusammenhangs auf Klassenebene anhand jeweils zweifaktorieller Modelle (*Thematische Motivierung* und *Verständlichkeit* im Fach Deutsch bzw. im Fach Englisch) mittels Wald-Test ergab allerdings, dass die Fixierung dieser Korrelationen auf 1 (im Sinne eines jeweils einfaktoriellen Modells) eine signifikante Verschlechterung der Modellpassung mit sich bringt[111]. Die Differenzierung der Faktoren ist also aus statistischer Sicht angemessen. Aus praktischer Sicht ist eine solche Unterscheidung kaum noch möglich: Entsprechend wurden auch die drei am höchsten korrelierten Konstrukte auf Klassenebene (*Thematische Motivierung*, *Verständlichkeit* und *Schülerorientierung*) in DESI – zusammen mit weiteren Merkmalen – einem einzigen Faktor („Prozessqualität“) zugeordnet (A. Helmke, et al., in Druck-a; Klieme et al., in Druck).

Betrachtet man die *fachspezifischen Interkorrelationen*, so zeigen sich im Fach Deutsch durchweg höhere Zusammenhänge zwischen den geteilten Unterrichtswahrnehmungen als im Fach Englisch. Der größte Unterschied bezüglich der Zusammenhänge der Merkmale in den Fächern Englisch und Deutsch zeigt sich bei den Unterrichtswahrnehmungen *Strukturiertheit* und *Klassenführung* (r = .60 im Fach Englisch vs. r = .75 im Fach Deutsch).

Erstaunlich sind die geringen bzw. nicht vorhandenen (nicht signifikanten) *fachübergreifenden Zusammenhänge* zwischen den verschiedenen Konstrukten auf Klassenebene. Die höchste Korrelation beträgt hier r = .36 (*Strukturiertheit* im Fach Englisch bzw. Deutsch). Demnach beurteilen Klassen den Unterricht in den beiden Fächern sehr unabhängig voneinander. Dies ist insofern bemerkenswert, als (1) die Klassen z.T. von derselben Lehrkraft in beiden Fächern unterrichtet wurden (vgl. Kap. 3.2) – und die meisten hier untersuchten Unterrichtsmerkmale in diesen Fällen wohl zumindest ähnlich hohe Ausprägungen besitzen sollten – und (2) sich auch darüber liegende Ebenen wie z.B. die Schulebene (im Sinne einer bestimmten „Unterrichtskultur“ an Schulen) oder Bildungsgangunterschiede in den Zusammenhängen niederschlagen sollten.

---

[111] Modell für das Fach Deutsch: Wald-Koeffizient = 24.848; DF = 1; $p < .001$; Modellgüte-Indizes: $Chi^2 = 274.894$, Scaling Correction Factor = 1.120; DF = 17; CFI = 0.980; TLI = 0.964; RMSEA = 0.044; SRMR (zwischen) = 0.031; SRMR (innerhalb) = 0.026. Die negative Residualvarianz von Item 1 der Skala *Thematische Motivierung* ($p < .05$) wurde auf Null fixiert.
Modell für das Fach Englisch: Wald-Koeffizient = 24.863; DF = 1; $p < .001$; Modellgüte-Indizes: $Chi^2 = 235.015$, Scaling Correction Factor = 1.148; DF = 17; CFI = 0.981; TLI = 0.966; RMSEA = 0.041; SRMR (zwischen) = 0.039; SRMR (innerhalb) = 0.023; Die negative Residualvarianz von Item 1 der Skala *Thematische Motivierung* (nicht signifikant) wurde auf Null fixiert.

Während die hohen fachspezifischen Interkorrelationen auf Klassenebene möglicherweise auf tatsächlich vorhandene Gemeinsamkeiten des Unterrichts verschiedener Lehrkräfte zurückgeführt werden können („gute“ Lehrkräfte unterrichten evtl. in allen hier untersuchten Dimensionen „gut“) sind die – wenngleich etwas niedrigeren – *fachspezifischen Interkorrelationen innerhalb von Klassen* hier weniger plausibel: Da es sich hier um Abweichungen von der geteilten Wahrnehmung (in gewissem Sinne: dem *true score*) handelt, wären hier bei einer „objektiven“ Beurteilung keine Zusammenhänge zu erwarten, es sei denn, diese wären auf über verschiedene bzw. alle Unterrichtsmerkmale konstante Milde- bzw. Strengeeffekte der einzelnen Beurteiler zurückzuführen. Dem widersprechen allerdings die relativ niedrigen *fachübergreifenden Zusammenhänge* zwischen den Unterrichtsmerkmalen. Auf das jeweilige Unterrichtsmerkmal bzw. auf verschiedene Unterrichtsmerkmale bezogene Milde- oder Strengeeffekte müssten sich auch fachübergreifend zeigen: Wer beispielsweise das Merkmal *Klassenführung* generell streng bewertet, sollte sowohl im Fach Deutsch als auch im Fach Englisch niedrigere Urteile vergeben als seine Klassenkameraden. Dies würde zu hohen Korrelationen zumindest zwischen den korrespondierenden Unterrichtsmerkmalen führen. Mit Ausnahme des Merkmals *Strukturiertheit* (r = .47) sind die Zusammenhänge der korrespondierenden Merkmale aber eher niedrig. Demnach deuten die fachspezifisch hohen Interkorrelationen innerhalb von Klassen hier eher auf eine globale Beurteilung entweder des jeweiligen Fachs oder der jeweiligen Lehrkraft hin.

Um dieser Frage nachzugehen, wurden die Zusammenhänge der Unterrichtsmerkmale in der Teilstichprobe (24 Klassen) mit *identischen Lehrkräften* in beiden Fächern untersucht. Aufgrund des geringen Stichprobenumfangs auf Klassenebene sind hier latente Analysen nicht sinnvoll. Deshalb wurden sogenannte Faktor*scores* für die jeweiligen Unterrichtsmerkmale innerhalb eines Ein-Ebenen-Modells[112] auf der Basis der Gesamtstichprobe erzeugt. Diese wurden anschließend in Zwei-Ebenen-Modellen analysiert (aufgrund von Konvergenzproblemen wegen des geringen Stichprobenumfangs und der hohen Interkorrelationen auf beiden Ebenen jeweils nur zwei Merkmale pro Modell).

Die Ergebnisse in Tabelle 29 zeigen – wenngleich bei der Interpretation der Zusammenhänge aufgrund der genannten Probleme Vorsicht geboten ist – auf *Klassenebene* erwartungsgemäß hohe positive Interkorrelationen der Unterrichtsmerkmale auch fachübergreifend. Die-

[112] Dazu wurde in Mplus eine KFA mit freien Ladungen und auf den Wert 1 fixierten Varianzen berechnet. Modellgüte-Indizes: $Chi^2 = 2257.913$, Scaling Correction Factor = 1.916; DF = 419; CFI = 0.970; TLI = 0.964; RMSEA = 0.024; SRMR = 0.022
Die hier angegebenen Indizes unterscheiden sich minimal von denen des analogen Modells, auf dessen Basis die *factor determinacies* berechnet wurden (s. Kap. 4.2, Fußnote 80), da dort auf die Berücksichtigung der *cluster*-Struktur der Daten verzichtet werden musste.

se Zusammenhänge lassen jedoch mindestens zwei unterschiedliche – auch simultane – Erklärungen zu: (1) Lehrkräfte gestalten ihren Unterricht bezüglich der hier untersuchten (allgemeinen) Unterrichtsmerkmale in beiden Fächern sehr ähnlich. Oder: (2) Die globale Beurteilung der Lehrkraft durch die Klasse schlägt sich in den Unterrichtswahrnehmungen in beiden Fächern nieder.

Tabelle 29: Manifeste Interkorrelationen[113] der Unterrichtsmerkmale innerhalb von Klassen und zwischen Klassen mit identischen Lehrkräften in den Fächern Englisch und Deutsch (Korrelationen innerhalb[114] von Klassen sind unterhalb, Korrelationen zwischen Klassen oberhalb der Hauptdiagonale dargestellt; Interkorrelationen innerhalb eines Fachs sind fett gedruckt; Korrelationen analoger Merkmale in beiden Fächern sind grau hinterlegt)

		Englisch					Deutsch				
	Unterrichtsmerkmal	1	2	3	4	5	6	7	8	9	10
Englisch	1. Themat. Motivierung		**.94***	**.91***	**.91***	**.91***	.93**	.87**	.77**	.92**	.82**
	2. Verständlichkeit	**.84**		**.96***	**.87***	**.99****	.89**	.97**	.91**	.88*	.99**
	3. Schülerorientierung	**.75**	**.83**		**.73**	**n.k.**	n.k.	n.k.	n.k.	.84**	n.k.
	4. Strukturiertheit	**.73**	**.67**	**.64**		**.82***	.78**	.78*	.65*	n.k.	.79**
	5. Klassenführung	**.63**	**.75**	**n.k.**	**.53**		.91***	.98**	1.00**	.93**	n.k.
Deutsch	6. Themat. Motivierung	.58	.49	n.k.	.50	.37		**.95****	**.90****	**.90****	**.90****
	7. Verständlichkeit	.46	.60	n.k.	.43	.43	**.84**		**.98****	**.87***	**1.00****
	8. Schülerorientierung	.47	.54	n.k.	.43	.41	**.77**	**.88**		**.75***	**.96****
	9. Strukturiertheit	.40	.41	.38	n.k.	.30	**.68**	**.68**	**.60**		**.90***
	10. Klassenführung	.44	.51	n.k.	.44	n.k.	**.69**	**.76**	**.68**	**.58**	

Für die letztgenannte Hypothese sprechen auch die – verglichen mit den Ergebnissen auf der Basis aller Klassen, also im Wesentlichen mit unterschiedlichen Lehrkräften (s. Tabelle 28) – hohen Interkorrelationen der jeweils auf ein unterschiedliches Fach bezogenen Unterrichtsmerkmale *innerhalb von Klassen*. Wenn Lehrkräfte in beiden Fächern ähnlich unterrichten, dann sollte dies innerhalb der Klasse (also im Sinne einer Abweichung von der geteilten Wahrnehmung) kaum eine Rolle spielen. Diese Zusammenhänge legen hier folgende Interpretation nahe: Die individuelle Wahrnehmung der Lehrkraft (Globalurteil) beeinflusst die individuellen Unterrichtswahrnehmungen.

Die Übertragung dieser Zusammenhänge innerhalb von Klassen auf die Zusammenhänge auf Klassenebene ist jedoch nur dann möglich, wenn eine kollektive globale Wahrnehmung der Lehrkraft angenommen wird. Es wäre möglich, dass in die individuellen Urteile zwar globale Sichtweisen der Lehrkraft eingehen, diese dann aber – eine vollkommen idiosynkratische Beurteilung vorausgesetzt[115] – auf Klassenebene ohne Bedeutung sind, da sie durch die Aggregation „herausgemittelt" werden. Dies würde lediglich zu einer Verringerung der Interra-

[113] Die jeweiligen Modelle in den mit „n.k." gekennzeichneten Zellen konvergierten nicht.
[114] Da alle Korrelationen innerhalb von Klassen hoch signifikant sind ($p < .001$; zweiseitige Tests) wurde auf eine Kennzeichnung verzichtet. Bei den Korrelationen auf Klassenebene wurde die übliche Kennzeichnung verwendet (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$; einseitige Tests).
[115] In Kapitel 4.6 wird jedoch gezeigt, dass auch von einer globalen Wahrnehmung der Lehrkraft auf Klassenebene ausgegangen werden muss.

ter-Reliabilität führen; die Validität der aggregierten Unterrichtswahrnehmungen bliebe davon unberührt.

Um die Frage der *innerklasslichen* Zusammenhänge weiter aufzuhellen, wurden die beiden Schülergruppen miteinander verglichen, die aus Klassen stammen, die von derselben Lehrkraft bzw. von unterschiedlichen Lehrkräften in den Fächern Englisch und Deutsch unterrichtet wurden (vgl. Kap. 3.2). Die in Tabelle 30 dargestellten Ergebnisse zeigen Folgendes: Die Korrelationen der Unterrichtsmerkmale innerhalb eines Fachs sind in beiden Gruppen identisch (keine signifikanten Unterschiede). Dagegen sind in der Gruppe mit identischen Lehrkräften *sämtliche* Korrelationen zwischen Unterrichtsmerkmalen, die sich auf unterschiedliche Fächer beziehen (Deutsch bzw. Englisch), deutlich höher ausgeprägt als in der Gruppe mit unterschiedlichen Lehrkräften in beiden Fächern.

Tabelle 30: Manifeste Interkorrelationen der Unterrichtsmerkmale innerhalb von Klassen; Vergleich der Gruppen von Schülerinnen und Schülern, die von derselben Lehrkraft bzw. unterschiedlichen Lehrkräften unterrichtet werden[116] (Gruppe mit identischen Lehrkräften unterhalb, Gruppe mit unterschiedlichen Lehrkräften oberhalb der Hauptdiagonale; Interkorrelationen, die sich in beiden Gruppen signifikant unterscheiden sind fett gedruckt[117]; Korrelationen analoger Merkmale in beiden Fächern sind grau hinterlegt)

		Englisch					Deutsch				
	Unterrichtsmerkmal	1	2	3	4	5	6	7	8	9	10
Englisch	1. Themat. Motivierung		.85	.78	.74	.65	**.24**	**.17**	**.18**	**.20**	**.18**
	2. Verständlichkeit	.83		.85	.67	.74	**.16**	**.28**	**.24**	**.19**	***.24***
	3. Schülerorientierung	.73	.83		.64	.66	**.18**	**.23**	**.30**	**.20**	***.21***
	4. Strukturiertheit	.75	.67	.64		.53	**.25**	**.21**	**.21**	***.53***	**.22**
	5. Klassenführung	.62	.74	.70	.55		**.17**	**.24**	**.25**	***.16***	**.33**
Deutsch	6. Themat. Motivierung	**.62**	**.52**	**.47**	**.51**	**.41**		.86	.78	.70	.67
	7. Verständlichkeit	**.47**	**.62**	**.50**	**.42**	**.46**	.81		.89	.69	.75
	8. Schülerorientierung	**.48**	**.56**	**.59**	**.42**	**.45**	.76	.88		.61	.68
	9. Strukturiertheit	**.46**	**.44**	**.40**	***.73***	***.35***	.67	.65	.58		.57
	10. Klassenführung	**.43**	***.48***	***.40***	**.43**	**.49**	.67	.73	.65	.61	

4.4 Fachspezifität: Unterscheiden sich die Messmodelle analoger Unterrichtsmerkmale in den Fächern Deutsch und Englisch?

Analog zu der in Kapitel 4.2 behandelten Frage nach identischen Messmodellen von Unterrichtswahrnehmungen auf beiden Ebenen (innerhalb und zwischen Klassen) geht es im Fol-

[116] Die Analysen erfolgten in Einebenen-Mischverteilungsmodellen in Mplus (KNOWNCLASS-Option; TYPE IS MIXTURE) auf der Basis der gruppenzentrierten (also Abweichungen vom Klassenmittelwert) Unterrichtswahrnehmungen (Basis: Faktorscores). Alle Interkorrelationen in beiden Gruppen (identische vs. unterschiedliche Lehrkräfte) sind hoch signifikant ($p < .001$; zweiseitige Tests).

[117] In beiden Gruppen (identische vs. unterschiedliche Lehrkräfte) wurden unterschiedliche Varianzen bezüglich der Unterrichtswahrnehmungen zugelassen. Die Signifikanz wurde anhand eines Wald-Tests (identische Korrelationen in beiden Gruppen mittels Fishers z-Transformation $z = ½ \ln((1 + r)/(1 - r))$, also : $z_1 = z_2$) überprüft. Alle fett markierten Unterschiede sind hoch signifikant ($p < .001$); Ausnahme: fett und kursiv markierte Unterschiede ($p < .01$).

genden um die Frage, ob die Annahme isomorpher Konstrukte in den Fächern Deutsch und Englisch gerechtfertigt ist. Dabei sind hier die jeweiligen Messmodelle sowohl ebenenspezifisch als auch ebenenübergreifend zu prüfen. Es ist durchaus möglich, dass bei einem Unterrichtsmerkmal das Messmodell der subjektiven Komponente (also innerhalb von Klassen) fachübergreifend gilt, während bezüglich der geteilten Wahrnehmung unterschiedliche Messmodelle vorliegen, oder umgekehrt. Weiterhin können ebenenspezifisch identische Messmodelle auf beiden Ebenen vorliegen. Darüber hinaus ist auch eine „perfekte" Isomorphie im Sinne sowohl fachübergreifend als auch ebenenübergreifend identischer Messmodelle möglich.

Zur Überprüfung der auf das Unterrichtsfach bezogenen Invarianz der Messmodelle wurden entsprechend für jedes Unterrichtsmerkmal insgesamt vier Modelle überprüft: (1) Invarianz der Messmodelle auf der Ebene innerhalb von Klassen, (2) zwischen Klassen, (3) innerhalb und zwischen Klassen (jedoch ebenenspezifische Messmodelle) sowie (4) ebenen- *und* fachübergreifend invariante Messmodelle. Die Ergebnisse sind in Tabelle 31 dargestellt. Bei den Unterrichtsmerkmalen *Strukturiertheit* und *Klassenführung* ist die Annahme sowohl fach- als auch ebenenübergreifender Messmodelle gerechtfertigt. Weiterhin führt bei der *Verständlichkeit* auf *Klassenebene* die Konstanthaltung der Messmodelle in beiden Fächern zu keiner signifikant schlechteren Modellpassung.

Tabelle 31: Fachspezifität bzw. Invarianz der Messmodelle der wahrgenommenen Unterrichtsmerkmale innerhalb von Klassen bzw. zwischen Klassen[118]

	Invarianz der Messmodelle bezüglich des Unterrichtsfachs											
	jeweils eine Ebene						jeweils beide Ebenen					
	innerhalb von Klassen			zwischen Klassen			ebenenspezifische Messmodelle			ebenenübergreifendes Messmodell		
	Koeffizient	DF	p	Koeffizient	DF	p	Koeffizient	DF	p	Koeffizient	DF	p
Thematische Motivierung	4.990	1	0.026	5.351	1	0.021	8.753	2	0.013	69.348	3	0.000
Verständlichkeit	11.948	3	0.008	4.640	3	**0.200**	17.168	6	0.009	137.597	9	0.000
Schülerorientierung	11.710	4	0.020	14.346	4	0.006	25.379	8	0.001	80.677	12	0.000
Strukturiertheit	1.042	2	**0.594**	4.373	2	**0.112**	5.166	4	**0.271**	7.323	6	**0.292**
Klassenführung	0.368	1	**0.544**	0.055	1	**0.815**	0.448	2	**0.800**	3.402	3	**0.334**

Auch hier wurden – analog zu den ebenenübergreifenden Analysen zur Messinvarianz in Kapitel 4.2 – die Unterschiede zwischen den Ladungen der fachspezifischen Messmodelle der einzelnen Unterrichtsmerkmale überprüft, um die praktische Bedeutsamkeit der Unterschiede einschätzen zu können. Die in Tabelle 32 dargestellten prozentualen Abweichungen der jeweiligen auf das Fach Deutsch bezogenen von den auf das Fach Englisch bezogenen unstan-

[118] Grundlage war jeweils das in Tabelle 28 (latente Interkorrelationen) verwendete Modell. Die Überprüfung der Messinvarianz im Sinne identischer unstandardisierter Ladungen erfolgte mittels Wald-Test. Nicht signifikante ($p \geq .05$) Ergebnisse sind fett gedruckt.

dardisierten Ladungen liegen im Bereich von -11.8% (Item 3 der Skala *Verständlichkeit*, innerhalb von Klassen) bis 14.1% (Item 2 der Skala *Strukturiertheit*, Klassenebene). Obwohl sich – absolut betrachtet – die höchsten prozentualen Abweichungen bei der Skala *Strukturiertheit* (Klassenebene) finden, sind diese Unterschiede nicht signifikant (vgl. Tabelle 31). Die relativ niedrigen ICCs der Indikatoren dieser Faktoren (in beiden Fächern; vgl. Kap. 4.2, Tabelle 18) – d.h. relativ niedrige Varianzanteile auf Klassenebene – führen zu großen Standardfehlern der Koeffizienten auf Ebene 2. Entsprechend ist die Teststärke des Wald-Tests hier relativ gering (d.h. nur relativ große Unterschiede führen hier zu einer signifikant schlechteren Passung eines Modells mit itemspezifisch identischen Ladungen in beiden Fächern).

Tabelle 32: Fachspezifische prozentuale Abweichung der unstandardisierten Ladungen: innerhalb von Klassen und zwischen Klassen

	prozentuale Abweichung[119] (Deutsch vs. Englisch)	
Unterrichtsmerkmal[120]	innerhalb von Klassen	zwischen Klassen
Thematische Motivierung		
Item 2	-4.9	9.5
Verständlichkeit		
Item 2	-5.3	10.4
Item 3	-11.8	2.8
Item 4	-7.2	0.4
Schülerorientierung		
Item 2	-1.7	8.0
Item 3	-2.4	9.2
Item 4	4.9	-1.1
Item 5	1.5	9.0
Strukturiertheit		
Item 2	-1.2	14.1
Item 3	-3.3	12.1
Klassenführung		
Item 2	2.1	1.3

Weiterhin wurde bei den beiden bezüglich ihrer Ladungen über beide Ebenen und Fächer hinweg vollständig invarianten Merkmalen *Strukturiertheit* und *Klassenführung* sowie bei der auf Klassenebene ladungsinvarianten *Verständlichkeit* überprüft, ob zusätzlich identische Interzepte und Residualvarianzen vorliegen (*strong* bzw. *strict factorial invariance*; vgl. Kap. 3.4.2). Die in Tabelle 33 dargestellten Ergebnisse zeigen, dass *strict factorial invariance* le-

[119] Die Abweichung bezieht sich auf die Differenz der Ladungen für das Fach Deutsch und für das Fach Englisch im Verhältnis zum Mittelwert der beiden fachspezifischen Ladungen. Der Mittelwert der Ladungen wurde deshalb als Referenz verwendet, da hier kein theoretisch plausibles Referenzmodell vorliegt (vgl. dazu auch Fußnote 95). Die zugrunde liegenden unstandardisierten Ladungen können Tabelle 22 entnommen werden.

[120] Da in den zugrunde liegenden Modellen jeweils die erste Ladung – also die Ladung des ersten Items – auf beiden Ebenen auf 1 fixiert wurde, sind diese jeweils identisch.

diglich für das Merkmal *Klassenführung* angenommen werden kann. Bei der wahrgenommenen *Strukturiertheit* unterscheiden sich interessanterweise nur die Interzepte, nicht jedoch die Ladungen und Residualvarianzen auf beiden Ebenen. Gleiches gilt für die geteilte Wahrnehmung der *Verständlichkeit*.

Tabelle 33: Fach- und ebenenübergreifende Tests auf Messinvarianz der Ladungen, der Interzepte und der Residualvarianzen[121]

Unterrichtsmerkmal	Test auf Äquivalenz	Koeffizient	DF	p
Verständlichkeit	Ladungen und Interzepte	27.710	6	0.000
(nur Klassenebene)	Ladungen und Residualvarianzen	13.460	7	**0.062**
	Ladungen, Interzepte, Residualvarianzen	42.574	10	0.000
Strukturiertheit	Ladungen und Interzepte	47.600	8	0.000
	Ladungen und Residualvarianzen	12.308	12	**0.421**
	Ladungen, Interzepte, Residualvarianzen	63.428	14	0.000
Klassenführung	Ladungen und Interzepte	3.532	4	**0.473**
	Ladungen und Residualvarianzen	8.146	7	**0.320**
	Ladungen, Interzepte, Residualvarianzen	8.203	8	**0.414**

Diese drei Faktoren lassen sich jeweils hinsichtlich ihrer fachspezifischen Varianzen miteinander vergleichen. Die Ergebnisse können Tabelle 34 entnommen werden. Auf Klassenebene zeigen sich höhere Streuungen im Fach Deutsch (Ausnahme: *Verständlichkeit*), während innerhalb von Klassen erwartungsgemäß keine Unterschiede vorliegen. Die subjektiven Varianz-Komponenten bleiben also konstant, obwohl sich die geteilten Wahrnehmungen, die ja zumindest von objektiven Faktoren des Unterrichts beeinflusst sein sollten, in ihrer Varianz unterscheiden.

Was Mittelwertunterschiede anbelangt, so lassen sich hier nur die Unterrichtswahrnehmungen (auf Klassenebene) bezüglich der *Klassenführung* miteinander vergleichen, da nur hier Invarianz auch bezüglich der Interzepte vorliegt. Der Unterschied ist allerdings nicht signifikant[122].

[121] Grundlage war jeweils das in Tabelle 28 (latente Interkorrelationen) verwendete Modell. Die Überprüfung der Messinvarianz im Sinne identischer unstandardisierter Ladungen, Interzepte und Residualvarianzen erfolgte mittels Wald-Test. Zur Überprüfung der Übereinstimmung der Interzepte wurden jeweils die Interzepte des ersten Indikators eines Faktors in beiden Fächern gleichgesetzt und der Mittelwert des jeweiligen Deutsch-Faktors frei geschätzt (die Voreinstellung, die den Faktormittelwert auf Null setzt, wurde also aufgehoben). Die frei geschätzten Interzepte wurden auf fachübergreifende Identität geprüft. Nicht signifikante ($p \geq .05$) Ergebnisse sind fett gedruckt.

[122] Der Faktormittelwert für *Klassenführung* im Fach Englisch wurde auf Null fixiert. Der entsprechende frei geschätzte Mittelwert im Fach Deutsch beträgt $M = -0.028$ ($p \geq .05$).

Tabelle 34: Fachspezifische Varianzen der faktorinvarianten Unterrichtsmerkmale[123]

	Innerhalb von Klassen				Zwischen Klassen			
	Varianz		Test auf identische Varianzen[124]		Varianz		Test auf identische Varianzen	
Unterrichtsmerkmal	Englisch	Deutsch	WALD-Koeffizient	p-Wert	Englisch	Deutsch	WALD-Koeffizient	p-Wert
Verständlichkeit	–	–	–	–	0.043	0.052	2.125	0.145
Strukturiertheit	0.298	0.311	2.231	0.135	0.064	0.089	7.347	**0.007**
Klassenführung	0.391	0.381	0.427	0.514	0.152	0.207	5.076	**0.024**

Insgesamt sind die fachspezifischen Unterschiede der Faktorladungen wohl eher von geringer praktischer Bedeutung, sofern nicht Mittelwerts- oder Varianzvergleiche bezüglich korrespondierender Unterrichtsmerkmale angestrebt werden. Für fachspezifische Zusammenhangsanalysen sind diese Differenzen (vermutlich) vernachlässigbar.

Bei Vergleichen bezüglich der Varianz bzw. der Mittelwerte von Faktoren, deren Messmodelle sich auch nur minimal (aber signifikant) unterscheiden, muss m.E. zumindest berücksichtigt werden, dass diese in den meisten Fällen dieselbe Stichprobengröße nutzen (und somit eine vergleichbare Teststärke besitzen). Wird theoretisch etwa ein deutlicher Mittelwertunterschied bezüglich eines bestimmten Merkmals erwartet, so ließe sich wohl bei relativ geringen Ladungs- und Interzept-Differenzen dennoch argumentieren, dass die entsprechenden Faktoren zu einem Vergleich herangezogen werden. In einem solchen Fall sollte dann aber die Differenz so groß sein, dass die Signifikanzgrenze deutlich unterschritten wird. Aus statistischer Sicht besteht eine gewisse Gefahr darin, bei der Überprüfung der Faktorinvarianz Unterschiede auf die Stichprobengröße zurückzuführen (d.h. die Teststärke des jeweiligen Verfahrens), während eine analoge Teststärke zum Nachweis von Unterschieden bezüglich der Faktorvarianzen bzw. -mittelwerte genutzt wird.

4.5 Itemformulierung: Ich- vs. Klassen-Bezug

Im Folgenden wird der Frage nachgegangen, ob sich die Itemformulierung (Ich- vs. Klassen-Bezug) in Form unterschiedlicher Varianzkomponenten innerhalb von Klassen niederschlägt, ob also je nach Itemformulierung neben dem zu messenden Konstrukt eine zusätzliche Dimension erfasst wird. Diese zusätzliche Dimension kann im Sinne des auf den Arbeiten von Campbell und Fiske (1959) basierenden *Multitrait-Multimethod*-Ansatzes mit latenten Variablen (MTMM, vgl. z.B. Marsh & Grayson, 1995) als Methodenkomponente betrachtet

[123] Modellgüte-Indizes: $Chi^2 = 4532.326$, Scaling Correction Factor = 1.032; DF = 868; CFI = 0.961; TLI = 0.955; RMSEA = 0.023; SRMR (zwischen) = 0.022; SRMR (innerhalb) = 0.054

[124] Die Zahl der Freiheitsgrade beträgt jeweils DF = 1.

werden. Das heißt, ichbezogene vs. klassenbezogene Formulierungen lassen sich als unterschiedliche Methoden der Erfassung desselben Konstrukts interpretieren.

Meist werden bei solchen Modellen die Methoden- und die *trait*-Komponenten als unabhängig (also unkorreliert) betrachtet (ebd.), während bei den *traits* Zusammenhänge im Modell berücksichtigt werden. Bei den Methodenfaktoren besteht ebenfalls – im sogenannten *correlated trait / correlated method model* (CTCM) – die Möglichkeit, Zusammenhänge zu modellieren. Solche Modelle führen jedoch häufig zu unzulässigen Lösungen (*heywood cases*). Im *correlated trait / uncorrelated method model* (CTUM) werden deshalb keine Zusammenhänge zwischen den Methodenfaktoren geschätzt. Da auch aus diesen Modellen häufig unzulässige Lösungen resultieren, schlägt Eid (2000) das sogenannte CTC(M-1)-Modell vor, in dem analog zum CTCM-Modell sowohl innerhalb der *traits* als auch innerhalb der Methodenfaktoren alle Korrelationen modelliert werden. Der einzige Unterschied zum CTCM-Modell besteht darin, dass beim CTC(M-1)-Modell ein Methodenfaktor weniger modelliert wird (daher „M-1"): Wurden verschiedene Merkmale z.B. insgesamt mittels dreier verschiedener Methoden erhoben, so werden im CTC(M-1)-Modell nur zwei Methodenfaktoren modelliert, während eine Methode als „Referenzmethode" fungiert[125].

Bezogen auf die zwei hier zu untersuchenden Methoden ist die Wahl einer Referenzmethode theoretisch nicht schlüssig zu klären: Einerseits erscheint es naheliegend, den Klassen-Bezug der Itemformulierung als Referenz zu betrachten, während der Ich-Bezug eine stärker idiosynkratische Komponente darstellen sollte. Wenn man andererseits annimmt, dass bei der Einschätzung der Urteile anderer (*proxy reporting*; vgl. dazu Kap. 2.1.3) zunächst auf das eigene Urteil zurückgegriffen wird, dann erscheint auch die Verwendung des Ich-Bezugs der Itemformulierung als Referenzmethode plausibel. Im Folgenden werden daher zwei CTC(M-1)-Modelle mit jeweils unterschiedlicher Referenzmethode sowie ein CTUM- und ein CTCM-Modell – das zusätzlich jeweils einen fachspezifischen Globalfaktor enthält – dargestellt, wobei die beiden fachspezifischen Modelle jeweils in ein (fachübergreifendes) Gesamtmodell integriert werden. Die fachspezifische Modellierung ist deshalb erforderlich, weil auch die Methodenfaktoren mit dem Fach bzw. der jeweiligen Lehrkraft konfundiert sein können. Wie bereits in Kapitel 2.1.2 erwähnt, kann Kenny (2004) zufolge die subjektive Komponente der (Unterrichts-)Wahrnehmung eine bedeutsame Größe darstellen – wenn es um spezifische Interaktionen der Lehrkraft und der jeweiligen Schülerin bzw. dem jeweiligen Schüler geht – die sich insbesondere bei ichbezogenen Itemformulierungen (also hier: beim Methodenfaktor „Ich-Bezug") zeigen sollte. Das heißt, es kann nicht davon ausgegangen werden, dass die

[125] Ein Nachteil dieses Modells besteht darin, dass es nicht symmetrisch ist: Die Modellpassung hängt von der jeweils gewählten Referenzmethode ab (vgl. Eid, 2000).

Methode „Ich-Bezug“ einen fachübergreifenden Faktor darstellt. Gleiches gilt für die Methode Klassen-Bezug: Die subjektive Einschätzung der Unterrichtswahrnehmungen der Mitschüler sollte sich ebenfalls zwischen beiden Fächern unterscheiden.

In den im Folgenden dargestellten Modellen wurden auf Klassenebene jeweils die drei Faktoren pro Fach modelliert (ohne Berücksichtigung von Itemformulierungen). Die hohen fachspezifischen Interkorrelationen der einzelnen Faktoren sowie die hohen Reliabilitäten der Indikatoren auf Klassenebene lassen keine zuverlässige Schätzung der Methodenfaktoren zu. Um die Abbildungen möglichst übersichtlich zu halten, wird jeweils nur das Modell innerhalb von Klassen dargestellt.

Vergleicht man die beiden CTC(M-1)-Modelle in Abbildung 3 bzw. Abbildung 4, so zeigen sich standardisierte Ladungen der Methodenfaktoren auf die jeweiligen Indikatoren in vergleichbarer Größenordnung (mit tendenziell höheren Ladungen des Klassen-Bezug-Faktors). Der Zusammenhang zwischen den fachspezifischen Ich-Bezug-Faktoren ist etwas höher (r = .56) als der zwischen den Klassen-Bezug-Faktoren (r = .35). Diese Zusammenhänge werden auch annähernd im CTUM-Modell (Abbildung 5) reproduziert. Während dort auch die standardisierten Ladungen der Klassen-Bezug-Faktoren fast identisch mit denen aus dem entsprechenden CTC(M-1)-Modell sind, zeigen sich deutliche Unterschiede bei den Ich-Bezug-Faktoren: Die Ladungen im CTUM-Modell sind hier teilweise nicht signifikant, teilweise sogar negativ.

Beim CTCM-Modell (hier nicht dargestellt), das eine Erweiterung des CTUM-Modells um die beiden Korrelationen zwischen den fachspezifischen Methodenfaktoren darstellt, zeigen sich extrem hohe Korrelationen zwischen den fachspezifischen Methodenfaktoren (r = .95 im Fach Englisch, r = .93 im Fach Deutsch). Dies spricht – zusammen mit durchweg hohen standardisierten Ladungen auf den Methodenfaktoren (.34 bis .73 im Fach Englisch, .40 bis .73 im Fach Deutsch) dafür, dass die Methodenfaktoren hier nicht mehr die unterschiedliche Itemformulierung messen, sondern einen fachspezifischen Globalfaktor der subjektiven Unterrichtswahrnehmung. Deshalb wurde das Modell um solche fachspezifischen Globalfaktoren (G-Faktoren) erweitert. Das resultierende Modell zeigt Abbildung 6. Aus Gründen der Übersichtlichkeit werden die entsprechenden Koeffizienten in nachfolgenden Tabellen dargestellt.

Die in Tabelle 35 dargestellten Ergebnisse zeigen durchweg hohe Faktorladungen auf den Globalfaktoren sowie auf den spezifischen Faktoren. Die Ladungen auf den auf die Itemformulierung bezogenen Faktoren sind hingegen eher niedrig, teilweise sogar nicht signifikant. Insbesondere zeigt sich, dass diese Faktoren in beiden Fächern im Wesentlichen jeweils nur

auf ein Item besonders hoch laden (jeweils das vierte Item der Skala *Verständlichkeit* beim Ich-Bezug-Faktor bzw. das zweite Item der Skala *Thematische Motivierung* bei den Klassen-Bezug-Faktoren). Das heißt, es kann hier kaum noch von itemübergreifenden *Faktoren* gesprochen werden. Dazu tragen auch die erwartungswidrig negativen Ladungen der Klassen-Bezug-Faktoren auf einzelne Indikatoren bei. Es handelt sich hier also vielmehr um diffuse Residualvarianzkorrelationen, die inhaltlich nicht interpretierbar sind.

Da die Faktoren, die jeweils auf einen Indikator laden, unkorreliert sind, kann jedes Item in entsprechende Varianzkomponenten[126] zerlegt werden (vgl. Tabelle 36). Erstaunlicherweise sind die Varianzanteile auf dem Globalfaktor bei den ichbezogenen Itemformulierungen niedriger und auf den spezifischen Faktoren deutlich höher als bei den Items mit Klassen-Bezug (s. Zeilen „Mittelwerte", „Items mit Ich-Bezug" bzw. „ Items mit Klassen-Bezug"). Offensichtlich führt die klassenbezogene Itemformulierung stärker zu einer globalen Beurteilung des Unterrichts. Dies ist möglicherweise darauf zurückzuführen, dass bei der Integration der vielfältigen Meinungen der Mitschüler – die aus dem Gedächtnis abgerufen werden müssen – ein Gesamturteil der Mitschüler bezüglich des Unterrichts bei der entsprechenden Lehrkraft gebildet wird – anstatt separater Urteile für jedes Unterrichtsmerkmal.

Die in der letzten Zeile der Tabelle aufgeführten Mittelwerte der Varianzkomponenten aller Items (separat für Englisch und Deutsch) zeigen, dass bei diesen drei Unterrichtsmerkmalen (die ohnehin hoch miteinander korreliert sind; vgl. Tabelle 28) die Globalfaktoren durchschnittlich mehr zur Varianzaufklärung bezüglich der Indikatoren beitragen als die spezifischen Faktoren. Mit einem Varianzanteil von etwa 10% – bezogen auf die *true score*-Varianzkomponenten – sind die auf die Itemformulierung bezogenen „Faktoren" von eher geringer Bedeutung.

Tabelle 37 zeigt die latenten Korrelationen der Faktoren des in Abbildung 6 dargestellten Modells. Erstaunlich sind die immer noch substanziellen Interkorrelationen der fachspezifischen Unterrichtswahrnehmungen – obwohl die jeweiligen Globalfaktoren einen relativ großen Teil der gemeinsamen Varianz erklären (s. Tabelle 36). Diese Zusammenhänge lassen sich also nicht allein auf ein globales Urteil bezüglich der Lehrkraft zurückführen. Hier könnten auf die eigene Person bezogene Wahrnehmungen eine wichtige Rolle spielen. Besonders

[126] Auf eine Adjustierung der Varianzanteile anhand des in Kapitel 3.4.2 vorgeschlagenen Verfahrens wurde hier aus zwei Gründen verzichtet: (1) Die *grouping error*-Varianzanteile liegen bei allen Indikatoren in ähnlicher Größenordnung (vgl. dazu Tabelle 20). Es handelt sich also um einen nahezu konstanten Fehlerterm. (2) In einem Fall (Item 1 der Skala *Thematische Motivierung* im Fach Englisch) musste die Residualvarianz auf Null fixiert werden. Hier sind offensichtlich die *grouping error*-Varianzanteile verschiedener Indikatoren miteinander korreliert, was mit einer Überschätzung der *true score*-Varianzen – bzw. einer Unterschätzung der Residualvarianz – einhergeht. Die Adjustierung würde dazu führen, dass die auf die zugrunde liegende metrische Variable bezogenen Varianzanteile in der Summe die theoretische Obergrenze von 100% überschreiten.

deutlich wird dies bei der relativ hohen fachübergreifenden Korrelation der wahrgenommenen *Verständlichkeit* des Unterrichts, der möglicherweise auf den Einfluss des globalen akademischen Selbstkonzepts der jeweiligen Schülerinnen und Schüler zurückgeführt werden kann. Bei allen drei hier untersuchten Unterrichtswahrnehmungen (*Thematische Motivierung*, *Verständlichkeit* und *Schülerorientierung*) ist es sehr viel naheliegender als bei den Merkmalen *Strukturiertheit* und *Klassenführung*, dass die „eigene Person" der Schülerin bzw. des Schülers besonders in das Urteil involviert ist.

Der relativ schwache Zusammenhang zwischen den Globalurteilen kann in dem Sinne interpretiert werden, dass Schülerinnen und Schüler relativ deutlich zwischen den entsprechenden (hier in einigen Fällen in beiden Fächern identischen) Lehrkräften in den Unterrichtsfächern Deutsch und Englisch differenzieren. Eine Interpretation der Interkorrelationen der Ich- bzw. Klassen-Bezug-Faktoren erscheint aus den oben genannten Gründen wenig sinnvoll. (Der Vollständigkeit halber sind sie jedoch in Tabelle 37 aufgeführt.)

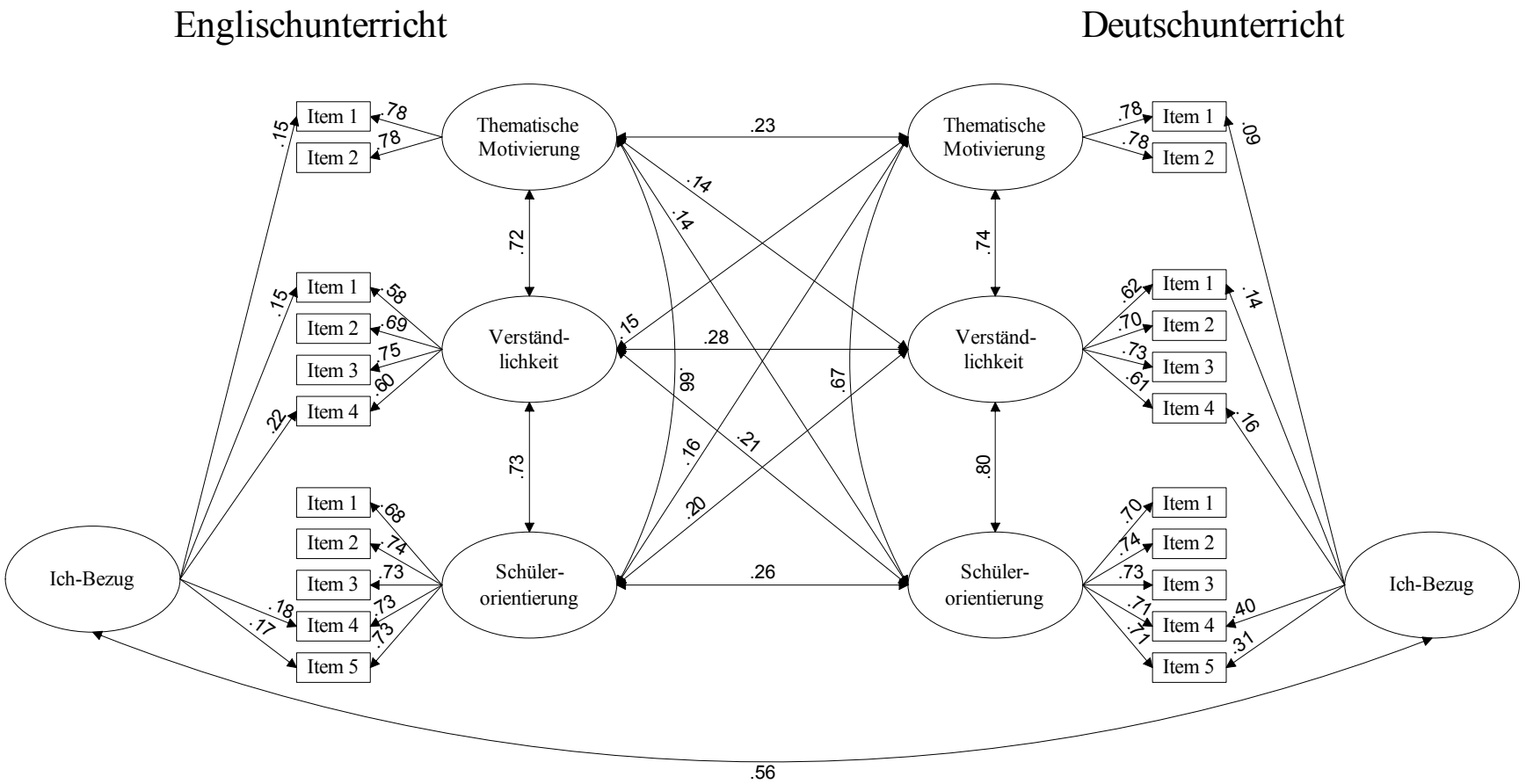

Abbildung 3: Einfluss der Itemformulierung: CFA-CTC(M-1)-Modell mit Ich-Bezug als Methodenfaktor und Klassen-Bezug als Referenzmethode[127]

[127] Alle Koeffizienten sind standardisiert. Alle Ladungen wurden frei geschätzt. Auf die Darstellung der Residualvarianzen (Fehlerterme) und der Varianzen der latenten Variablen (alle auf 1 fixiert) wurde aus Gründen der Übersichtlichkeit verzichtet. Modellgüte-Indizes: $Chi^2 = 1930.284$, Scaling Correction Factor = 1.067; DF = 379; CFI = 0.975; TLI = 0.970; RMSEA = 0.023; SRMR (zwischen) = 0.047; SRMR (innerhalb) = 0.019

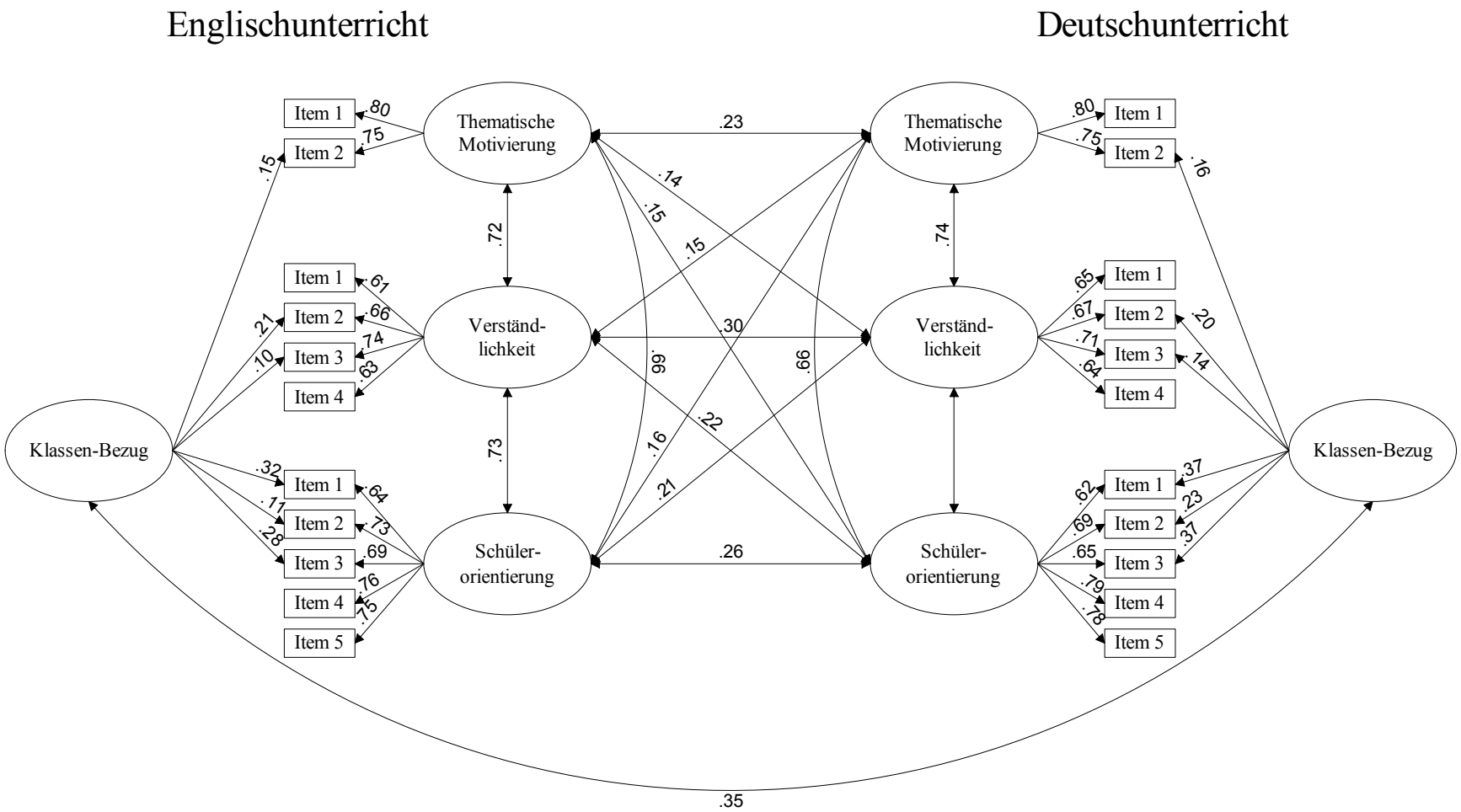

Abbildung 4: Einfluss der Itemformulierung: CFA-CTC(M-1)-Modell mit Klassen-Bezug als Methodenfaktor und Ich-Bezug als Referenzmethode[128]

[128] Alle Koeffizienten sind standardisiert. Alle Ladungen wurden frei geschätzt. Auf die Darstellung der Residualvarianzen (Fehlerterme) und der Varianzen der latenten Variablen (alle auf 1 fixiert) wurde aus Gründen der Übersichtlichkeit verzichtet. Modellgüte-Indizes: $Chi^2 = 1781.629$, Scaling Correction Factor = 1.069; DF = 377; CFI = 0.978; TLI = 0.973; RMSEA = 0.022; SRMR (zwischen) = 0.047; SRMR (innerhalb) = 0.019

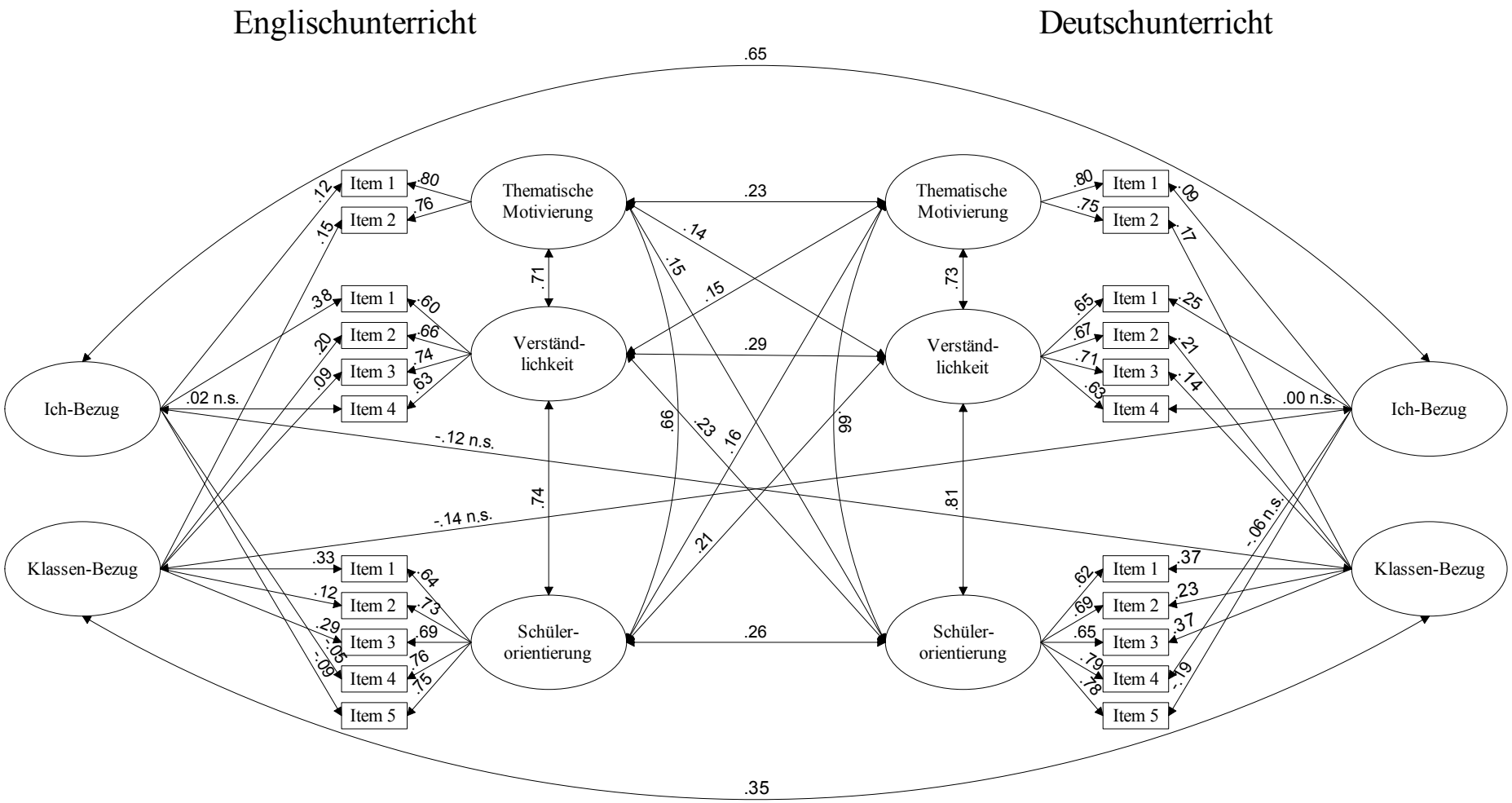

Abbildung 5: Einfluss der Itemformulierung: CFA-CTUM-Modell[129]

[129] Alle Koeffizienten sind standardisiert. Alle Ladungen wurden frei geschätzt. Auf die Darstellung der Residualvarianzen (Fehlerterme) und der Varianzen der latenten Variablen (alle auf 1 fixiert) wurde aus Gründen der Übersichtlichkeit verzichtet. Modellgüte-Indizes: $Chi^2 = 1607.145$, Scaling Correction Factor = 1.042; DF = 364; CFI = 0.980; TLI = 0.975; RMSEA = 0.021; SRMR (zwischen) = 0.047; SRMR (innerhalb) = 0.018

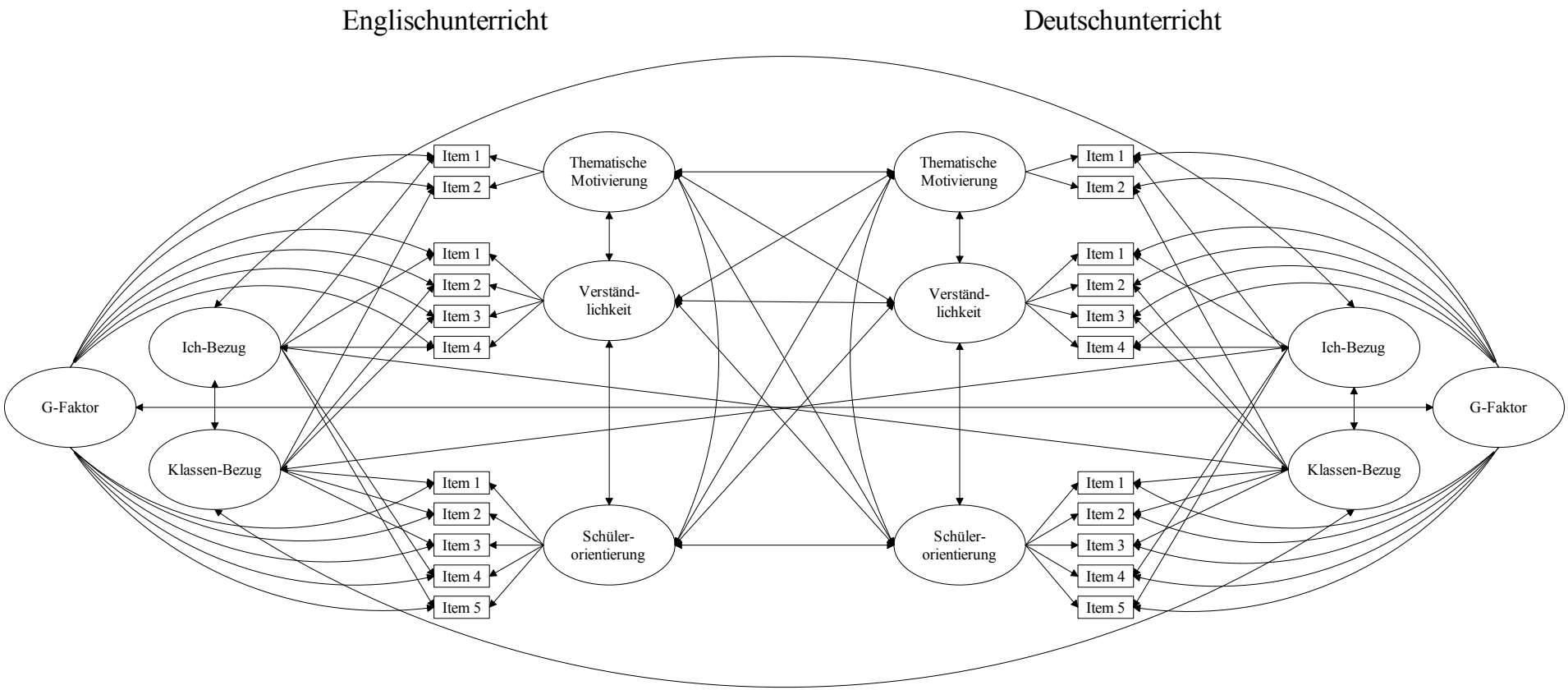

Abbildung 6: Einfluss der Itemformulierung: CFA-CTCM-Modell mit fachspezifischen Globalfaktoren der Unterrichtswahrnehmung

Tabelle 35: Standardisierte Ladungen[130] des CFA-CTCM-Modells mit fachspezifischen Globalfaktoren

		Englisch				Deutsch			
Spezifische Faktoren	Itemnummer (I) = Ich-Bezug (K) = Klassen-Bezug	Global-faktor	spezifische Faktoren	Ich-Bezug	Klassen-Bezug	Global-faktor	spezifische Faktoren	Ich-Bezug	Klassen-Bezug
Thematische Motivierung	1 (I)	.46	.88	.13	–	.49	.77	.15	–
	2 (K)	.58	.36	–	.43	.60	.38	–	.36
Verständlichkeit	1 (I)	.33	.56	.07	–	.35	.61	.07	–
	2 (K)	.51	.45	–	.09	.54	.42	–	.08
	3 (K)	.51	.53	–	.12	.53	.49	–	.08
	4 (I)	.44	.37	.58	–	.49	.30	.61	–
Schülerorientierung	1 (K)	.70	.16	–	-.08 n.s.	.69	.19	–	-.09 n.s.
	2 (K)	.61	.42	–	-.04 n.s.	.62	.38	–	-.12
	3 (K)	.70	.25	–	-.13	.68	.27	–	-.17
	4 (I)	.52	.63	.04	–	.46	.72	.09	–
	5 (I)	.61	.40	.05	–	.56	.50	.09	–

[130] Alle Ladungen – mit Ausnahme der mit „n.s.“ (nicht signifikant) gekennzeichneten – sind mindestens auf dem 5%-Niveau (zweiseitige Tests) signifikant. Modellgüte-Indizes: $Chi^2 = 1037.851$, Scaling Correction Factor = 1.024; DF = 340; CFI = 0.989; TLI = 0.985; RMSEA = 0.016; SRMR (zwischen) = 0.047; SRMR (innerhalb) = 0.014. Die nicht-signifikante negative Residualvarianz des ersten Items der Skala *Thematische Motivierung* im Fach Englisch wurde auf Null fixiert.

Tabelle 36: Varianzkomponenten der Indikatoren des CFA-CTCM-Modells mit fachspezifischen Globalfaktoren (Angaben in Prozent)

		Englisch					Deutsch				
Spezifische Faktoren	Itemnummer (I) = Ich-Bezug (K) = Klassen-Bezug	Global-faktor	spezifische Faktoren	Ich-Bezug	Klassen-Bezug	Residual-varianz	Global-faktor	spezifische Faktoren	Ich-Bezug	Klassen-Bezug	Residual-varianz
Thematische Motivierung	1 (I)	21.0	77.3	1.8	–	0.0[131]	24.0	59.4	2.2	–	14.3
	2 (K)	33.4	13.0	–	18.7	34.9	35.5	14.4	–	12.6	37.6
Verständlichkeit	1 (I)	10.7	30.8	0.5	–	58.0	12.3	37.5	0.4	–	49.8
	2 (K)	25.7	20.0	–	0.8	53.5	29.5	17.8	–	0.6	52.1
	3 (K)	26.3	28.3	–	1.4	44.0	28.2	24.1	–	0.6	47.0
	4 (I)	19.0	13.6	33.5	–	33.7	23.9	9.2	37.6	–	29.3
Schülerorientierung	1 (K)	48.7	2.6	–	0.6	48.1	47.9	3.6	–	0.7	47.7
	2 (K)	36.8	17.5	–	0.1	45.6	38.3	14.1	–	1.3	46.2
	3 (K)	48.9	6.2	–	1.8	43.1	46.1	7.4	–	2.8	43.8
	4 (I)	27.1	39.7	0.2	–	33.0	21.5	52.0	0.7	–	25.7
	5 (I)	37.6	16.0	0.3	–	46.1	31.1	25.3	0.8	–	42.7
Mittelwerte											
Items mit Ich-Bezug		23.1	35.5	7.3	–	34.2	22.6	36.7	8.4	–	32.4
Items mit Klassen-Bezug		36.6	14.6	–	3.9	44.9	37.6	13.6	–	3.1	45.7
				Ich- bzw. Klassen-Bezug					Ich- bzw. Klassen-Bezug		
gesamt		30.5	24.1	5.4		40.0	30.8	24.1	5.5		39.7

[131] Die nicht-signifikante Residualvarianz wurde auf Null fixiert.

Tabelle 37: Interkorrelationen der Faktoren des CFA-CTCM-Modells mit fachspezifischen Globalfaktoren

		Englisch						Deutsch		
		Global-faktor	Thematische Motivierung	Verständ-lichkeit	Schüler-orientierung	Ich-Bezug-Faktor	Klassen-Bezug-Faktor	Thematische Motivierung	Verständ-lichkeit	Ich-Bezug-Faktor
Englisch	Verständlichkeit	–	.38***							
	Schülerorientierung	–	.27***	.42***						
	Ich-Bezug-Faktor	–	–	–	–					
	Klassen-Bezug-Faktor	–	–	–	–	.33***				
Deutsch	Globalfaktor	.23***	–	–	–	–	–			
	Thematische Motivierung	–	.17***	.08***	.09**	–	–			
	Verständlichkeit	–	.09***	.38***	.22***	–	–	.39***		
	Schülerorientierung	–	.09***	.22***	.29***	–	–	.28***	.55***	
	Ich-Bezug-Faktor	–	–	–	–	.23**	-.01	–	–	
	Klassen-Bezug-Faktor	–	–	–	–	.13*	.35***	–	–	.33**

4.6 Determinanten der geteilten und der nicht-geteilten Wahrnehmungskomponenten

Im folgenden Abschnitt geht es um die Frage, ob Unterrichtswahrnehmungen durch die in diesem Zusammenhang häufig diskutierten Merkmale Geschlecht, Fachleistung und Schulnoten sowie durch das Globalurteil bezüglich der Lehrkraft „vorhergesagt" werden können. Es sei darauf hingewiesen, dass „Vorhersagen" hier nicht im Sinne kausaler Einflüsse interpretiert werden kann, da dazu wenigstens alle Merkmale zu zwei Messzeitpunkten hätten erfasst werden müssen (wobei sich Kausalität im strikten Sinne auch dann nicht nachweisen lässt). Entsprechend sind hier auch die Begriffe „Prädiktor" bzw. „Determinante" im Sinne *möglicher* kausaler Einflussgrößen zu verstehen. Mit Ausnahme des Geschlechts (bzw. des Mädchenanteils auf Klassenebene) ist hier die Richtung des Zusammenhangs theoretisch nicht eindeutig: Sowohl bei den Testleistungen und Noten als auch beim Globalurteil bezüglich der Lehrkraft sind Zusammenhänge – verglichen mit den Analysen hier – in umgekehrter Richtung ebenfalls plausibel (also: Unterrichtswahrnehmung als Prädiktor) und stellen die „übliche" Analysestrategie dar. Beim auf die Lehrkraft bezogenen Globalurteil wird dies am Ende dieses Kapitels deutlich: Dort wird untersucht, ob die individuellen Unterrichtswahrnehmungen das auf die Lehrkraft bezogene Globalurteil beeinflussen und vor allem, ob diese Einflüsse über Lehrkräfte (also über Klassen) hinweg variieren.

In Tabelle 38 sind die um Bildungsgangeffekte bereinigten Interkorrelationen der Prädiktoren auf beiden Ebenen (innerhalb von Klassen und zwischen Klassen) dargestellt. Betrachtet man die Leistungsindikatoren (Testleistungen bzw. Schulnoten), so zeigen sich mittlere bis sehr hohe Stabilitäten: Der niedrigste Zusammenhang beträgt r = .46 (Deutschnote, innerhalb von Klassen), der höchste r = .94 (Englischleistung, Klassenebene). Die fachübergreifenden Zusammenhänge der jeweiligen Leistungsmaße zum jeweiligen Messzeitpunkt liegen zwischen r = .41 (Testleistung zu Beginn der neunten Klassenstufe, innerhalb von Klassen) und r = .74 (Testleistung am Ende der neunten Klassenstufe, Klassenebene). Bei den Korrelationen zwischen den Leistungstests und den Schulnoten besteht ein deutlicher Unterschied zwischen den beiden Analyseebenen: Innerhalb von Klassen zeigen sich moderate Zusammenhänge der beiden „Fachleistungsperspektiven" (von r = .29 bei der Deutschleistung zum ersten[132] bis r = .42 bei der Englischleistung zum zweiten Messzeitpunkt), während auf

[132] Es sei darauf hingewiesen, dass die Schulnoten am Ende der achten Klassenstufe erteilt wurden, während die Testleistung am Anfang der neunten Klassenstufe erfasst wurde. Insofern handelt es sich hier streng genommen um zwei verschiedene Messzeitpunkte. Daneben ist zu beachten, dass die Noten am Ende der achten Klassenstufe teilweise von ehemaligen Lehrkräften erteilt wurden.

Klassenebene lediglich der auf die Englischleistung am Ende der neunten Klassenstufe bezogene Zusammenhang signifikant von Null verschieden ist. Eine mögliche Erklärung hierfür könnte ein klasseninternes Bezugssystem bei der Benotung sein: Die leistungsstarken Schülerinnen und Schüler innerhalb einer Klasse erhalten „gute“ Noten, (relativ) unabhängig davon, wie gut diese Schülerinnen und Schüler im absoluten Vergleich mit anderen Schülerinnen und Schülern aus anderen Klassen des gleichen Bildungsganges sind. Oder anders formuliert: „Gute“ Schülerinnen und Schüler erhalten – je nach Leistungsniveau der jeweiligen Klasse – unterschiedliche Noten. Die Unterschiede hinsichtlich der Benotung auf Klassenebene wären dann eher auf Milde- bzw. Strengetendenzen der jeweiligen Lehrkraft zurückzuführen. Die relativ hohen Korrelationen auf Klassenebene zwischen den jeweiligen Noten in den Fächern Deutsch und Englisch – die ja größtenteils von unterschiedlichen Lehrkräften stammen – sind möglicherweise Ausdruck eines übergreifenden „Benotungsmaßstabes“ der jeweiligen Schule.

Bezogen auf das Geschlecht lässt sich Tabelle 38 entnehmen, dass Mädchen innerhalb von Klassen durchweg höhere Werte (wenn auch das Ausmaß nicht besonders ausgeprägt ist) bei allen aufgeführten Prädiktoren aufweisen. Beim Mädchenanteil (Klassenebene) zeigen sich analoge Korrelationen, wobei einige Zusammenhänge nicht signifikant sind.

Bei den auf die jeweilige Englisch- bzw. Deutschlehrkraft bezogenen Globalurteilen liegen ebenfalls durchweg positive Zusammenhänge mit den übrigen Prädiktoren auf der Ebene innerhalb von Klassen vor (Ausnahmen: Globalurteil der Deutschlehrkraft und Deutsch- bzw. Englischleistung am Anfang der neunten Klassenstufe). Auf der Klassenebene hingegen sind hier nur zwei signifikante Korrelationen zu verzeichnen: Das auf die Deutschlehrkraft bezogene Globalurteil der Klasse korreliert schwach negativ (!) mit der Deutschleistung zu Beginn ($r = -.16$) sowie schwach positiv mit der Deutschnote am Ende der neunten Klassenstufe ($r = .20$).

Abschließend sei noch auf die erwartungswidrig negative Korrelation ($r = -.21$) auf Klassenebene zwischen Englischleistung und Deutschnote zum ersten Messzeitpunkt hingewiesen. Alle übrigen fachübergreifenden Korrelationen zwischen Leistungstests und Schulnoten auf Klassenebene sind nicht signifikant. Insofern erscheint eine Interpretation dieses Ergebnisses wenig sinnvoll (zumindest ohne umfassende, weiterführende Analysen).

Tabelle 38: Interkorrelationen der Prädiktoren[133] (innerhalb von Klassen unterhalb, zwischen Klassen oberhalb der Hauptdiagonale dargestellt; Stabilitäten grau markiert)

	Geschlecht	Englischleistung Anfang 9. Klassenstufe	Englischleistung Ende 9. Klassenstufe	Deutschleistung Anfang 9. Klassenstufe	Deutschleistung Ende 9. Klassenstufe	Englischnote (umgepolt) Ende 8. Klassenstufe	Englischnote (umgepolt) Ende 9. Klassenstufe	Deutschnote (umgepolt) Ende 8. Klassenstufe	Deutschnote (umgepolt) Ende 9. Klassenstufe	Globalurteil bzgl. der Englischlehrkraft	Globalurteil bzgl. der Deutschlehrkraft
Geschlecht[134]		.23***	.28***	.19*	.29***	.14	.14	.19**	.11	.00	-.05
Englischleistung Anfang 9. Klassenstufe	.11***		.94***	.66***	.68***	.05	.12	-.21**	-.11	.02	-.04
Englischleistung Ende 9. Klassenstufe	.14***	.75***		.65***	.74***	.09	.15*	-.12	-.07	.04	-.08
Deutschleistung Anfang 9. Klassenstufe	.14***	.41***	.39***		.74***	-.05	.03	-.12	-.10	-.09	-.16*
Deutschleistung Ende 9. Klassenstufe	.19***	.42***	.44***	.69***		.00	.05	-.02	-.08	-.02	-.11
Englischnote (umgepolt) Ende 8. Klassenstufe	.10***	.40***	.39***	.24***	.25***		.67***	.48***	.45***	-.03	.03
Englischnote (umgepolt) Ende 9. Klassenstufe	.08***	.44***	.42***	.26***	.28***	.53***		.46***	.54***	.10	-.02
Deutschnote (umgepolt) Ende 8. Klassenstufe	.21***	.28***	.28***	.29***	.31***	.55***	.37***		.70***	-.05	.08
Deutschnote (umgepolt) Ende 9. Klassenstufe	.18***	.27***	.26***	.28***	.31***	.34***	.57***	.46***		-.07	.20**
Globalurteil bzgl. der Englischlehrkraft	.08***	.09***	.12***	.06***	.07***	.10***	.18***	.07***	.07***		-.04
Globalurteil bzgl. der Deutschlehrkraft	.11***	.00	.04**	.02	.07***	.04**	.04*	.09***	.17***	.16***	

[133] Die Interkorrelationen wurden auf der Basis eines Gesamtmodells mit allen (latenten) Unterrichtsmerkmalen in beiden Fächern geschätzt. Dabei wurden auf Klassenebene Bildungsgangeffekte kontrolliert (*dummy*-Kodierung für die Bildungsgänge Realschule und Gymnasium). Die von Mplus ausgegebenen standardisierten Koeffizienten (hier Korrelationen der Residuen) sind auf die Gesamtstreuung der jeweiligen Variablen auf der jeweiligen Ebene bezogen. Entsprechend stellen diese Koeffizienten – sofern die Residualvarianzen kleiner als die ursprünglichen Varianzen sind – Unterschätzungen der üblicherweise angegebenen Partialkorrelationen dar. Deshalb wurden die Korrelationskoeffizienten auf Klassenebene anhand der standardisierten Residualvarianzen (s^2_{std,ζ_x} bzw. s^2_{std,ζ_y}, die multipliziert mit den entsprechenden Varianzen – s^2_x bzw. s^2_y – die Residualvarianzen – $s^2_{\zeta_x}$ bzw. $s^2_{\zeta_y}$ – ergeben) wie folgt adjustiert:

$$r_{xy,adj} = \frac{Cov(x,y)}{\sqrt{s^2_x \cdot s^2_{std,\zeta_x}}\sqrt{s^2_y \cdot s^2_{std,\zeta_y}}} = \frac{Cov(x,y)}{\sqrt{s^2_x}\sqrt{s^2_y}} \cdot \frac{1}{\sqrt{s^2_{std,\zeta_x} \cdot s^2_{std,\zeta_y}}} = \frac{r_{xy}}{\sqrt{s^2_{std,\zeta_x} \cdot s^2_{std,\zeta_y}}}$$

Es sei darauf hingewiesen, dass die Zusammenhänge zwischen Englisch- und Deutschleistung (insbesondere auf der Ebene innerhalb von Klassen, aufgrund der – verglichen mit der Klassenebene – niedrigeren Reliabilität der *scores*) die wahren Zusammenhänge unterschätzen, da die PVs aus getrennten Skalierungen stammen (und nicht aus einer einzigen mit allen Deutsch- und Englischtests zu beiden Messzeitpunkten). Die Signifikanzangaben beziehen sich jeweils auf zweiseitige Tests (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$).

Modellgüte-Indizes (gemittelte Angaben für fünf *plausible values* der Testleistungen): $Chi^2 = 6540.975$, DF = 1369; CFI = 0.961; TLI = 0.947; RMSEA = 0.021; SRMR (zwischen) = 0.072; SRMR (innerhalb) = 0.022

[134] *dummy*-Kodierung: 0 = Jungen, 1 = Mädchen. Auf Klassenebene ist dieses Merkmal als Mädchenanteil zu interpretieren.

Die Zusammenhänge zwischen den Prädiktoren und den Unterrichtsmerkmalen aus Schülersicht sind in Tabelle 39 (innerhalb von Klassen) bzw. Tabelle 40 (Klassenebene) dargestellt. Bezogen auf das Geschlecht zeigen sich hier die bereits in Kapitel 4.2 auf Itemebene berichteten Befunde: Mädchen schätzen den Unterricht etwas positiver ein als Jungen (Ausnahmen sind hier die *Thematische Motivierung* im Fach Englisch sowie die *Strukturiertheit* in beiden Fächern), während der Mädchenanteil (Klassenebene) keine Rolle bei der geteilten Unterrichtswahrnehmung spielt.

Tabelle 39: Zusammenhänge zwischen Unterrichtswahrnehmungen und Prädiktoren innerhalb von Klassen[135]

	Unterrichtsmerkmal	Geschlecht[136]	Englischleistung Anfang 9. Klassenstufe	Englischleistung Ende 9. Klassenstufe	Deutschleistung Anfang 9. Klassenstufe	Deutschleistung Ende 9. Klassenstufe	Englischnote (umgepolt) Ende 8. Klassenstufe	Englischnote (umgepolt) Ende 9. Klassenstufe	Deutschnote (umgepolt) Ende 8. Klassenstufe	Deutschnote (umgepolt) Ende 9. Klassenstufe	Globalurteil bzgl. der Englischlehrkraft	Globalurteil bzgl. der Deutschlehrkraft
Englisch	Thematische Motivierung	.03	$.09^{***}$	$.12^{***}$	$.05^{**}$	$.05^{**}$	$.10^{***}$	$.18^{***}$	$.06^{***}$	$.04^{*}$	$.56^{***}$	$.12^{***}$
	Verständlichkeit	$.06^{**}$	$.24^{***}$	$.29^{***}$	$.14^{***}$	$.20^{***}$	$.22^{***}$	$.31^{***}$	$.13^{***}$	$.11^{***}$	$.55^{***}$	$.10^{***}$
	Schülerorientierung	$.07^{***}$	$.10^{***}$	$.13^{***}$	$.09^{***}$	$.11^{***}$	$.13^{***}$	$.17^{***}$	$.11^{***}$	$.10^{***}$	$.52^{***}$	$.12^{***}$
	Strukturiertheit	-.03	.02	.03	$-.03^{*}$	-.02	$.05^{***}$	$.09^{***}$	.02	.00	$.37^{***}$	$.11^{***}$
	Klassenführung	$.05^{**}$	$.06^{***}$	$.10^{***}$	$.05^{**}$	$.08^{***}$	$.07^{***}$	$.11^{***}$	$.04^{*}$	$.07^{***}$	$.42^{***}$	$.16^{***}$
Deutsch	Thematische Motivierung	$.11^{***}$	-.01	.00	.03	$.07^{***}$	.01	.01	$.10^{***}$	$.16^{***}$	$.13^{***}$	$.59^{***}$
	Verständlichkeit	$.09^{***}$	$.06^{***}$	$.12^{***}$	$.09^{***}$	$.18^{***}$	$.08^{***}$	$.07^{***}$	$.15^{***}$	$.22^{***}$	$.13^{***}$	$.57^{***}$
	Schülerorientierung	$.11^{***}$	$.05^{**}$	$.08^{***}$	$.06^{***}$	$.11^{***}$	$.06^{***}$	$.06^{***}$	$.11^{***}$	$.17^{***}$	$.13^{***}$	$.53^{***}$
	Strukturiertheit	.02	-.01	.01	$-.05^{**}$	.00	.03	.02	.03	$.06^{***}$	$.10^{***}$	$.38^{***}$
	Klassenführung	$.08^{***}$	.02	$.06^{***}$	.02	$.08^{***}$	$.05^{***}$	.03	$.07^{***}$	$.08^{***}$	$.11^{***}$	$.45^{***}$

Bei den Leistungsindikatoren sind auf der Ebene innerhalb von Klassen erwartungsgemäß überwiegend positive Korrelationen zu verzeichnen (einzige negative Korrelationen: Deutschtestleistung zu Beginn der neunten Klassenstufe und *Strukturiertheit* im Fach Englisch bzw. Deutsch). Das heißt, leistungsstärkere Schülerinnen und Schüler beurteilen den Unterricht positiver als ihre Mitschüler. Da die Leistungen in beiden Fächern relativ hoch miteinander korreliert sind (s. Tabelle 38), ist es auch nicht erstaunlich, dass in vielen Fällen Leistungsindikatoren in einem Fach mit Unterrichtswahrnehmungen im jeweils anderen Fach zusammenhängen.

[135] Modellgüte-Indizes vgl. Fußnote 133. Die Signifikanzangaben beziehen sich jeweils auf zweiseitige Tests (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$).

[136] *dummy*-Kodierung: 0 = Jungen, 1 = Mädchen. Auf Klassenebene ist diese Variable als Mädchenanteil zu interpretieren.

Tabelle 40: Zusammenhänge zwischen Unterrichtswahrnehmungen und Prädiktoren auf Klassenebene[137]

	Unterrichtsmerkmal	Geschlecht[138]	Englischleistung Anfang 9. Klassenstufe	Englischleistung Ende 9. Klassenstufe	Deutschleistung Anfang 9. Klassenstufe	Deutschleistung Ende 9. Klassenstufe	Englischnote (umgepolt) Ende 8. Klassenstufe	Englischnote (umgepolt) Ende 9. Klassenstufe	Deutschnote (umgepolt) Ende 8. Klassenstufe	Deutschnote (umgepolt) Ende 9. Klassenstufe	Globalurteil bzgl. der Englischlehrkraft	Globalurteil bzgl. der Deutschlehrkraft
Englisch	Thematische Motivierung	-.05	.01	.04	-.13*	-.05	.00	.10	-.13	-.13	.92***	-.07
	Verständlichkeit	.00	.06	.09	-.07	.02	.07	.16*	-.05	-.01	.87***	-.01
	Schülerorientierung	.09	.10	.16*	-.04	.05	.05	.17*	-.07	-.10	.87***	-.04
	Strukturiertheit	-.06	-.09	-.08	-.17*	-.14*	.07	.11	.06	.03	.72***	.13
	Klassenführung	.07	.08	.12	-.05	.13	-.11	-.12	-.13	-.09	.54***	-.07
Deutsch	Thematische Motivierung	-.07	-.07	-.09	-.18**	-.13*	.10	.03	.11	.21**	-.03	.90***
	Verständlichkeit	-.05	-.05	-.04	-.16*	-.07	.10	.05	.08	.15*	.01	.87***
	Schülerorientierung	.01	.05	.06	-.06	.04	.10	.09	.10	.20**	.01	.87***
	Strukturiertheit	-.16	-.15*	-.18**	-.24**	-.21**	.14	.07	.13	.18*	.03	.73***
	Klassenführung	-.03	-.04	-.05	-.11	-.01	.13*	.14*	.03	.01	.09	.56***

Dies gilt auch für die Klassenebene. Dort finden sich – bezogen auf die signifikanten Korrelationen – ebenfalls durchweg positive Zusammenhänge bei den Schulnoten. Bei den Leistungstests hingegen sind – ebenfalls nur auf die signifikanten Korrelationen bezogen – sowohl positive als auch negative (Englischleistung) bzw. durchweg negative (!) Korrelationen (Deutschleistung) zu verzeichnen. Theoretisch wären unkorrelierte Testleistungen und Unterrichtswahrnehmungen hier noch erklärbar, weil das fachspezifische akademische Selbstkonzept (im Sinne einer internen Repräsentation des individuellen Leistungsniveaus) von Schülerinnen und Schülern auf Klassenebene zwar mit den Fachnoten zusammenhängt – wenngleich in geringerem Maße als auf der Ebene innerhalb von Klassen – (vgl. dazu auch A. Helmke, Schrader, Wagner, Nold & Schröder, in Druck-b; Wagner, Helmke, Schrader, Eichler, Thomé & Willenberg, in Druck), nicht aber mit der Testleistung im jeweiligen Fach (die wiederum nicht oder nur sehr schwach mit der Fachnote zusammenhängt). Dies ist insofern plausibel, als die (parallel zu den Schülerbefragungen erhobenen und nicht individuell rückgemeldeten[139]) Testleistungen – im Gegensatz zu den Schulnoten – keinen direkten Einfluss auf das akademische Selbstkonzept von Schülerinnen und Schülern ausüben können.

Betrachtet man die Zusammenhänge zwischen den auf die jeweilige Fachlehrkraft bezogenen Globalurteilen und den Unterrichtswahrnehmungen, so zeigen sich hier die mit Ab-

[137] Zur Adjustierung der um Bildungsgangeffekte kontrollierten Korrelationen sowie der gemittelten Modellgüte-Indizes vgl. Fußnote 133. Die Signifikanzangaben beziehen sich jeweils auf zweiseitige Tests (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$).

[138] *dummy*-Kodierung: 0 = Jungen, 1 = Mädchen. Auf Klassenebene ist diese Variable als Mädchenanteil zu interpretieren.

[139] Zur Rückmeldung und Rezeption von Daten im Rahmen von DESI vgl. Rolff und von der Gathen (in Druck).

stand höchsten Korrelationen – und zwar auf beiden Ebenen –, während fachübergreifend auf der Ebene innerhalb von Klassen relativ niedrige, auf der Klassenebene keine Zusammenhänge vorliegen. Dabei sind die auf die relativ objektiven und eher niedrig-inferenten Unterrichtsmerkmale *Strukturiertheit* und *Klassenführung* bezogenen Korrelationen jeweils niedriger als die auf die mit den Eigenschaften und Einstellungen der Schülerinnen und Schüler konfundierten Merkmale (*Thematische Motivierung*, *Verständlichkeit* und *Schülerorientierung*) bezogenen.

Der negative Zusammenhang von Deutschtestleistung und verschiedenen Unterrichtswahrnehmungen auf Klassenebene zeigt sich auch in der negativen Korrelation (r = -.16) zwischen Deutschtestleistung (Anfang der neunten Klassenstufe[140]) und dem auf die Deutschlehrkraft bezogenen – und hoch mit den Unterrichtswahrnehmungen korrelierten – Globalurteil (vgl. Tabelle 38). Bei den Testleistungen im Fach Englisch hingegen bestehen auf Klassenebene keine Zusammenhänge mit dem entsprechenden Globalurteil. Analog dazu kovariieren die Deutschnote am Ende der neunten Klassenstufe und das Globalurteil bezüglich der Deutschlehrkraft positiv miteinander (r = .20), während kein entsprechender Zusammenhang im Fach Englisch besteht. Das zeigt, dass es sich hierbei nicht um eine „Besonderheit" des spezifischen Deutschtests handelt, sondern dass sich die Einflüsse der Fachleistung auf die Unterrichtswahrnehmungen in den Fächern Deutsch und Englisch offensichtlich unterscheiden.

In Tabelle 41 sind die (metrischen) ICCs[141] der hier untersuchten Prädiktoren aufgeführt. Neben den erwartungsgemäß hohen ICCs bei den Testleistungen – hierfür sind zu einem großen Teil Bildungsgangunterschiede verantwortlich – zeigen sich bei den Noten deutlich niedrigere relative Varianzanteile auf Klassenebene, da dort u.a. wesentlich geringere Differenzen zwischen den Bildungsgängen bestehen als bei den Tests. Beachtlich sind aber vor allem die deutlich über Null liegenden relativen Übereinstimmungen der Schülerinnen und Schüler hinsichtlich des auf ihre jeweiligen Lehrkräfte bezogenen Globalurteils. Diese Niveauunterschiede auf Klassenebene zeigen, dass komplette *Klassen* – und nicht nur individuelle Schülerinnen und Schüler – bestimmte Lehrkräfte insgesamt positiver bzw. negativer beurteilen. Die Einflüsse des Globalurteils auf die individuelle Wahrnehmung des Unterrichts werden entsprechend bei der Aggregation, wie in Tabelle 40 bereits dargestellt, *nicht* effektiv eliminiert.

[140] Die entsprechende Korrelation (r = -.11) mit der Deutschtestleistung am Ende der neunten Klassenstufe ist nur marginal signifikant.
[141] Da es sich bei der Geschlechtszugehörigkeit eindeutig um eine kategoriale Variable handelt, wurde diese hier nicht mit aufgeführt. Die für nominale Variablen bestimmte ICC bezüglich der Geschlechtszugehörigkeit beträgt ICC = .16.

Der Effekt auf Klassenebene ist sogar stärker ausgeprägt als auf der Ebene innerhalb von Klassen. Dieses Ergebnis unterstützt ausgesprochen deutlich die *general impression* Halo-Hypothese, und das auf beiden Analyseebenen (geteilte und nicht-geteilte Unterrichtswahrnehmungen).

Tabelle 41: ICCs der Prädiktoren und gemeinsame Varianzanteile der kategorisierten Variablen mit den zugrunde liegenden intervallskalierten, normalverteilten Variablen innerhalb von Klassen

	ICC(metrisch)		r_g		r_{gw}^2	
	Englisch	Deutsch	Englisch	Deutsch	Englisch	Deutsch
Globalurteil bzgl. der jeweiligen Lehrkraft	0.182	0.209	0.927	0.930	0.829	0.830
Note (Ende 8. Klassenstufe)	0.153	0.141	0.950	0.938	0.885	0.860
Note (Ende 9. Klassenstufe)	0.120	0.123	0.948	0.934	0.884	0.854
Testleistung (Anfang 9. Klassenstufe)	0.633	0.554	–	–	–	–
Testleistung (Ende 9. Klassenstufe)	0.632	0.604	–	–	–	–

Da das Globalurteil bezüglich der Lehrkraft sowie die Schulnoten auch als kategorisierte Variablen einer zugrunde liegenden intervallskalierten, normalverteilten Variable betrachtet werden können, wurden hierfür die entsprechenden Korrekturfaktoren (r_g, r_{gw}^2) berechnet, mit deren Hilfe sich die in Tabelle 39 dargestellten Korrelationen adjustieren lassen (vgl. Tabelle 42). Insgesamt zeigen sich eher geringfügige Unterschiede zwischen den adjustierten und den unadjustierten Korrelationen (die größte Differenz findet sich bei der Korrelation zwischen Thematischer Motivierung und dem Globalurteil – jeweils im Fach Deutsch: $r = .59$ vs. $r_{adj} = .65$).

Tabelle 42: Adjustierte Zusammenhänge zwischen Unterrichtswahrnehmungen und Prädiktoren innerhalb von Klassen[142]

	Unterrichtsmerkmal	Englischnote (umgepolt) Ende 8. Klassenstufe	Englischnote (umgepolt) Ende 9. Klassenstufe	Deutschnote (umgepolt) Ende 8. Klassenstufe	Deutschnote (umgepolt) Ende 9. Klassenstufe	Globalurteil bzgl. der Englischlehrkraft	Globalurteil bzgl. der Deutschlehrkraft
Englisch	Thematische Motivierung	.10***	.19***	.06***	.04*	.61***	.13***
	Verständlichkeit	.23***	.33***	.14***	.12***	.60***	.11***
	Schülerorientierung	.13***	.18***	.11***	.10***	.57***	.13***
	Strukturiertheit	.06***	.09***	.02	.00	.40***	.12***
	Klassenführung	.07***	.12***	.04*	.07***	.46***	.17***
Deutsch	Thematische Motivierung	.01	.01	.11***	.17***	.14***	.65***
	Verständlichkeit	.09***	.08***	.16***	.24***	.14***	.62***
	Schülerorientierung	.07***	.06***	.12***	.18***	.14***	.58***
	Strukturiertheit	.03	.02	.03	.06***	.11***	.41***
	Klassenführung	.05***	.03	.07***	.09***	.12***	.50***

[142] Modellgüte-Indizes vgl. Fußnote 133. Die Signifikanzangaben beziehen sich jeweils auf zweiseitige Tests (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$).

Um die Zusammenhänge zwischen den verschiedenen, voneinander nicht unabhängigen Prädiktoren (vgl. Tabelle 38) und den einzelnen Unterrichtswahrnehmungen in ihrer Gesamtheit betrachten zu können, wurde zusätzlich zu den einfachen Korrelationen ein mehrebenenanalytisches Pfadanalysemodell berechnet. Dabei wurde jedes Unterrichtsmerkmal in jedem Fach als latente Variable (Ladung des jeweils ersten Indikators auf 1 fixiert), die von *allen* Prädiktoren beeinflusst wird, modelliert. Das heißt, die „Effekte" sind jeweils unter Konstanthaltung aller übrigen Prädiktoren zu interpretieren. Aufgrund der extrem hohen Kollinearität der Englischtestleistungen zu Beginn und am Ende des Schuljahres auf Klassenebene (r = .94; vgl. Tabelle 38) wurden alle Testleistungen hier durch die entsprechenden Leistungszuwächse (PV-Differenzwerte; vgl. Kap. 4.1) ersetzt. Die Interkorrelationen der Unterrichtsmerkmale (Residualvarianzen) sowie der Prädiktoren wurden zugelassen[143]. Dieses Modell ist als explorative Analyse zu verstehen, da angesichts der Modellkomplexität und der Datenlage (hohe Kollinearitäten zwischen den Prädiktoren, Unterrichtswahrnehmungen nur am Ende des Schuljahres erhoben) ein strikt konfirmatorisches Modell unangemessen erscheint.

Betrachtet man die Ergebnisse in Tabelle 43 auf der Ebene *innerhalb von Klassen* (untere Hälfte der Tabelle), dann zeigt sich, dass der Effekt des Geschlechts (erste Spalte) auch unter Kontrolle der übrigen Prädiktoren weitgehend erhalten bleibt – wenn auch etwas abgeschwächt. Lediglich bei der *Verständlichkeit* und der *Strukturiertheit* im Fach Englisch ergeben sich Veränderungen: Während der Einfluss auf das erstgenannte Merkmal – im Gegensatz zu den einfachen Korrelationen (vgl. Tabelle 39) – hier das 5%-Signifikanzniveau nicht mehr erreicht, wird der schwach negative Zusammenhang auf die *Strukturiertheit* hier signifikant.

Bei den Leistungszuwächsen (Testleistungen) zeigen sich signifikant positive Einflüsse auf die Unterrichtswahrnehmungen im jeweiligen Fach – Ausnahmen: *Strukturiertheit* im Fach Englisch und *Thematische Motivierung* im Fach Deutsch. Die auf die Unterrichtsmerkmale im jeweils anderen Fach bezogenen Regressionskoeffizienten der Leistungszuwächse hingegen sind nicht signifikant. Ausnahmen: Die *Verständlichkeit* wird durch den Leistungszuwachs im jeweils anderen Fach, die wahrgenommene *Schülerorientierung* im Fach Deutsch durch den Zuwachs der Englischleistung positiv beeinflusst.

[143] Es handelt sich hier um ein sogenanntes *bow-free pattern*-Modell, das als rekursives Modell eingestuft werden kann (vgl. Kline, 2005), weshalb sich der Nachweis der Identifikation des Strukturmodells erübrigt.

Tabelle 43: Vorhersage der Unterrichtswahrnehmungen innerhalb von Klassen und zwischen Klassen anhand der Prädiktoren: mehrebenenanalytisches Pfadanalysemodell[144] (Testleistungen und Schulnoten, die sich auf Unterrichtsmerkmale im selben Fach beziehen, sind grau markiert)

			Prädiktor (standardisierte Regressionsgewichte)								
		Unterrichtsmerkmal	Geschlecht[145]	Englischleistung (Zuwachs) Anfang/Ende 9. Klassenstufe	Deutschleistung (Zuwachs) Anfang/Ende 9. Klassenstufe	Englischnote (umgepolt) Ende 8. Klassenstufe	Englischnote (umgepolt) Ende 9. Klassenstufe	Deutschnote (umgepolt) Ende 8. Klassenstufe	Deutschnote (umgepolt) Ende 9. Klassenstufe	Bildungsgang[146] Realschule	Bildungsgang Gymnasium	R^2 in Prozent
zwischen Klassen	Englisch	Thematische Motivierung	-.07	.11	.12	-.07		-.16	-.19	-.10	-.02	10.1
		Verständlichkeit	-.04	.09	.12	-.03		-.18	-.04	.01	.12	8.8
		Schülerorientierung	.06	.17*	.08	-.09	.37***	-.12	-.22	-.02	.21*	17.7
		Strukturiertheit	-.09	.01	.04		.16	.08	-.09	-.19	-.45***	12.9
		Klassenführung	.05	.10	.25***	-.06	-.07	-.23	.09	.03	.10	12.0
	Deutsch	Thematische Motivierung	-.09	-.11	.13*	.14	-.16	-.08	.33*	-.13	-.23**	10.9
		Verständlichkeit	-.08	-.02	.14*	.12	-.09	-.08	.22	-.12	-.12	5.9
		Schülerorientierung	-.02	-.03	.16**	.05	-.04		.32*	-.12	-.05	9.1
		Strukturiertheit	-.17	-.16	.09	.16	-.09	.05	.19	-.24**	-.44***	18.9
		Klassenführung	-.08	-.10	.17**	.10	.17	-.02	-.08	-.03	-.09	6.2
innerhalb von Klassen	Englisch	Thematische Motivierung	.02	.05**	-.01	.00	.23***	.02	-.10***	-	-	4.1
		Verständlichkeit	.03	.08***	.05**	.07**	.33***	.01	-.11***	-	-	11.9
		Schülerorientierung	.05**	.05**	.02	.03	.15***	.04	-.02	-	-	3.7
		Strukturiertheit	-.03*	.02	.02	.01	.12***	.01	-.08***	-	-	1.4
		Klassenführung	.04*	.05**	.02	.01	.11***	-.02	.00	-	-	1.8
	Deutsch	Thematische Motivierung	.07***	.01	.04	-.04	-.12***	.05**		-	-	4.7
		Verständlichkeit	.04*	.07***	.09***	.01	-.09***		.23***	-	-	7.4
		Schülerorientierung	.07***	.04*	.06**	.02	-.07***	.03	.17***	-	-	4.2
		Strukturiertheit	.01	.03	.06***	.02	-.02	-.01	.06***	-	-	0.8
		Klassenführung	.06**	.04	.08**	.03	-.04*	.02	.08***	-	-	2.1

Eine entscheidende Rolle für die Unterrichtswahrnehmungen innerhalb von Klassen spielen die Schulnoten, und zwar insbesondere die am Ende der neunten Klassenstufe erhobenen erwarteten Fachnoten: Dort zeigen sich – unter Kontrolle der Noten am Ende der achten Klassenstufe sowie des jeweiligen Leistungszuwachses – durchweg hoch signifikante positive Einflüsse auf die Unterrichtswahrnehmungen im jeweiligen Fach und teilweise negative Einflüsse auf die Wahrnehmungen im jeweils anderen Fach. Diese negativen Einflüsse der Noten im

[144] Da es sich hier nicht um eine konfirmatorische Analyse handelt, bei der strikte Erwartungen bezüglich vorliegender (und nicht vorliegender!) Effekte mit erwarteten Richtungen vorgegeben sind, sondern um eine explorative Analyse möglicher Einflussgrößen, beziehen sich die Signifikanzangaben jeweils auf zweiseitige Tests (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$).
Modellgüte-Indizes (gemittelte Angaben für fünf *plausible values* der Testleistungszuwächse): $Chi^2 = 5808.776$, DF = 1193; CFI = 0.957; TLI = 0.944; RMSEA = 0.021; SRMR (zwischen) = 0.062; SRMR (innerhalb) = 0.021

[145] *dummy*-Kodierung: 0 = Jungen, 1 = Mädchen. Auf Klassenebene ist diese Variable als Mädchenanteil zu interpretieren.

[146] Die Kontrolle der Bildungsgänge erfolgte mithilfe von *dummy*-Kodierungen: Bildungsgang Realschule (0 = trifft nicht zu, 1 = trifft zu), Bildungsgang Gymnasium (0 = trifft nicht zu, 1 = trifft zu)

einen Fach auf die Unterrichtswahrnehmung im anderen Fach sind vermutlich darauf zurückzuführen, dass hier neben der auf den *Klassendurchschnitt* bezogenen auch die auf den *individuellen Notendurchschnitt* einer Schülerin bzw. eines Schülers bezogene Note eine Rolle spielt: Eine – gemessen am Klassendurchschnitt – besonders gute Englischnote wird in ihrem Effekt abgeschwächt, wenn gleichzeitig eine – ebenfalls am Klassendurchschnitt gemessen – besonders gute Deutschnote vorliegt, wohingegen der Effekt beim Vorliegen einer eher schlechten Deutschnote verstärkt wird.

Der Einfluss der Noten am Ende der achten Klassenstufe auf die Unterrichtswahrnehmungen ist eher gering. Nur bei drei Unterrichtsmerkmalen sind positive Einflüsse der Noten im jeweils identischen Fach zu verzeichnen (*Verständlichkeit* im Fach Englisch bzw. Deutsch sowie *Thematische Motivierung* im Fach Deutsch). Die Varianzaufklärung (rechte Spalte) ist erwartungsgemäß am höchsten bei den mit den Eigenschaften und Einstellungen der Schülerinnen und Schüler konfundierten Merkmale (*Thematische Motivierung*, *Schülerorientierung* und insbesondere *Verständlichkeit*), während die eher objektiv zu beurteilenden Merkmale (*Strukturiertheit* und *Klassenführung*) kaum durch die Prädiktoren erklärt werden.

Betrachtet man die Effekte auf der Klassenebene, so zeigen sich beim Mädchenanteil keine signifikanten Koeffizienten. Bei den Leistungszuwächsen sind große Unterschiede zwischen den Inhaltsbereichen zu beobachten: Während der auf die Englischtestleistung bezogene Zuwachs nur in einem Fall einen signifikant positiven Effekt aufweist (*Schülerorientierung* im Fach Englisch), sind beim Deutschtest fünf signifikant positive Regressionsgewichte zu verzeichnen: Mit Ausnahme der *Strukturiertheit* sind alle auf das Fach Deutsch bezogenen Unterrichtswahrnehmungen positiv durch den Leistungszuwachs beeinflusst; daneben zeigt sich ein erwartungswidrig hoch signifikant positiver Effekt auf die *Klassenführung* im Fach Englisch. Dieser Effekt deutet sich bereits in den Korrelationen der Prädiktoren mit den Unterrichtsmerkmalen auf Klassenebene an (vgl. Tabelle 40), und zwar in Form der numerisch relativ großen Differenz zwischen den (nicht signifikanten) Korrelationskoeffizienten von *Klassenführung* im Fach Englisch und den Deutschtestleistungen am Anfang bzw. Ende des Schuljahres ($r = -.05$ vs. $r = .13$). Eine inhaltliche Deutung dieses Effekts erscheint aber auf der Basis der vorliegenden Analysen nicht angebracht.

Bezogen auf die Schulnoten zeigt sich auf der Klassenebene ein vergleichbares Bild wie auf der Ebene innerhalb von Klassen: Die Noten am Ende des achten Schuljahres sind hier jedoch *generell* unbedeutend für die Vorhersage der Unterrichtswahrnehmungen. Die erwarteten Noten am Ende des neunten Schuljahres weisen bedeutsame Zusammenhänge mit den Merkmalen im jeweils identischen Fach auf, die mit Eigenschaften und Einstellungen der

Schülerinnen und Schüler konfundiert sind (*Thematische Motivierung*, *Verständlichkeit* und *Schülerorientierung*; Ausnahme: Der Effekt der Deutschnote am Ende der neunten Klassenstufe auf die *Verständlichkeit* im Fach Deutsch ist mit $p < .10$ nur marginal signifikant), nicht jedoch mit den eher objektiv beurteilbaren.

Dieser Einfluss der Schulnote auf die Unterrichtswahrnehmungen lässt sich – aufgrund der oben genannten Einschränkungen jedoch mit Vorsicht – im Sinne der *grading leniency*-Hypothese interpretieren: Der Unterricht von Lehrkräften, die verglichen mit dem tatsächlichen Leistungszuwachs relativ „gute" Noten vergeben, wird teilweise positiver bewertet. Dies ist hier besonders bedeutsam, da es sich um Effekte auf Klassenebene handelt. Das heißt, die Einflüsse auf Individualebene werden durch die Aggregation nicht eliminiert. Es sei allerdings darauf hingewiesen, dass sich die mithilfe der Leistungstests erfassten Inhaltsbereiche – die hier ohnehin nur durch jeweils ein Testmodul pro Fach repräsentiert sind – und die der Benotung zugrunde liegenden nur zum Teil überlappen (können). Insofern ist der Effekt der Konstanthaltung des Leistungszuwachses durch die Testergebnisse nur teilweise gewährleistet. Andererseits legen die auf Klassenebene (im Gegensatz zur Ebene innerhalb von Klassen!) weitgehend entkoppelten Testleistungen und Schulnoten (s.o.) aber auch eine Interpretation der Noten im Sinne eines klasseninternen Bezugssystems nahe, bei der die Klassenebene kaum eine Rolle spielt.

Die Bildungsgang-Prädiktoren existieren nur auf Klassenebene (da jeweils komplette Klassen einem bestimmten Bildungsgang angehören) und sind aufgrund ihrer *dummy*-Kodierung als Abweichung bezüglich der Unterrichtswahrnehmung innerhalb des entsprechenden Bildungsganges vom Bildungsgang Hauptschule (dem Referenzbildungsgang) zu interpretieren. Im Bildungsgang Realschule zeigt sich – bei Konstanthaltung der übrigen Prädiktoren – nur bei der *Strukturiertheit* im Fach Deutsch ein negativer Effekt, während im Bildungsgang Gymnasium negative Effekte bei der *Strukturiertheit* (in beiden Fächern) sowie der Thematischen Motivierung im Fach Deutsch bzw. ein positiver Effekt bei der *Schülerorientierung* im Fach Englisch vorliegen.

Die höchste Varianzaufklärung auf Klassenebene findet sich bei der *Strukturiertheit* im Fach Englisch ($r^2 = 18.9\%$). Diese ist allerdings im Wesentlichen auf Bildungsgangunterschiede zurückzuführen. Im Gegensatz zu allen übrigen Unterrichtswahrnehmungen finden sich bei der *Strukturiertheit* (in beiden Fächern) ansonsten keine weiteren signifikanten Effekte. Auch der relativ große Unterschied bezüglich der Varianzaufklärung der *Schülerorientierung* in beiden Fächern ($r^2 = 17.7\%$ im Fach Englisch, $r^2 = 9.1\%$ im Fach Deutsch) ist we-

sentlich auf Bildungsgangunterschiede zurückzuführen (ein signifikant positiver Effekt des Bildungsganges Gymnasium findet sich nur im Fach Englisch).

Wie bereits zu Beginn des Kapitels erwähnt, soll abschließend untersucht werden, ob das auf die Lehrkraft bezogene Globalurteil (als abhängige Variable) in einem über Klassen hinweg konstanten oder variierenden Zusammenhang mit den individuellen Unterrichtswahrnehmungen (hier Prädiktoren) steht. Dazu wurden sogenannte *random slope*-Modelle auf der Basis von Ebene-1-Faktor*scores*[147] (*within factor scores*) mit HLM[148] berechnet[149]. Dabei wurden jeweils fachspezifisch alle hier untersuchten Unterrichtswahrnehmungen in ein Gesamtmodell aufgenommen. Alle Variablen (inklusive der abhängigen) wurden der besseren Interpretierbarkeit der Ergebnisse halber z-standardisiert. Da hier davon ausgegangen werden muss, dass alle Regressionsgewichte, die sich auf die miteinander relativ hoch korrelierten Prädiktoren beziehen, über Klassen hinweg variieren, ist dieses Verfahren hier einer schrittweisen Vorgehensweise (jeweils ein Modell für jeden Prädiktor) vorzuziehen. Im Falle (hoch) korrelierter Prädiktoren erscheinen konstante Regressionskoeffizienten als variierend, wenn (tatsächlich) ein anderer, im Modell nicht berücksichtigter Prädiktor einen variierenden Einfluss auf die abhängige Variable ausübt. Dies lässt sich an folgendem Beispiel verdeutlichen: Eine abhängige Variable y wird von zwei miteinander korrelierten Ebene-1-Prädiktoren x_v und x_k vorhergesagt, wobei der Einfluss von x_v auf y über Klassen hinweg variiert, während x_k einen konstanten Einfluss auf y ausübt. Das heißt, x_k übt einen konstanten Einfluss auf y aus, *wenn der (variierende) Einfluss von x_v kontrolliert wird*. Wird hingegen der Einfluss von x_v auf y nicht kontrolliert, dann zeigen sich bezogen auf x_k variierende Regressionsgewichte. Sind die Prädiktoren hingegen unkorreliert, dann sind solche Artefakte nicht zu erwarten, da der Einfluss unkorrelierter Prädiktoren nicht kontrolliert werden muss[150].

Die auf das Fach Englisch bezogenen Ergebnisse des Einflusses der Unterrichtswahrnehmungen auf das Globalurteil bezüglich der Lehrkraft sind in Tabelle 44 dargestellt. Das zu-

[147] Die Faktor*scores* basieren auf dem in Kap. 4.2, Fußnote 89 beschriebenen Modell. Im Prinzip hätten hier auch die gruppenzentrierten Faktor*scores* aus dem Ein-Ebenen-Modell verwendet werden können. Die *within factor scores* stellen lediglich etwas präzisere Schätzungen dar, da sich die ebenenspezifischen Ladungen meist unterscheiden (vgl. Kap. 4.2), was im Ein-Ebenen-Modell nicht berücksichtigt wird.

[148] In HLM 6.04 stehen leider keine Angaben zur Signifikanz der Korrelationen bzw. Kovarianzen auf Klassenebene zur Verfügung. Die fehlenden Kennzeichnungen in den folgenden Tabellen bedeuten deshalb nicht, dass die berichteten Korrelationen statistisch nicht bedeutsam sind.

[149] Prinzipiell sind hier auch Analysen auf der Basis latenter Faktoren in Mplus möglich. Aufgrund der vorliegenden Modellkomplexität sind solche Analysen aber (derzeit) nicht möglich.

[150] Theoretisch ist aber auch hier ein Artefakt im Sinne eines statistischen Nachweises eines tatsächlich nicht vorliegenden variierenden Zusammenhangs denkbar: Wenn die beiden Prädiktoren einen konstanten Einfluss auf die abhängige Variable ausüben, der Zusammenhang der Prädiktoren aber über Klassen hinweg variiert (wenngleich der mittlere Zusammenhang Null beträgt), dann besteht die Möglichkeit, dass in Modellen, die jeweils nur einen Prädiktor berücksichtigen, *random slopes* nachgewiesen werden.

grunde liegende Modell enthält nur noch vier Prädiktoren: Da im vollständigen Modell für den Faktor *Strukturiertheit* weder ein signifikanter *fixed effect* noch ein signifikanter *random effect* vorlag, d.h. weder das mittlere Regressionsgewicht noch die Variation der Regressionsgewichte statistisch bedeutsam waren, wurde dieser Prädiktor aus dem Modell entfernt[151]. Bei den verbleibenden vier Prädiktoren zeigen sich bezüglich der Regressionsgewichte durchweg signifikante Mittelwerte und Streuungen. Im Mittel sind die Effekte der klasseninternen Abweichungen der Unterrichtswahrnehmungen positiv mit dem Globalurteil verknüpft, wobei – legt man das 95%-Intervall für die Variation der Regressionsgewichte zugrunde (also: Mittelwert ± 1.96 Standardabweichungen) – bei allen Prädiktoren in einem nennenswerten Teil der Klassen auch negative Koeffizienten vorliegen. Daneben zeigen sich sowohl negative (*Thematische Motivierung*, *Schülerorientierung*) als auch positive (*Verständlichkeit*, *Klassenführung*) Zusammenhänge zwischen der globalen Wahrnehmung der Englischlehrkraft auf Klassenebene und den jeweiligen Regressionsgewichten. Der absolut betrachtet stärkste Zusammenhang findet sich bei der Thematischen Motivierung (r = -.49): Mit zunehmend positiver globaler Wahrnehmung der Lehrkraft auf Klassenebene sinkt der Einfluss der subjektiv wahrgenommenen Thematischen Motivierung auf das Globalurteil.

Tabelle 44: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Englisch[152]: Mittelwert und Streuung der Regressionsgewichte, Interzept-Regressionsgewicht-Korrelationen

	Regressionsgewicht		Korrelation Interzept (Globalurteil), Regressionsgewicht
	Mittelwert	Variation zwischen Klassen (SD)	
Thematische Motivierung	0.20***	0.19***	-.49
Verständlichkeit	0.19***	0.25***	.33
Schülerorientierung	0.10***	0.14**	-.26
Klassenführung	0.07***	0.16***	.18

Betrachtet man die Interkorrelationen der über Klassen hinweg variierenden Regressionsgewichte (Tabelle 45), so zeigen sich stark negative bis mäßig positive Zusammenhänge im Bereich von r = -.73 (*Verständlichkeit*, *Thematische Motivierung*) bis r = .29 (*Klassenführung*, *Schülerorientierung*). Die Tatsache, dass die Interkorrelationen nicht durchweg sehr hoch sind und in die gleiche Richtung zeigen, ist hier insofern von Bedeutung, als es sich bei der Auswahl an Unterrichtsmerkmalen lediglich um einen sehr kleinen Ausschnitt denkbarer Ein-

[151] Wählt man zur Vorhersage des Globalurteils lediglich den Prädiktor *Strukturiertheit* (Fach Englisch), dann zeigt sich – entsprechend den obigen Anmerkungen zu korrelierten Prädiktoren in *random slope*-Modellen – eine hoch signifikante *slope*-Variation.

[152] Ein *Loglikelihood-Ratio*-Test zwischen dem *random intercept*- und dem *random slope*-Modell zeigt (zusätzlich zu den signifikanten *slope*-Varianzen) eine signifikant bessere Passung des *random slope*-Modells ($Chi^2 = 120.008$, DF = 14, $p < .001$). Die Signifikanzangaben zu den mittleren Regressionsgewichten beziehen sich jeweils auf zweiseitige Tests (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$).

flüsse auf die globale Wahrnehmung der Lehrkraft handelt. Insofern wäre es prinzipiell möglich, dass beispielsweise in Wirklichkeit nur ein einziges, hier nicht erfasstes (und entsprechend nicht kontrolliertes), – aber mit den verwendeten Prädiktoren hoch korreliertes Merkmal in seinem Einfluss auf das Globalurteil variiert. Für das hier zugrunde gelegte Mehrebenen-Regressionsmodell könnte dies bedeuten, dass sich substantielle Variationen der Regressionsgewichte zeigen – obwohl bezüglich der untersuchten Prädiktoren in Wirklichkeit keine *random slopes* vorliegen. In diesem Falle aber müssten die Interkorrelationen der *slope*-Variationen sehr hoch positiv bzw. negativ ausfallen.

Tabelle 45: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Englisch: Interkorrelationen der Regressionsgewichte

	Thematische Motivierung	Verständlichkeit	Schülerorientierung
Verständlichkeit	-.73		
Schülerorientierung	.16	-.54	
Klassenführung	-.20	-.46	.29

Wie im Fach Englisch, so zeigt sich auch im Fach Deutsch kein signifikanter mittlerer Einfluss der *Strukturiertheit* auf das Globalurteil bezüglich der Lehrkraft (vgl. Tabelle 46), wohingegen hier die entsprechenden Regressionsgewichte mit einer Standardabweichung von SD = 0.13 hoch signifikant über Klassen hinweg variieren. Verglichen mit der Analyse im Fach Englisch sind die mittleren Regressionsgewichte bezüglich der Thematischen Motivierung bzw. der *Verständlichkeit* deutlich höher bzw. niedriger[153]. Die Streuungen der Regressionsgewichte sowie die Korrelationen der Interzepte mit den Regressionsgewichten sind insgesamt betrachtet auf einem etwas niedrigeren Niveau als im Fach Englisch. Gleiches gilt für die in Tabelle 47 dargestellten Interkorrelationen der Regressionsgewichte.

Zusammengenommen sprechen diese Ergebnisse dafür, dass für die Bildung des Globalurteils der Lehrkraft die hier untersuchten Unterrichtswahrnehmungen im Fach Englisch eine größere Rolle spielen als im Fach Deutsch. Ausgenommen davon ist jedoch die wahrgenommene *Strukturiertheit*, die im Fach Englisch weder im Mittel noch hinsichtlich der Variation über Klassen hinweg für die Bildung des Globalurteils bedeutsam ist. Im Fach Deutsch hingegen ist hier eine hoch signifikante Variation der Regressionsgewichte auf Klassenebene zu verzeichnen, wenngleich der mittlere Effekt auch hier statistisch nicht bedeutsam ist.

[153] Aufgrund der etwas größeren ICC des auf die Deutschlehrkraft bezogenen Globalurteils (vgl. Tabelle 41) sind hier insgesamt minimal niedrigere Koeffizienten – die sich auf die Gesamtstreuung beziehen – zu erwarten.

Tabelle 46: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Deutsch[154]: Mittelwert und Streuung der Regressionsgewichte, Interzept-Regressionsgewicht-Korrelationen

	Regressionsgewicht		Korrelation Interzept (Globalurteil), Regressionsgewicht
	Mittelwert	Variation zwischen Klassen (SD)	
Thematische Motivierung	0.27***	0.17**	-.31
Verständlichkeit	0.08*	0.16**	.30
Schülerorientierung	0.12***	0.13*	.04
Strukturiertheit	-0.01	0.13***	-.19
Klassenführung	0.11***	0.09**	-.09

Tabelle 47: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Deutsch: Interkorrelationen der Regressionsgewichte

	Thematische Motivierung	Verständ-lichkeit	Schülerorien-tierung	Strukturiert-heit
Verständlichkeit	-.44			
Schülerorientierung	-.42	-.38		
Strukturiertheit	-.18	-.37	.01	
Klassenführung	-.16	-.38	.08	.16

Interessant wäre hier auch eine parallele Modellierung der auf die Englisch- bzw. Deutschlehrkraft bezogenen Globalurteile in der Teilstichprobe von Klassen, die in beiden Fächern von unterschiedlichen Lehrkräften unterrichtet wurden. Eine hohe Korrelation zwischen den jeweiligen *slopes* analoger Unterrichtswahrnehmungen in beiden Fächern auf das jeweilige Globalurteil würde dafür sprechen, dass sich *Klassen* in ihrer Gewichtung des entsprechenden Merkmals unterscheiden, dass also die Gewichtung *nicht* auf Merkmale der Lehrkraft zurückgeführt werden kann. Ein entsprechendes Mplus-Modell konvergierte allerdings nicht – vermutlich aufgrund der hohen Modellkomplexität.

[154] Ein *Loglikelihood-Ratio*-Test zwischen dem *random intercept*- und dem *random slope*-Modell zeigt (zusätzlich zu den signifikanten *slope*-Varianzen) eine signifikant bessere Passung des *random slope*-Modells ($Chi^2 = 64.348$, $DF = 20$, $p < .001$). Die Signifikanzangaben zu den mittleren Regressionsgewichten beziehen sich jeweils auf zweiseitige Tests (* für $p < .05$, ** für $p < .01$ und *** für $p < .001$).

5. Diskussion

Im Folgenden werden die zentralen Ergebnisse dieser Untersuchung zur Frage der Unterrichtswahrnehmung aus Schülersicht zusammengefasst. Es folgt ein Ausblick auf noch zu klärende Forschungsfragen sowie methodische Probleme.

5.1 Zusammenfassung der Ergebnisse

In der vorliegenden Untersuchung wurde der Frage nachgegangen, wie Schülerinnen und Schüler der neunten Klassenstufe ihren Englisch- bzw. Deutschunterricht bezüglich fünf verschiedener theoretischer Konstrukte[155] wahrnehmen. Es konnte gezeigt werden, dass die Unterrichtswahrnehmungen über Klassen hinweg systematisch variieren. Die Interrater-Reliabilitäten (ICCs) lagen, verglichen mit den Ergebnissen anderer Untersuchungen, insgesamt in einem mittleren bis hohen Bereich.

Im Rahmen von mehrebenenanalytischen konfirmatorischen Faktorenanalysen wurde der Frage nachgegangen, inwiefern sich die Konstrukte auf der Ebene innerhalb von Klassen bzw. zwischen Klassen miteinander vergleichen lassen. Zunächst zeigte sich, dass eine identische Zuordnung von Indikatoren zu Faktoren auf beiden Ebenen bei akzeptabler Modellpassung möglich ist. Dabei unterschieden sich allerdings häufig die Messmodelle der jeweils im Hinblick auf die Indikatorenzuordnung identischen Faktoren auf beiden Ebenen, so dass die Annahme isomorpher Konstrukte auf beiden Ebenen nicht gerechtfertigt erscheint. In diesen Fällen sind die parallelen Faktoren auf beiden Ebenen eher als analoge Konstrukte im Sinne des *fuzzy composition process models* (Bliese, 2000) zu interpretieren. Dies ist insofern nachteilig, als analoge Konstrukte auf Aggregatebene inhaltlich in der Regel schwieriger zu beschreiben sind, weshalb solche Aggregatmerkmale wohl häufig mithilfe des Begriffs „Klima“ umschrieben werden.

Um den Einfluss der Kommunikation mit Mitschülerinnen bzw. Mitschülern auf die individuelle Unterrichtswahrnehmung zu untersuchen, wurde die Geschlechtszugehörigkeit als „Kommunikationsbarriere“ betrachtet, d.h. es wurde aufgrund der Ergebnisse von Studien zu sozialen Netzwerken in Schulklassen davon ausgegangen, dass Mädchen nicht bzw. nur wenig mit Jungen in ihrer Klasse kommunizieren und umgekehrt. Eine höhere Kommunikation innerhalb von „Subgruppen“ sollte eigentlich zu höheren Übereinstimmungen bzw. Interrater-Reliabilitäten führen, wenn man davon ausgeht, dass die Meinungen der Mitschülerinnen bzw. Mitschüler in die individuelle Urteilsbildung einfließen. Erwartungsgemäß zeigten sich

[155] Es handelte sich dabei um die folgenden Unterrichtswahrnehmungen aus Schülersicht: Thematische Motivierung, Verständlichkeit, Schülerorientierung, Strukturiertheit, Klassenführung.

– verglichen mit den auf die gesamten Klassen bezogenen ICCs – deutlich höhere Übereinstimmungen in der Gruppe der Mädchen, nicht hingegen bei den Jungen, wo die ICCs fast durchweg signifikant niedriger waren als bei den Mädchen. Möglicherweise kommunizieren Jungen (insgesamt bzw. unterrichtsbezogen) weniger als Mädchen oder gewichten ihr eigenes Urteil stärker.

Die unterschiedlichen Unterrichtswahrnehmungen bezüglich der Gruppe der Mädchen bzw. der Gruppe der Jungen innerhalb von Klassen wurden auch anhand sogenannter *random slope*-Modelle nachgewiesen: Die Regressionsgewichte für die *dummy*-kodierte Prädiktorvariable „Geschlecht" variierte bei allen hier untersuchten Items signifikant über Klassen hinweg, d.h. die Itemmittelwerte für die Gruppe der Mädchen unterscheiden sich jeweils von denen der Jungen.

Was die Differenziertheit der Unterrichtswahrnehmungen – und die damit verbundene Frage der diskriminanten Validität – anbelangt, so zeigten sich mittlere bis hohe Zusammenhänge zwischen den nicht-geteilten („subjektiven") und hohe bis sehr hohe zwischen den geteilten („objektiven") Wahrnehmungskomponenten im jeweiligen Unterrichtsfach. Die hohen Interkorrelationen auf der Ebene innerhalb von Klassen lassen sich jedoch nicht einfach im Sinne unterschiedlicher Milde- bzw. Strengetendenzen der Schülerinnen und Schüler erklären, da die fachübergreifenden Zusammenhänge insgesamt nur sehr schwach ausgeprägt sind. Ob also Schülerinnen und Schüler beispielsweise ihren Englischunterricht verglichen mit ihren Mitschülern eher „streng" bewerten, sagt relativ wenig über das jeweilige Urteil bezüglich des Deutschunterrichts aus.

Bei den teilweise extrem hohen Zusammenhängen der jeweils fachspezifischen Konstrukte auf Klassenebene lässt sich ohne Einbeziehung anderer Datenquellen (z.B. Videoratings) nicht beurteilen, ob diese Interkorrelationen den „tatsächlichen" Zusammenhängen (aus einer hypothetisch „objektiven" Sicht) entsprechen oder ob sie Methodenartefakte im Sinne einer verzerrten geteilten Wahrnehmung des Unterrichts aus Schülersicht darstellen. Allerdings sprechen die relativ niedrigen bis nicht vorhandenen fachübergreifenden Zusammenhänge der Konstrukte auf der Klassenebene zumindest für ein gewisses Maß an verzerrter Wahrnehmung, da hier eigentlich aufgrund darüber liegender Ebenen (wie z.B. die Schulebene; aufgrund unterschiedlicher Lehrerausbildungen aber auch die Länderebene bzw. die Ebene der Bildungsgänge) zumindest moderat positive Zusammenhänge zu erwarten wären.

Dass diese niedrigen bzw. nicht vorhandenen fachübergreifenden Interkorrelationen der geteilten Unterrichtswahrnehmungen nicht auf unterschiedliche Wahrnehmungen des jeweiligen Fachs zurückzuführen sind, wird an den teilweise extrem hohen fachübergreifenden Zu-

sammenhängen in der Teilstichprobe der Klassen deutlich, die in den beiden Fächern Deutsch und Englisch von derselben Lehrkraft unterrichtet worden sind. Hier zeigen sich auch bei den nicht-geteilten Unterrichtswahrnehmungen durchgängig signifikant höhere fachübergreifende Zusammenhänge als in der Gruppe der Schülerinnen und Schüler mit unterschiedlichen Lehrkräften in beiden Fächern. Diese Ergebnisse sprechen für eine stark lehrkraftbezogene geteilte sowie nicht-geteilte Unterrichtswahrnehmung.

Betrachtet man die Wahrnehmungen der verschiedenen Unterrichtsmerkmale in beiden Fächern, so kann lediglich bei der *Klassenführung* von vollständig messinvarianten Faktoren auf beiden Ebenen in beiden Fächern ausgegangen werden, während bei der *Strukturiertheit* unterschiedliche Interzepte vorliegen. Bei allen übrigen Merkmalen sind bereits die Ladungen der analogen Faktoren nicht identisch (Ausnahme: *Verständlichkeit* auf Klassenebene). Dies zeigt, dass – mit Ausnahme des Fachbezugs („Englischlehrkraft" bzw. „Deutschlehrkraft") – identisch formulierte Items je nach Fach teilweise unterschiedlich wahrgenommen werden. Das heißt, in einigen Fällen kann hier nur von analogen und nicht von isomorphen Konstrukten ausgegangen werden. Als praktische Konsequenz ergibt sich daraus, dass Unterschiede bezüglich der Mittelwerte bzw. Varianzen der fachspezifischen Unterrichtswahrnehmungen teilweise nicht miteinander vergleichbar sind. Bei den beiden bezüglich ihrer Ladungen invarianten Faktoren (*Strukturiertheit* und *Klassenführung*) zeigen sich keine Unterschiede bezüglich der subjektiven Varianzkomponenten, während diese bei den geteilten Wahrnehmungen im Fach Deutsch jeweils größer als im Fach Englisch sind. Dieses Ergebnis spricht für eine gewisse „Objektivität" der geteilten Wahrnehmungen, da trotz größerer wahrgenommener Unterschiede im Fach Deutsch die subjektiven Varianz-Komponenten in beiden Fächern identisch bleiben.

Der Einfluss des Adressatenbezugs der Itemformulierung (Ich- vs. Klassen-Bezug) ist insgesamt beurteilt wohl eher von geringer Bedeutung. Dies ist insofern ein interessantes Ergebnis, als dieser Differenzierung aus theoretischer Sicht durchaus eine gewisse Bedeutsamkeit zukommt. So unterscheidet etwa Chan (1998) im Rahmen seiner Kompositionsmodelle hier zwischen *direct consensus* und *referent-shift consensus model*, wobei ersterem die ich-, letzterem die klassenbezogene Itemformulierung zuzuordnen ist. Auch aus Sicht der kognitiv fundierten Survey-Forschung sollte es eine gewisse Rolle spielen, ob Schülerinnen und Schüler lediglich ihre eigene Sichtweise wiedergeben, oder ob sie die Sichtweise ihrer Mitschüler mit einbeziehen sollen, was eine erheblich anspruchsvollere kognitive Aufgabe darstellt. Allerdings wird dort angenommen, dass auch im Falle von klassenbezogenen Formulierungen zunächst ein eigenes Urteil gebildet wird, das dann lediglich aufgrund vorliegender Informatio-

nen von Mitschülern und der wahrgenommenen Ähnlichkeit mit den Mitschülern adjustiert wird.

Möglicherweise ist dieses Ergebnis zumindest teilweise darauf zurückzuführen, dass hier Items untersucht wurden, die sich neben dem Adressatenbezug auch hinsichtlich ihres Inhaltes unterscheiden. Allerdings erscheint es auch unplausibel, dass Schülerinnen und Schüler über derart spezifische Informationen ihrer Mitschüler verfügen, die – je nach konkretem Iteminhalt – zu stark unterschiedlichen Bewertungen führen. Dagegen sprechen auch die deutlich höheren Varianzanteile der klassenbezogenen Itemformulierungen auf einem jeweiligen fachspezifischen Globalfaktor. Offensichtlich führt der Einbezug der Mitschüler-Perspektive zu einer stärker globalen Beurteilung des Unterrichts. Dies ist insofern erwartungsgemäß, als die vorliegenden Informationen von Mitschülern zu konkreten Iteminhalten vermutlich häufig eine geringe „Passung" aufweisen und somit eher globale Einschätzungen der Mitschüler in die Urteile einfließen.

Bei den in der Literatur häufig genannten potentiellen Prädiktoren der Unterrichtswahrnehmungen (Geschlecht, Testleistungen, Schulnoten), die im Sinne eines verzerrenden Einflusses diskutiert werden, lassen sich folgende Ergebnisse aus einem Regressionsmodell mit allen Kovariaten – die Ergebnisse sind entsprechend jeweils unter Kontrolle aller übrigen Prädiktoren zu interpretieren – festhalten:

Auf der Ebene *innerhalb von Klassen* werden die hier untersuchten Unterrichtsmerkmale…

1. …von Mädchen meist etwas positiver als von Jungen beurteilt.
2. …von Schülerinnen und Schülern mit höheren Testleistungszuwächsen (insbesondere im jeweiligen Fach) in der Regel positiver beurteilt.
3. …zumeist positiver eingeschätzt, wenn „bessere" Noten im jeweiligen Fach am Ende des Schuljahres erwartet werden, während die jeweilige Fachnote am Ende des vorangegangenen Schuljahres (die allerdings z.T. auch nicht von derselben Lehrkraft stammt) kaum eine Rolle spielt.
4. …tendenziell weniger günstig eingeschätzt, wenn „bessere" Noten im jeweils anderen Fach am Ende des Schuljahres erwartet werden.

Auf der *Klassenebene* werden die hier untersuchten Unterrichtsmerkmale…

1. …mit zunehmenden Leistungszuwächsen beim Deutschtest positiver beurteilt, während der Leistungszuwachs im Englischtest hier nahezu keine Rolle spielt.

2. …mit zunehmend „besseren“ Noten am Ende des Schuljahres teilweise positiver beurteilt, während auch hier – analog zu den Ergebnissen auf der Ebene innerhalb von Klassen – die Note am Ende des vorangegangenen Schuljahres keine Rolle spielt.

Das letztgenannte Ergebnis ist hier von größter Bedeutung, da es mit der *grading leniency*-Hypothese, wonach „gute“ Noten zu einer „guten“ Beurteilung des Unterrichts führen, in Einklang steht. Wenn man eine – aufgrund der Datenlage – vorsichtige Interpretation in dieser Richtung wagt, dann könnte der vorliegende Zusammenhang dafür sprechen, dass der Unterricht von Lehrkräften, die (bezogen auf den tatsächlichen Leistungszuwachs) „zu gute“ Noten vergeben, von Schülerinnen und Schülern positiver bewertet wird. Dieses Ergebnis ist deshalb besonders bedeutsam, da es sich hierbei um einen Effekt auf *Klassenebene* handelt, d.h. der „Effekt“ der Benotung wird durch die Aggregation der Daten nicht eliminiert. Insofern erscheint es möglich, dass Lehrkräfte tatsächlich die Beurteilung ihres Unterrichts durch die Benotung – hier im Sinne des Notenschnitts der Klasse – beeinflussen können.

Bei dieser Interpretation muss allerdings beachtet werden, dass die Testleistungen (bei weitem) nicht den gesamten Inhaltsbereich eines Fachs widerspiegeln können. Insofern ist eine Kontrolle des (tatsächlichen) Leistungszuwachses aufgrund der Testleistungen nur teilweise gewährleistet.

Die mit Abstand stärksten Zusammenhänge sowohl mit den geteilten als auch mit den nicht-geteilten Unterrichtswahrnehmungen finden sich beim Globalurteil der jeweiligen Fachlehrkraft. Dies spricht einerseits sehr stark für eine wenig differenzierte, also eher globale Wahrnehmung des Unterrichts von Schülerinnen und Schülern, die mit der *generell impression*-Halo-Hypothese im Einklang steht: Das globale Urteil bezüglich einer Lehrkraft überlagert die spezifischen Unterrichtswahrnehmungen.

Andererseits kann dieser Zusammenhang auch umgekehrt betrachtet werden: Die globale Wahrnehmung einer Lehrkraft, die ja überwiegend auf Erfahrungen mit der Lehrkraft in Unterrichtssituationen basiert, ist wohl zu einem großen Teil auch Resultat von Unterrichtswahrnehmungen. In der vorliegenden Untersuchung konnte gezeigt werden, dass bei einer solchen Betrachtung des Globalurteils alle hier untersuchten Unterrichtsmerkmale – mit Ausnahme der in diesem Kontext nicht bedeutsamen *Strukturiertheit* – einen im Mittel positiven Einfluss auf die globale Beurteilung der Lehrkraft ausüben. Darüber hinaus variieren diese Zusammenhänge über Klassen hinweg (einzige Ausnahme ist hier die *Strukturiertheit* im Fach Englisch – nicht aber im Fach Deutsch). Das heißt, in die globale Urteilsbildung gehen die einzelnen Unterrichtsmerkmale jeweils mit einem unterschiedlichen Gewicht ein.

Ein aus methodischer Sicht bedeutsames Ergebnis dieser Arbeit ist der auf der Basis einer Simulationsstudie erbrachte Nachweis der relativ guten Approximation der Parameter eines Populationsmodells mit einer Faktorstruktur auf zwei Ebenen und ordinalen Indikatoren durch ein analoges metrisches Modell. Die etwas unterschätzten Reliabilitäten der Indikatoren auf der Ebene innerhalb von Klassen lassen sich mithilfe eines auf Überlegungen von O'Brien (1985) basierenden Verfahrens erstaunlich gut korrigieren. Interessanterweise – und auch für die praktische Anwendung von größter Bedeutung – „reagieren" die für intervallskalierte Indikatoren (und nicht für ordinale!) zur Verfügung stehenden Modellanpassungsindizes auch hier sehr sensibel auf Modell-Fehlspezifikationen.

Damit sind Analysen deutlich umfangreicherer Zweiebenen-Faktor-Modelle möglich als bei der Verwendung theoretisch angemessenerer IRT-Modelle. Hier muss im Einzelfall entschieden werden, ob nicht – wie hier – Aussagen aus komplexen Modellen den methodisch korrekteren Ergebnissen aus kleinen „Teil"-Modellen vorzuziehen sind. So lässt sich beispielsweise das Ergebnis, dass in der vorliegenden Untersuchung in einem Gesamtmodell der Unterrichtswahrnehmung insgesamt zehn Faktoren unterschieden werden können, nicht einfach durch verschiedene Analysen von Teilen dieses Gesamtmodells ersetzen, da jeder Indikator Informationen für *alle* Faktoren liefert. Anders formuliert: Je komplexer das Modell, desto komplexer ist auch die Kovarianzmatrix, die durch das Modell (adäquat) repliziert werden muss.

Es muss jedoch einschränkend darauf hingewiesen werden, dass die ordinalen Daten in der vorliegenden Simulationsstudie nur wenig von der Normalverteilung abweichen. Bei größeren Verletzungen dieser Voraussetzung muss mit einem entsprechend größeren *bias* der Parameter und im Extremfall auch mit unzuverlässigen Ergebnissen der Modell-Fit-Indizes gerechnet werden.

5.2 Ausblick

Aufgrund der zentralen Bedeutung der globalen Wahrnehmung der Lehrkraft im Kontext der geteilten und nicht-geteilten Unterrichtswahrnehmung wäre eine Klärung der Zusammenhänge dringend erforderlich. Da aus theoretischer Sicht eine gegenseitige Beeinflussung der Unterrichts- und der globalen Wahrnehmung angenommen werden kann, lassen sich die zugrunde liegenden Prozesse wohl nur im Rahmen einer Längsschnittuntersuchung mit mehreren (mindestens zwei bzw. drei) Messzeitpunkten angemessen untersuchen.

Einen zusätzlich erschwerenden Faktor bei der Untersuchung dieser Zusammenhänge stellen die über Klassen hinweg variierenden Einflüsse verschiedener Unterrichtsmerkmale auf das Globalurteil dar. Wenn man davon ausgeht, dass die Variation auf bestimmte Eigenschaften der Lehrkräfte zurückzuführen ist – was ebenfalls zu überprüfen wäre –, dann müsste die Stichprobe eine hinreichende Zahl an Lehrkräften bzw. Klassen umfassen, um über ausreichende Teststärke sowohl für den Nachweis variierender Regressionsgewichte als auch potentieller *cross-level*-Effekte (also Einflüsse bestimmter Lehrermerkmale auf die Regressionsgewichte auf der Ebene innerhalb von Klassen) zu verfügen.

Idealerweise sollten nur Lehrkräfte einbezogen werden, die die jeweilige Klasse erst übernehmen und dort zuvor noch nie unterrichtet haben, da so der Entwicklungsverlauf des zirkulären Prozesses – also der gegenseitigen Beeinflussung des Globalurteils und der Unterrichtswahrnehmungen – von Beginn an untersucht werden könnte. Hier ließe sich auch die Übertragbarkeit der von Kenny (2004) für die interpersonale Wahrnehmung nachgewiesenen Stabilität von Urteilen auf die Unterrichtswahrnehmung überprüfen.

Bezogen auf die zirkuläre Beeinflussung von Globalurteil und Unterrichtswahrnehmungen wäre zunächst zu klären, ob der Einfluss des Globalurteils auf die nachfolgende Unterrichtswahrnehmung oder der Einfluss der Unterrichtswahrnehmungen auf das spätere Globalurteil größer ist. Wenn das Globalurteil überwiegend ein Resultat verschiedener Unterrichtswahrnehmungen darstellt, so wäre eine Interpretation des Unterrichts aus Schülersicht im Sinne eines G-Faktor-Modells[156] mit spezifischen Unterrichtsmerkmalen angebracht, wobei der G-Faktor als globales Urteil bezüglich der „Unterrichtsqualität“ betrachtet werden könnte. Überwiegt hingegen der Einfluss des Globalurteils auf die Unterrichtswahrnehmung, so wäre die Interpretation des G-Faktors eher schwierig: Hier könnten beispielsweise Persönlichkeitsmerkmale, physische Attraktivität der Lehrkraft und andere unterrichtsferne Merkmale eine Rolle spielen.

Weiterhin ließen sich – bei mindestens drei Messzeitpunkten – hohe Korrelationen der Regressionsgewichte[157] identischer Unterrichtsmerkmale zu verschiedenen Messzeitpunkten auf die globale Wahrnehmung der Lehrkraft (zum jeweils nächsten Messzeitpunkt) als Hinweis auf die Stabilität klassenspezifischer Gewichtungen bewerten, die sich möglicherweise auf Lehrer- oder auf Klassenmerkmale zurückführen lassen. Eine niedrige Korrelation der *slopes* hingegen spräche für eine Variabilität der Gewichtungen verschiedener Unterrichts-

[156] G-Faktor steht hier für „General-Faktor“.
[157] Da ein ähnlich komplexes *random slope*-Modell in der vorliegenden Untersuchung nicht konvergierte, wären hierfür entweder völlig andere Verfahren (z.B. Mischverteilungsmodelle) oder approximierende Lösungen erforderlich (wie z.B. die Erzeugung von *scores* für die jeweiligen *slopes* auf Klassenebene; aufgrund der üblicherweise geringen Reliabilität der *random slopes* ist dieses Verfahren allerdings problematisch).

merkmale bei der Bildung des Globalurteils, die etwa auf einen besonders starken Einfluss von Ereignissen aus der unmittelbaren Vergangenheit hinweisen könnte: Wenn eine Lehrkraft beispielsweise in der letzten Unterrichtsstunde vor der Erhebung in der Klasse besonders schülerorientiert unterrichtet hat, dann könnte dies einen starken Effekt auf das aktuelle Globalurteil ausüben. Bei der nächsten Erhebung stünde dann aber vermutlich ein anderes Unterrichtsmerkmal bei der Bildung des Globalurteils im Vordergrund.

Darüber hinaus wäre der postulierte über Klassen bzw. Lehrkräfte hinweg *konstante* Einfluss des Globalurteils auf die verschiedenen Unterrichtswahrnehmungen zum jeweils nächsten Messzeitpunkt zu überprüfen. Theoretisch wäre ein solcher invarianter Einfluss wohl am plausibelsten. Dennoch sind auch hier Interaktionen zwischen der Klassenzugehörigkeit und dem jeweiligen Regressionsgewicht denkbar. Möglicherweise neigen Schülerinnen und Schüler insbesondere bei den schwach ausgeprägten Unterrichtsmerkmalen zu einer „Überbewertung", wenn sie ihre Lehrkraft besonders gern mögen.

Als weiterer Zugang zur Erforschung des auf die Lehrkraft bezogenen Globalurteils kämen auch – z.B. im Rahmen von Voruntersuchungen – offene Fragen in Betracht, bei denen Schülerinnen und Schüler, die angeben, ihre Lehrkraft besonders zu mögen bzw. nicht zu mögen, dazu aufgefordert werden, die drei wichtigsten Gründe dafür anzugeben. Weiterhin könnte auf diese Weise erfasst werden, welche Eigenschaften eine „ideale" Lehrkraft aus Schülersicht besitzen sollte. Häufig genannte Angaben aus diesen offenen Fragen ließen sich dann in Form von vorgegebenen Antworten in umfangreicheren Studien einsetzen.

Auf der Basis solcher Daten ließe sich dann zunächst klären, ob sich *Klassen* hinsichtlich des „Idealbildes" einer Lehrkraft unterscheiden. Wenn ja, dann wäre dies im Sinne einer *Eigenschaft* von Schulklassen zu interpretieren – die „Rater" wären dann nicht „theoretisch austauschbar" zwischen Klassen –, was einen objektiven Vergleich der Globalurteile, die sich auf konkrete Lehrkräfte beziehen, zumindest einschränken würde. Würden sich hingegen die Angaben zu einer hypothetischen „idealen" Lehrkraft nicht unterscheiden, dann würden eventuelle Unterschiede bezüglich der Bewertungen der konkreten Lehrkraft die Hypothese unterstützen, dass von der Lehrkraft selbst Einflüsse auf die Gewichtungen unterschiedlicher Faktoren bei der Bildung eines globalen Urteils ausgehen.

Einen anderen Zugang zu der Frage, welche Komponenten bei der Erfassung des auf die Lehrkraft bezogenen Globalurteils eine Rolle spielen, bieten Verfahren der kognitiv orientierten Survey-Forschung. So ließen sich etwa mithilfe sogenannter retrospektiver Protokolle (vgl. z.B. Sudman et al., 1996), bei denen Schülerinnen und Schüler im Rahmen von Inter-

views nach dem Inhalt bzw. einer Paraphrasierung des entsprechenden Items (bzw. mehrerer Items bei Erfassung des Globalurteils mithilfe einer Skala) befragt werden könnten, Einflussfaktoren auf die globale Wahrnehmung der Lehrkraft „indirekt“ ergründen.

Zur Beantwortung der Frage nach eventuell vorliegenden Ankereffekten bei der Unterrichtswahrnehmung aus Schülersicht – Schülerinnen und Schüler beurteilen vermutlich den Unterricht der jeweiligen Lehrkraft auf der Basis des Unterrichts anderer, ihnen bekannter Lehrkräfte (vgl. dazu auch Clausen, 2002) – wäre an den Einsatz verschiedener Sequenzen aus unterschiedlichen Unterrichtsvideografien zu denken: Schülerinnen und Schüler könnten im Anschluss an die Vorführung der jeweiligen Sequenz eine Reihe von unterrichtsbezogenen Fragen dazu beantworten. Abschließend würden Schülerinnen und Schüler dann aufgefordert, den Unterricht ihrer eigenen Lehrkraft auf der Basis der zuvor gesehenen Videosequenzen einzuordnen. Idealerweise würde jeweils ein Teil der Klasse als Kontrollgruppe die Fragen beantworten *ohne* zuvor die Unterrichtsvideografien gesehen zu haben. Zusätzlich könnten verschiedene „Video“-Gruppen (mit unterschiedlichen Reihenfolgen der Filmsequenzen) unterschieden werden.

Beim Vorliegen von Ankereffekten bei der Beurteilung des Unterrichts sollten sich die auf die jeweiligen Videosequenzen bezogenen Fragen in ihrem Niveau klassenweise unterscheiden. Reihenfolgeeffekte der Unterrichtsvideografien würden dagegen für – aufgrund unzureichender Kenntnis des möglichen Spektrums (z.B. Unterricht in anderen Bildungsgängen) – „variable“ Verankerungen sprechen, die neuen Erfahrungen angepasst werden. Weiterhin würden – verglichen mit der Kontrollgruppe – höhere Übereinstimmungen der auf den eigenen Unterricht bezogenen Wahrnehmungen in der „Video“-Gruppe erwartet, da dort die Antwortskala präziser verankert ist. Dies könnte auch dazu führen, dass unterschiedliche Konstrukte weniger global beantwortet werden, was sich in niedrigeren Interkorrelationen der Unterrichtswahrnehmungen in der „Video“-Gruppe verglichen mit der Kontrollgruppe zeigen sollte. Darüber hinaus wäre auf Klassenebene zu erwarten, dass sich die Unterrichtswahrnehmungen der Kontroll- und der „Video“-Gruppe deutlich von einer perfekten Korrelation unterscheiden.

Bei einer solchen Untersuchung sollte auch das auf die jeweilige Lehrkraft (auch in den Videosequenzen!) bezogene Globalurteil erfasst werden. Besonders interessant wären die Effekte in der „Video“-Gruppe: Variieren die auf das jeweilige Globalurteil bezogenen Regressionsgewichte eines bestimmten Unterrichtsmerkmals in gleicher Weise über die zu beurteilenden Lehrkräfte hinweg, sind also die *slopes* hoch korreliert, dann wäre die Variation der

Regressionsgewichte auf Eigenschaften der Klassenkomposition und *nicht* auf Eigenschaften der zu beurteilenden Lehrkraft zurückzuführen. Niedrige Korrelationen der *slopes* jeweils identischer Unterrichtskonstrukte (auf unterschiedliche Lehrkräfte bezogen) hingegen würden für den Einfluss bestimmter Lehrermerkmale auf die Gewichtung von Unterrichtsmerkmalen bei der „Konstruktion" eines Globalurteils sprechen.

Um zu einer Einschätzung einer maximal möglichen Übereinstimmung bezüglich der einzusetzenden unterrichtsbezogenen Items zu gelangen – um gewissermaßen die Qualität der Items „an sich" zu bewerten –, könnten zuvor auch Experten (z.B. erfahrene Lehrkräfte) die verschiedenen Unterrichtssequenzen auf dieser Basis einschätzen. Dabei sollten exakt dieselben Vorgaben verwendet werden wie bei den Schülerinnen und Schülern (also keine zusätzlichen Einweisungen, Hinweise etc.). Damit ließe sich zusätzlich für jede Sequenz ein „Unterrichtsprofil" aus Expertensicht erstellen, welches mit den mittleren Einschätzungen von Schülerinnen und Schülern verglichen werden könnte. Auf dieser Basis wären Rückschlüsse auf die – je nach Konstrukt – unterschiedlich ausgeprägten generellen Fähigkeiten von Schülerinnen und Schülern zur Unterrichtsbeurteilung möglich.

Um den Einfluss der Kommunikation von Schülerinnen und Schülern auf die Unterrichtswahrnehmungen genauer untersuchen zu können, wäre eine gleichzeitige Erhebung des Unterrichts aus Schülersicht und der sozialen Netzwerke in Klassen (Kommunikationsaspekte und Globalurteile bezüglich der Mitschüler) von großem Vorteil. Auf der Basis einer solchen Untersuchung ließe sich der Effekt der (unterrichtsbezogenen bzw. allgemeinen) Kommunikation – die möglicherweise durch das ebenfalls zu erfassende, auf die jeweilige Schülerin bzw. den jeweiligen Schüler bezogene Globalurteil moderiert wird – auf die verschiedenen Unterrichtswahrnehmungen mithilfe spezieller mehrebenenanalytischer Verfahren zur Modellierung sozialer Netzwerke[158] (vgl. z.B. Snijders & Bosker, 1999) sehr präzise untersuchen. Dann ließe sich auch untersuchen, ob die unterschiedlichen Interrater-Reliabilitäten bei Mädchen und Jungen auf ein unterschiedliches Ausmaß an (allgemeiner bzw. unterrichtsbezogener) Kommunikation oder auf eine unterschiedliche Gewichtung der Urteile der Mitschüler zurückzuführen ist.

Was den Einfluss von häufig diskutierten Prädiktoren wie Geschlecht, Leistung und Schulnoten auf die Unterrichtswahrnehmungen anbelangt, so wären hier theoretische Modelle erforderlich, die die vermittelnden psychologischen Prozesse dieser „Trägervariablen" umfas-

[158] Diese Verfahren müssten allerdings auf eine Anwendung im Kontext mehrebenenanalytischer KFAs verallgemeinert werden.

sen. In diesem Zusammenhang sind vermutlich globale und fachspezifische Selbstkonzepte, die Einstellung zur Schule sowie motivationale Faktoren (z.B. intrinsische vs. extrinsische Motivation) von zentraler Bedeutung.

Im Kontext geschlechtstypischer Unterschiede der Unterrichtswahrnehmung, die sich möglicherweise auf kommunikative Aspekte bzw. unterschiedliche Gewichtungen der Urteile der Mitschüler zurückführen lässt, sollte auch die Messinvarianz der Faktoren in beiden Gruppen (Jungen und Mädchen) überprüft werden, die in der vorliegenden Arbeit bei verschiedenen Analysen vorausgesetzt wurde. Dazu geben die relativ großen Unterschiede bezüglich der Interrater-Reliabilitäten Anlass. Liegen aber unterschiedliche Messmodelle vor, dann sind Aussagen bezüglich der Niveauunterschiede von Unterrichtswahrnehmungen in beiden Gruppen nicht oder nur bedingt möglich.

Weiterhin wäre auch ein Vergleich der Interkorrelationen der Unterrichtswahrnehmungen innerhalb der Gruppe der Mädchen bzw. der Jungen aufschlussreich. Wenn man annimmt, dass die Kommunikation (zumindest die auf unterrichtliche Aspekte bezogene) innerhalb von Klassen bei Mädchen größer ist als bei Jungen, dann sollte sich dies in höheren Interkorrelationen der Unterrichtsmerkmale in der Gruppe der Mädchen zeigen: Da die aus Sicht der Mitschüler zur Verfügung stehenden Informationen meist wohl nur eine unzureichende Passung mit den erfragten Inhalten aufweisen, sollte sich hier ein stärkerer Effekt im Sinne eines Globalfaktors zeigen. Zusätzlich kann davon ausgegangen werden, dass das Gewicht der Meinungen der Mitschüler mit zunehmendem Umfang der verfügbaren Informationen zunimmt. Da tendenziell Schülerinnen und Schüler, die sich mit Mitschülern häufig austauschen, diese wohl auch als ihnen sehr ähnlich empfinden, sollten die Informationen der Mitschüler ein höheres Gewicht bei der Urteilsbildung erhalten (vgl. dazu Kap. 2.1.3).

Bezogen auf die Frage der Effekte unterschiedlicher Itemformulierungen (Adressatenbezug, Wahrnehmungsperspektive, Inferenzgrad; vgl. Kap. 2.1.1) auf die Unterrichtswahrnehmungen wären weitere Untersuchungen erforderlich, bei denen Schülerinnen und Schüler unterschiedliche Formulierungen *identischer* Iteminhalte erhalten. Auf dieser Basis ließen sich die Methodenanteile, die auf die Formulierung zurückzuführen sind, exakt bestimmen, vorausgesetzt, es liegen keine (oder nur geringfügige) Item-Methoden-Interaktionen vor, die eine Repräsentation der „Methode“ in Form eines (gemeinsamen) Faktors zulassen. Um die Zahl der von individuellen Schülern zu bearbeitenden Items möglichst gering zu halten, wäre die Erfassung in Form eines sogenannten Multi-Matrix-Designs angebracht, bei dem Schülerinnen und Schüler jeweils verschiedene Kombinationen von Itemformulierungen erhalten.

Ein solches Design wird derzeit im Rahmen des Projekts „Diagnostik des Unterrichts als Bestandteil diagnostischer Kompetenz und pädagogischer Professionalität“ eingesetzt, bei dem der Autor dieser Untersuchung Mitantragsteller ist.

Neben den Itemformulierungen sollten auch verschiedene Verankerungen der Antwortkategorien untersucht werden. Es ist denkbar, dass etwa niedrig-inferente Items in Kombination mit möglichst exakten Verankerungen (z.B. „weniger als 1 Mal pro Unterrichtseinheit“, „1-2 Mal pro Unterrichtseinheit“ etc.) – wenngleich aufgrund von Erinnerungseffekten die Interrater-Reliabilitäten nicht besonders hoch sein mögen – auf der Aggregatebene z.B. weniger hoch mit dem Globalurteil bezüglich der Lehrkraft korrelieren und insofern eher das eigentlich zu erfassende Konstrukt messen als hoch-inferente Items mit „vagen“ Antwortalternativen. Selbstverständlich ist eine solche Erfassung nicht bei allen Unterrichtsmerkmalen in gleichem Umfang möglich.

Aus methodischer Sicht sind – zumindest theoretisch – im Zusammenhang mit der Unterrichtswahrnehmung aus Schülersicht eine Reihe aufschlussreicher exploratorischer Analysen möglich. So könnten beispielsweise bei Faktoren mit ebenenspezifischen Messmodellen (also unterschiedlichen Ladungen auf identischen Indikatoren auf den Ebenen) anhand von Mischverteilungsmodellen Gruppen (sogenannte latente Klassen) mit identischen und nicht-identischen Messmodellen ermittelt werden. Die Zugehörigkeit zu einer bestimmten Gruppe ließe sich dann möglicherweise durch verschiedene relevante Klassenmerkmale (wie etwa Bildungsgangzugehörigkeit, Leistungsniveau etc.) erklären.

Neben den Ladungen lassen sich auch für verschiedene (latente) Teilpopulationen unterschiedliche Interzepte und Residualvarianzen der Indikatoren modellieren. So ließen sich eventuell bei auf verschiedene Unterrichtsfächer bezogenen, inhaltlich identischen Indikatoren *latente Klassen* (also hier: Gruppen von Schulklassen) ermitteln, in denen *strict factorial invariance* vorliegt – also konstante itemspezifische Ladungen, Residualvarianzen (auf beiden Ebenen) und Interzepte (auf Klassenebene). Innerhalb der jeweiligen Gruppe von Schulklassen wären dann fachübergreifende Mittelwert- und Varianzvergleiche (diese auch ebenenübergreifend) möglich.

Von großer Bedeutsamkeit wären auch Simulationsstudien zu mehrebenenanalytischen KFAs, die sich mit den Effekten von Verletzungen der Annahme unabhängiger Urteile auf der Ebene innerhalb von Klassen[159] beschäftigen. Da eine solche Varianzheterogenität auf Ebene 1 meist nicht oder nur sehr grob kontrolliert werden kann, wäre eine Einschätzung des somit

[159] Diese Annahme ist in der vorliegenden Untersuchung beispielsweise aufgrund der in der Regel größeren Ähnlichkeit der Urteile der Mädchen, verglichen mit den Jungen, verletzt.

in Kauf genommenen *bias* der verschiedenen Schätzer (Ladungen, Residualvarianzen, Korrelationen etc. auf beiden Ebenen, Interzepte auf Ebene 2) von großem Interesse.

Als technisch mögliche, inhaltlich aber teilweise wohl nur schwer zu interpretierende Analysestrategien wären abschließend noch folgende Modelle zu erwähnen:

1. Faktormodelle mit nonlinearen Faktor-Indikator-Beziehungen: Messinvarianz in linearen Faktormodellen ist möglicherweise auf das Vorliegen nicht-linearer Faktor-Indikator-Beziehungen zurückzuführen. In solchen Fällen würde sich bei nonlinearer Modellierung Messinvarianz zeigen (vgl. Bauer, 2005).
2. Faktormodelle mit heteroskedastischen Residualvarianzen: Mit solchen Modellen ließe sich die Hypothese überprüfen, ob die Residualvarianz der Indikatoren von potentiellen Prädiktoren beeinflusst wird. Beispielsweise könnte hierbei die Fachnote eine Rolle spielen: Schülerinnen und Schüler mit besonders „guten“ bzw. „schlechten“ Noten könnten dazu neigen, alle Items besonders positiv bzw. besonders negativ zu bewerten. Dies entspräche einer Abhängigkeit der Residualvarianz von der Note. Solche Modelle lassen sich in Mplus allerdings nur auf einer Ebene spezifizieren. Das heißt, es ließe sich hier nur das Faktormodell innerhalb von Klassen überprüfen. Dazu lassen sich die auf der Basis von Dagne, Howe, Brown und Muthén (2002) vorgeschlagenen Modelle adaptieren, die von Muthén (2002b) für eine Modellierung in Mplus umformuliert wurden.

6. Literatur

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13* (2), 153-166.

Anderson, N. H. (1996). *A functional theory of cognition*. Mahwah, NJ: Erlbaum.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43* (4), 561-573.

Ansari, A., Jedidi, K. & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika, 67* (1), 49-78.

Babad, E. (1996). How high is "high inference"? Within classroom differences in students' perceptions of classroom interaction. *Journal of Classroom Interaction, 31* (1), 1-9.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60* (3), 361-370.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods, 10* (3), 305-316.

Baumert, J., Kunter, M., Brunner, M., Krauss, S., Blum, W. & Neubrand, M. (2004). Schule und Unterricht. Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 314-349). Münster: Waxmann.

Beck, B., Bundt, S. & Gomolka, J. (in Druck). Ziele und Anlage der Studie. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Beck, B. & Klieme, E. (2006). Einleitung. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Ergebnisse Band 1* (S. 1-8). Weinheim: Beltz Pädagogik.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Hrsg.), *Multilevel theory, research, and methods in organisations* (S. 349-381). San Francisco: Jossey-Bass.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brown, R. D. & Hauenstein, N. M. (2005). Interrater agreement reconsidered: An alternative to the r-sub(wg) indices. *Organizational Research Methods, 8* (2), 165-184.

Burke, M. J., Finkelstein, L. M. & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2* (1), 49-68.

Burnkrant, S. R. (2003). *Interrater agreement of incumbent job specification importance ratings: Rater, occupation, and item effects*. Digital Library and Archives. Information: http://scholar.lib.vt.edu/theses/available/etd-10252003-132129/unrestricted/Burnkrant.pdf.

Byrne, B. M., Shavelson, R. J. & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105* (3), 456-466.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56* (2), 81-105.

Carkenord, D. M. & Stephens, M. G. (1994). Understanding student judgments of teaching effectiveness: A "policy capturing" approach. *Journal of Psychology, 128* (6), 675-682.

Cavanagh, R. F. & Romanoski, J. T. (2006). Rating scale instruments and measurement. *Learning Environments Research, 9* (3), 273-289.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83* (2), 234-246.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?* Münster: Waxmann.

Dagne, G. A., Howe, G. W., Brown, C. H. & Muthén, B. O. (2002). Hierarchical modeling of sequential behavioral data: An empirical Bayesian approach. *Psychological Methods, 7* (2), 262-280.

De Boeck, P. & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory item response models: A generalized linear and nonlinear approach* (S. 3-166). New York: Springer.

den Brok, P., Brekelmans, M. & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the Questionnaire on Teacher Interaction. *Learning Environments Research, 9* (3), 199-213.

DESI-Konsortium. (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI)*. Deutsches Institut für Internationale Pädagogische Forschung. Information: http://www.dipf.de/desi/DESI_Zentrale_Befunde.pdf.

Diehl, J. M. & Arbinger, R. (1990). *Einführung in die Inferenzstatistik*. Frankfurt (Main): Klotz.

Diehl, J. M. & Kohr, H. U. (1994). *Deskriptive Statistik* (11. Aufl.). Frankfurt (Main): Klotz.

Dillon, W. R., Kumar, A. & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin, 101* (1), 126-135.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. *Review of Research in Education, 5*, 163-198.

Dreesmann, H. (1979). *Das Unterrichtsklima als situative Bedingung für kognitive Prozesse und das Leistungsverhalten von Schülern.* Universität Heidelberg.

Dreesmann, H. (1982). *Unterrichtsklima – Wie Schüler den Unterricht wahrnehmen*. Weinheim: Beltz.

Dwight, L. A. (1957). The mean or average deviation is a minimum when taken from the median: A geometrical proof. *Journal of Experimental Education, 26* (1), 93-94.

Eder, F. (2006). Schul- und Klassenklima. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3. Aufl., S. 622-631). Weinheim: Beltz/PVU.

Eichler, W. (2006). Sprachbewusstheit. In B. Beck & E. Klieme (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Konzepte und Messung* (S. 147-157). Weinheim: Beltz Pädagogik.

Eichler, W. & Nold, G. (2006). Sprachbewusstheit. In B. Beck & E. Klieme (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Konzepte und Messung* (S. 63-82). Weinheim: Beltz Pädagogik.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65* (2), 241-261.

Fay, R. E. (Hrsg.). (1989). *Theory and application of replicate weighting for variance calculations* (Proceedings of the Section on Survey Research Methods, S. 212-217). Washington, D.C.: American Statistical Association.

Feeley, T. H. (2002). Comment on Halo effects in rating and evaluation research. *Human Communication Research, 28* (4), 578-586.

Funder, D. C., Kolar, D. C. & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69* (4), 656-672.

Gentry, M., Gable, R. K. & Rizza, M. G. (2002). Students' perceptions of classroom activities: Are there grade-level and gender differences? *Journal of Educational Psychology, 94* (3), 539-544.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52* (11), 1182-1186.

Grilli, L. & Rampichini, C. (2007). Multilevel factor models for ordinal variables. *Structural Equation Modeling, 14* (1), 1-25.

Gruehn, S. (2000). *Unterricht und schulisches Lernen*. Münster: Waxmann.

Guilford, J. P. (1954). *Psychometric methods* (2. Aufl.). New York: McGraw-Hill.

Haladyna, T. & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education, 35* (6), 669-687.

Harsch, C. & Schröder, K. (2006). Textrekonstruktion: C-Test. In B. Beck & E. Klieme (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Konzepte und Messung* (S. 212-225). Weinheim: Beltz Pädagogik.

Hartig, J., Jude, N. & Wagner, W. (in Druck). Methodische Grundlagen der Messung und Vorhersage sprachlicher Kompetenzen. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Harvey, R. J. & Hollander, E. (2004). *Benchmarking r-sub(wg) interrater agreement indices: Let's drop the .70 rule-of-thumb.* Beitrag präsentiert bei Annual Conference of the Society for Industrial and Organisational Psychology, Chicago.

Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marcoulides & R. E. Schumacker (Hrsg.), *New developments and techniques in Structural Equation Modeling* (S. 89-128). Mahwah, NJ: Erlbaum.

Helmke, A. (2003). Unterrichtsevaluation: Verfahren und Instrumente. *schulmanagement, 34* (1), 8-11.

Helmke, A. (2006). *Unterrichtsqualität: Erfassen, Bewerten, Verbessern* (4. Aufl.). Seelze: Kallmeyersche Verlagsbuchhandlung.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2002). *Skalenhandbuch zum Projekt SALVE*. Landau: Universität Landau.

Helmke, A., Hosenfeld, I., Schrader, F.-W. & Wagner, W. (2002). Unterricht aus der Sicht der Beteiligten. In A. Helmke & R. S. Jäger (Hrsg.), *Die Studie MARKUS - Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext. Grundlagen und Perspektiven* (S. 325-411). Landau: Verlag Empirische Pädagogik.

Helmke, A., Schrader, F.-W., Wagner, W., Klieme, E., Nold, G. & Schröder, K. (in Druck-a). Wirksamkeit des Englischunterrichts. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Helmke, A., Schrader, F.-W., Wagner, W., Nold, G. & Schröder, K. (in Druck-b). Selbstkonzept, Motivation und Englischleistung. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Helmke, T., Helmke, A., Schrader, F.-W., Wagner, W., Nold, G. & Schröder, K. (in Druck). Die Videostudie des Englischunterrichts. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53* (2), 221-234.

Hox, J. J. (2002). *Multilevel analysis: techniques and applications*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Hu, L.-T. & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Hrsg.), *Structural Equation Modeling: Concepts, issues, and applications* (S. 76-99). Thousand Oaks, CA: Sage Publications.

Hutchison, D. & Healy, M. (2001). The effect on variance component estimates of ignoring a level in a multilevel model. *Multilevel Modelling Newsletter, 13* (2), 4-5.

James, L. R., Demaree, R. G. & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69* (1), 85-98.

James, L. R., Demaree, R. G. & Wolf, G. (1993). r-sub(wg): An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78* (2), 306-309.

Jedidi, K. & Ansari, A. (2001). Bayesian Structural Equation Models for multilevel data. In G. A. Marcoulides & R. E. Schumacker (Hrsg.), *New developments and techniques in Structural Equation Modeling* (S. 129-157). Mahwah, NJ: Erlbaum.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59* (3), 381-389.

Jude, N., Klieme, E., Eichler, W., Lehmann, R., Nold, G., Schröder, K., Thomé, G. & Willenberg, H. (in Druck). Strukturmodelle sprachlicher Kompetenzen. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review, 8* (3), 265-280.

Kenny, D. A. & Bergman, J. S. (1980). Statistical approaches to the correction of correlational bias. *Psychological Bulletin, 88* (2), 288-295.

Klieme, E., Jude, N., Eichler, W. & Willenberg, H. (in Druck). Deutschunterricht aus der Sicht von Lehrpersonen sowie Schülerinnen und Schülern. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Kline, R. B. (2005). *Principles and practice of Structural Equation Modeling* (2. Aufl.). New York: Guilford.

Komaroff, E. (1997). Effect of simultaneous violations of essential /g=t/-equivalence and uncorrelated error on coefficient /g=a/. *Applied Psychological Measurement, 21* (4), 337-348.

Kreft, I. & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9* (3), 231-251.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tilmann, K.-J. & Weiß, M. (2002). *PISA 2000 : Dokumentation der Erhebungsinstrumente*. Berlin: MPI für Bildungsforschung.

La Du, T. J. & Tanaka, J. S. (1995). Incremental fit index changes for nested Structural Equation Models. *Multivariate Behavioral Research, 30* (3), 289-316.

Lance, C. E., LaPointe, J. A. & Steward, A. M. (1994). A test of the context dependency of three causal models of Halo rater error. *Journal of Applied Psychology, 79* (3), 332-340.

Lindell, M. K. (2001). Assessing and testing interrater agreement on a single target using multi-item rating scales. *Applied Psychological Measurement, 25* (1), 89-99.

Little, T. D. (1997). Mean and Covariance Structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32* (1), 53-76.
Longford, N. T. & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika, 57* (4), 581-597.
Lubbers, M. J. (2003). Group composition and network structure in school classes: a multilevel application of the p* model. *Social Networks, 25* (4), 309-332.
Lubke, G. H., Dolan, C. V., Kelderman, H. & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*, 543-566.
Lubke, G. H. & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10* (1), 21-39.
Lüdtke, O. & Köller, O. (2006). Mehrebenenanalyse. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3. Aufl., S. 469-474). Weinheim: Beltz/PVU.
Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9* (3), 215-230.
Malloy, T. E., Agatstein, F., Yarlas, A. & Albright, L. (1997). Effects of communication, information overlap, and behavioral consistency on consensus in social perception. *Journal of Personality and Social Psychology, 73* (2), 270-280.
Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57* (3), 519-530.
Marsh, H. W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74* (2), 264-279.
Marsh, H. W. & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Hrsg.), *Structural Equation Modeling: Concepts, issues, and applications* (S. 177-198). Thousand Oaks, CA: Sage Publications.
Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52* (11), 1187-1197.
Marsh, H. W. & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92* (1), 202-228.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47* (2), 149-174.
McIntyre, M. D. & James, L. R. (1995). The inconsistency with which raters weight and combine information across targets. *Human Performance, 8* (2), 95-111.
Mehta, P. D. & Neale, M. C. (2005). People are variables too: Multilevel Structural Equations Modeling. *Psychological Methods, 10* (3), 259-284.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58* (4), 525-543.
Meredith, W. & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44* (11 (Suppl. 3)), 69-77.
Minium, E. W. (1970). *Statistical reasoning in psychology and education*. New York: Wiley.
Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29* (2), 133-161.
Murphy, K. R. & Anhalt, R. L. (1992). Is Halo error a property of the rater, ratees, or the specific behaviors observed? *Journal of Applied Psychology, 77* (4), 494-500.
Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continous latent variable indicators. *Psychometrika, 49* (1), 115-132.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28* (4), 338-354.
Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22* (3), 376-398.
Muthén, B. O. (2002a). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29* (1), 81-117.
Muthén, B. O. (2002b). *Modeling heteroscedastic measurement errors*. Mplus Web Note #3. Information: http://www.statmodel.com/download/webnotes/mc3.pdf.
Muthén, B. O. (2004). *Mplus Technical Appendices*. Los Angeles, CA: Muthén & Muthén.
Muthén, B. O. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.
Muthén, B. O. & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.
Muthén, L. K. & Muthén, B. O. (2006). *Mplus User's Guide* (4 Aufl.). Los Angeles, CA: Muthén & Muthén.
O'Brien, R. M. (1985). The relationship between ordinal measures and their underlying values: Why all the disagreement. *Quality and Quantity, 19*, 265-277.
Ostini, R. & Nering, M. L. (2006). *Polytomous Item Response Theory models* (Quantitative applications in the social sciences, Vol. 07-144). Thousand Oaks: Sage.
Parducci, A. (1965). Category judgement: A range-frequency model. *Psychological Review, 72* (6), 407-418.
Peters, C. C. & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: MacGraw-Hill.
PISA-Konsortium Deutschland (Hrsg.). (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.
Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*.
Raudenbush, S., Bryk, A., Cheong, Y. F. & Congdon, R. (2004). *HLM 6: Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International.
Reise, S. P., Widaman, K. F. & Pugh, R. H. (1993). Confirmatory factor analysis and Item Response Theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114* (3), 552-566.
Renaud, R. D. & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education, 46* (8), 929-953.
Rindskopf, D. (1984). Structural Equation Models: Empirical identification, Heywood cases, and related problems. *Sociological Methods & Research, 13* (1), 109-119.
Rolff, H.-G. & von der Gathen, J. (in Druck). Rückmeldungen an Lehrkräfte und Rezeption. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.
Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Aufl.). Bern: Huber.
Saldern, M. v. (1987). *Sozialklima von Schulklassen: Überlegungen und mehrebenenanalytische Untersuchungen zur subjektiven Wahrnehmung von Lernumwelten*. Frankfurt (Main): Lang.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika (Monograph Supplement No. 17)*.
SAS Institute Inc. (2004). *SAS/STAT® 9.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2006). *Base SAS® 9.1.3 Procedures Guide, Second Edition, Volumes 1, 2, 3, and 4*. Cary, NC: SAS Institute Inc.

Satorra, A. (2000). *Scaled and adjusted restricted tests in multi-sample analysis of moment structures* (Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker). London: Kluwer Academic Publishers.

Satow, L. (1999). *Klassenklima und Selbstwirksamkeitsentwicklung: Eine Längsschnittstudie in der Sekundarstufe I*. Freie Universität Berlin. Information: http://www.diss.fu-berlin.de/2000/9/index.html.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Steyer, R. & Eid, M. (2001). *Messen und Testen* (2. Aufl.). Berlin: Springer.

Sudman, S., Bradburn, N. M. & Schwarz, N. (1996). *Thinking about answers: the application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Tuerlinckx, F. & Wang, W.-C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory item response models: A generalized linear and nonlinear approach* (S. 75-110). New York: Springer.

Wagner, W., Helmke, A., Schrader, F.-W., Eichler, W., Thomé, G. & Willenberg, H. (in Druck). Selbstkonzept und Motivation im Fach Deutsch. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.), *Deutsch Englisch Schülerleistungen International (DESI). Leistungsverteilungen und Bedingungsfaktoren*. Weinheim: Beltz.

Warr, P. B. & Knapper, C. (1968). *The perception of people and events*. London: Wiley.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64* (6), 956-972.

West, S. G., Finch, J. F. & Curran, P. J. (1995). Structural Equation Models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Hrsg.), *Structural Equation Modeling: Concepts, issues, and applications* (S. 56-75). Thousand Oaks, CA: Sage Publications.

Westat. (2002). *WesVar® 4.2 User's Guide*. Rockville, MD: Westat.

Whittaker, T. A. & Stapleton, L. M. (2006). The performance of cross-validation indices used to select among competing covariance structure models under multivariate nonnormality conditions. *Multivariate Behavioral Research, 41* (3), 295-335.

Widaman, K. F. & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Hrsg.), *The science of prevention: methodological advances from alcohol and substance abuse research* (S. 281-324). Washington, DC: American Psychological Association.

Winne, P. H. & Marx, R. W. (1982). Students' and teachers' views of thinking processes for classroom learning. *The Elementary School Journal, 82*, 493-518.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continous outcomes.* University of California, Los Angeles.

Verzeichnis der Abbildungen

Abbildung 1: Einflussgrößen der individuellen Wahrnehmung von Unterricht 27

Abbildung 2: Zusammenhang zwischen Interzept und Regressionsgewicht am Beispiel des ersten Items der Skala *Verständlichkeit* im Fach Englisch. Sortierung aufsteigend nach mittlerer Einschätzung der Jungen in Klassen.... 91

Abbildung 3: Einfluss der Itemformulierung: CFA-CTC(M-1)-Modell mit Ich-Bezug als Methodenfaktor und Klassen-Bezug als Referenzmethode 107

Abbildung 4: Einfluss der Itemformulierung: CFA-CTC(M-1)-Modell mit Klassen-Bezug als Methodenfaktor und Ich-Bezug als Referenzmethode..... 108

Abbildung 5: Einfluss der Itemformulierung: CFA-CTUM-Modell..... 109

Abbildung 6: Einfluss der Itemformulierung: CFA-CTCM-Modell mit fachspezifischen Globalfaktoren der Unterrichtswahrnehmung 110

Verzeichnis der Tabellen[160]

Tabelle 1: Adressat des Lehrerverhaltens, Wahrnehmungsperspektive und Inferenzgrad bei Itemformulierungen zur Unterrichtswahrnehmung aus Schülersicht 9

Tabelle 2: Skalen und Items zum Unterricht aus Schülersicht 34

Tabelle 3: Beispiel zur Bestimmung der optimalen Scores für eine vierstufig ordinalskalierte Variable bei zugrunde liegender intervallskalierter, normalverteilter Variable 55

Tabelle 4: Statistische Kennwerte der arbiträren bzw. optimalen *scores* einer vierstufig ordinalskalierten Variable bei zugrunde liegender intervallskalierter, normalverteilter Variable 55

Tabelle 5: Schwellenparameter (*thresholds*) des Zwei-Ebenen-IRT-Modells mit den Faktoren *Thematische Motivierung* und *Strukturiertheit* im Fach Englisch ... 60

Tabelle 6: Simulation I: IRT-Populationsmodell und identisches Analysemodell, jedoch mit intervallskalierten Indikatoren 61

Tabelle 7: Kategorienwahrscheinlichkeiten, Mittelwerte der arbiträren *scores* und Korrelationen der optimalen (r_{ku}) sowie der arbiträren *scores* (r_g) mit den hypothetisch zugrunde liegenden intervallskalierten, normalverteilten Variablen 62

Tabelle 8: Unadjustierte ICC, gemeinsame Varianz der kategorisierten Variable mit arbiträren *scores* und der hypothetischen intervallskalierten Variable (innerhalb), adjustierte Koeffizienten, *bias* in Prozent 63

Tabelle 9: Modellgüte-Indizes für verschiedene KFA-Modelle (Mittelwerte und Standardabweichungen bezüglich der 1000 Replikationen) 65

Tabelle 10: Schiefe und Exzess der auf der Basis des Zwei-Ebenen-IRT-Modells generierten Daten (Basis: Daten aus allen 1000 Replikationen) 65

Tabelle 11: Modellgüte-Indizes für verschiedene KFA-Modelle (Mittelwerte und Standardabweichungen bezüglich der 1000 Replikationen des zweiten Populationsmodells) 67

Tabelle 12: Wald-Test für itemspezifisch identische Ladungen (Basis: Modell B) für den Faktor *Strukturiertheit* auf beiden Ebenen (Erwartungswerte und beobachtete Werte auf der Basis von 1000 Replikationen) 68

Tabelle 13: Verteilungen der Kategorien, Mittelwerte, Standardabweichungen, Anzahl gültiger Werte, Anteil fehlender Werte, Kurtosis und Schiefe der Indikatoren 73

Tabelle 14: Verteilungen der Kategorien der Indikatoren (inkl. Normalverteilungskurven) 75

Tabelle 15: Reliabilität der Skalen bzw. Faktor*scores* zur Unterrichtswahrnehmung aus Schülersicht (Einebenen-Analyse) 76

[160] Die Tabellentitel wurden teilweise gekürzt

Tabelle 16: Verteilungen, Mittelwerte, Streuungen und Anzahl gültiger Werte der kategorialen Prädiktoren ... 77
Tabelle 17: Mittelwerte und Streuungen der testbasierten Leistungsindikatoren ... 78
Tabelle 18: Interrater-Reliabilität: ICCs (metrisch und ordinal) auf Itemebene (in Prozent) ... 79
Tabelle 19: Korrelationen der optimal bzw. arbiträr kategorisierten Variablen mit den (hypothetisch) zugrunde liegenden intervallskalierten, normalverteilten Variablen ... 80
Tabelle 20: Adjustierte metrische ICCs und gemeinsame Varianzanteile der kategorisierten Variablen mit den zugrunde liegenden intervallskalierten, normalverteilten Variablen innerhalb von Klassen ... 81
Tabelle 21: Überprüfung der Messinvarianz der Faktoren (identische itemspezifische unstandardisierte Ladungen) auf beiden Ebenen: Wald-Tests ... 83
Tabelle 22: Prozentuale Abweichung der unstandardisierten Ladungen innerhalb von Klassen vs. Klassenebene ... 84
Tabelle 23: (Adjustierte) standardisierte Ladungen und Indikatorreliabilitäten innerhalb von Klassen und auf Klassenebene ... 86
Tabelle 24: Interrater-Reliabilität bezüglich der messfehlerbereinigten Unterrichtswahrnehmungen auf Itemebene: ICCs der Kommunalitäten ... 87
Tabelle 25: Interrater-Reliabilität bezüglich messfehlerbereinigter Unterrichtswahrnehmungen auf Konstruktebene: Latente ICCs der ebenenbezogen messinvarianten Faktoren ... 88
Tabelle 26: Geschlechtsspezifische Interrater-Übereinstimmung innerhalb von Klassen: Zwei-Ebenen-Modell mit klassenspezifisch variierendem Regressionsgewicht für die Geschlechtszugehörigkeit ... 90
Tabelle 27: Geschlechtsunterschiede bezüglich der Interrater-Reliabilität von Unterrichtswahrnehmungen auf Itemebene: ICCs (metrisch, in Prozent) auf Itemebene nach Geschlecht ... 92
Tabelle 28: Latente Interkorrelationen der Unterrichtsmerkmale innerhalb von Klassen und zwischen Klassen ... 94
Tabelle 29: Manifeste Interkorrelationen der Unterrichtsmerkmale innerhalb von Klassen und zwischen Klassen mit identischen Lehrkräften in den Fächern Englisch und Deutsch ... 97
Tabelle 30: Manifeste Interkorrelationen der Unterrichtsmerkmale innerhalb von Klassen; Vergleich der Gruppen von Schülerinnen und Schülern, die von derselben Lehrkraft bzw. unterschiedlichen Lehrkräften unterrichtet werden ... 98
Tabelle 31: Fachspezifität bzw. Invarianz der Messmodelle der wahrgenommenen Unterrichtsmerkmale innerhalb von Klassen bzw. zwischen Klassen ... 99
Tabelle 32: Fachspezifische prozentuale Abweichung der unstandardisierten Ladungen: innerhalb von Klassen und zwischen Klassen ... 100
Tabelle 33: Fach- und ebenenübergreifende Tests auf Messinvarianz der Ladungen, der Interzepte und der Residualvarianzen ... 101
Tabelle 34: Fachspezifische Varianzen der faktorinvarianten Unterrichtsmerkmale ... 102

Tabelle 35: Standardisierte Ladungen des CFA-CTCM-Modells mit fachspezifischen Globalfaktoren 111
Tabelle 36: Varianzkomponenten der Indikatoren des CFA-CTCM-Modells mit fachspezifischen Globalfaktoren 112
Tabelle 37: Interkorrelationen der Faktoren des CFA-CTCM-Modells mit fachspezifischen Globalfaktoren 113
Tabelle 38: Interkorrelationen der Prädiktoren 116
Tabelle 39: Zusammenhänge zwischen Unterrichtswahrnehmungen und Prädiktoren innerhalb von Klassen 117
Tabelle 40: Zusammenhänge zwischen Unterrichtswahrnehmungen und Prädiktoren auf Klassenebene 118
Tabelle 41: ICCs der Prädiktoren und gemeinsame Varianzanteile der kategorisierten Variablen mit den zugrunde liegenden intervallskalierten, normalverteilten Variablen innerhalb von Klassen 120
Tabelle 42: Adjustierte Zusammenhänge zwischen Unterrichtswahrnehmungen und Prädiktoren innerhalb von Klassen 120
Tabelle 43: Vorhersage der Unterrichtswahrnehmungen innerhalb von Klassen und zwischen Klassen anhand der Prädiktoren: mehrebenenanalytisches Pfadanalysemodell 122
Tabelle 44: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Englisch: Mittelwert und Streuung der Regressionsgewichte, Interzept-Regressionsgewicht-Korrelationen 126
Tabelle 45: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Englisch: Interkorrelationen der Regressionsgewichte 127
Tabelle 46: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Deutsch: Mittelwert und Streuung der Regressionsgewichte, Interzept-Regressionsgewicht-Korrelationen 128
Tabelle 47: Globalurteil in Abhängigkeit von subjektiven Unterrichtswahrnehmungen mit über Klassen hinweg variierenden Regressionsgewichten im Fach Deutsch: Interkorrelationen der Regressionsgewichte 128

Anhang A

Transformation von Gleichung (7) zu Gleichung (8) (s. Kap. 3.4.2)

Setzt man bei dem ersten Summanden in (7), der sich auf die *between*-Kovarianz bezieht, für die jeweilige ICC(2) die zu Beschreibung der Reliabilität in der KTT übliche Formulierung (*true score*-Varianz – τ_x^2 bzw. τ_y^2 – geteilt durch Gesamtvarianz $\sigma_{M(x)}^2$ bzw. $\sigma_{M(y)}^2$) ein, so erhält man:

$$\sqrt{ICC(2)_x ICC(2)_y}\,\rho_{between} = \sqrt{\frac{\tau_x^2}{\sigma_{M(x)}^2} \cdot \frac{\tau_y^2}{\sigma_{M(y)}^2}} \cdot \frac{\mathrm{cov}(x,y)_{between}}{\sqrt{\tau_x^2 \tau_y^2}} =$$

$$\frac{\sqrt{\tau_x^2 \tau_y^2}}{\sqrt{\sigma_{M(x)}^2 \sigma_{M(y)}^2}} \cdot \frac{\mathrm{cov}(x,y)_{between}}{\sqrt{\tau_x^2 \tau_y^2}} = \frac{\sqrt{\tau_x^2 \tau_y^2} \cdot \mathrm{cov}(x,y)_{between}}{\sigma_{M(x)} \sigma_{M(y)} \sqrt{\tau_x^2 \tau_y^2}} = \tag{23}$$

$$\frac{\mathrm{cov}(x,y)_{between}}{\sigma_{M(x)} \sigma_{M(y)}}$$

Es ergibt sich die auf die Standardabweichungen der beiden Mittelwerte standardisierte *between*-Kovarianz. Analog dazu lässt sich der zweite, auf die *within*-Kovarianz bezogene Summand in (7) wie folgt umformulieren:

$$\sqrt{(1-ICC(2)_x)(1-ICC(2)_y)}\,\rho_{within} = \sqrt{(1-\frac{\tau_x^2}{\sigma_{M(x)}^2})(1-\frac{\tau_y^2}{\sigma_{M(y)}^2})} \cdot \frac{\mathrm{cov}(x,y)_{within}}{\sqrt{\sigma_x^2 \sigma_y^2}} =$$

$$\sqrt{\frac{\sigma_{M(x)}^2 - \tau_x^2}{\sigma_{M(x)}^2} \cdot \frac{\sigma_{M(y)}^2 - \tau_y^2}{\sigma_{M(y)}^2}} \cdot \frac{\mathrm{cov}(x,y)_{within}}{\sqrt{\sigma_x^2 \sigma_y^2}} \tag{24}$$

Ersetzt man nun die Varianzen der Mittelwerte in den Zählern der beiden auf die Unreliabilitäten bezogenen Terme durch den Ausdruck im Nenner der ICC(2) (s. (6)) – da die Definition der Reliabilität der KTT und der ICC(2) im Ergebnis identisch sind, und die Zähler der beiden Formulierungen jeweils die *between true score*-Varianz enthalten, müssen auch die Nenner gleich sein – dann ergibt sich:

$$\sqrt{\frac{(\tau_x^2+\frac{\sigma_x^2}{n})-\tau_x^2}{\sigma_{M(x)}^2}\cdot\frac{(\tau_y^2+\frac{\sigma_y^2}{n})-\tau_y^2}{\sigma_{M(y)}^2}}\cdot\frac{\mathrm{cov}(x,y)_{within}}{\sqrt{\sigma_x^2\sigma_y^2}}=$$

$$\sqrt{\frac{\frac{\sigma_x^2}{n}}{\sigma_{M(x)}^2}\cdot\frac{\frac{\sigma_y^2}{n}}{\sigma_{M(y)}^2}}\cdot\frac{\mathrm{cov}(x,y)_{within}}{\sqrt{\sigma_x^2\sigma_y^2}}=\sqrt{\frac{\frac{\sigma_x^2\sigma_y^2}{n^2}}{\sigma_{M(x)}^2\sigma_{M(y)}^2}}\cdot\frac{\mathrm{cov}(x,y)_{within}}{\sqrt{\sigma_x^2\sigma_y^2}}= \tag{25}$$

$$\frac{\frac{\sigma_x\sigma_y}{n}}{\sigma_{M(x)}\sigma_{M(y)}}\cdot\frac{\mathrm{cov}(x,y)_{within}}{\sigma_x\sigma_y}=\frac{\frac{1}{n}\cdot\mathrm{cov}(x,y)_{within}}{\sigma_{M(x)}\sigma_{M(y)}}$$

Setzt man nun die Ergebnisse für die *between*- bzw. *within*-Kovarianz-Komponente in (7) ein, so erhält man:

$$\rho(M(x),M(y))=\frac{\mathrm{cov}(M(x),M(y))}{\sigma_{M(x)}\sigma_{M(y)}}=\frac{\mathrm{cov}(x,y)_{between}}{\sigma_{M(x)}\sigma_{M(y)}}+\frac{\frac{1}{n}\cdot\mathrm{cov}(x,y)_{within}}{\sigma_{M(x)}\sigma_{M(y)}}= \tag{26}$$

$$\frac{\mathrm{cov}(x,y)_{between}+\frac{1}{n}\cdot\mathrm{cov}(x,y)_{within}}{\sigma_{M(x)}\sigma_{M(y)}}$$

Multipliziert man diese Gleichung mit den Standardabweichungen der Mittelwerte so ergibt sich (8).

Anhang B

Mithilfe des unten dargestellten SAS-Macros lassen sich die auf den Überlegungen von O'Brien (1985) basierenden Koeffizienten zum Zusammenhang von kategorisierten Variablen mit einer zugrunde liegenden intervallskalierten Variable auf der Basis empirischer Häufigkeitsverteilungen kategorisierter Variablen ermitteln (vgl. dazu Kap. 3.4.2).

Folgende Informationen müssen dazu an das Macro übermittelt werden:

1. der zu verwendende SAS-Datensatz[161] (datset)
2. die Liste der auszuwertenden Variablen[162] (varlist)
3. optional können auch Gewichte angegeben werden (weight); per Voreinstellung wird auf die Verwendung von Gewichten verzichtet (weight=off)

Das Macro wird wie folgt aufgerufen[163] (inkl. exemplarischen Angaben für Datensatz, Variablen und Gewicht):

```
%groupingerr(datset=desi, varlist=x1 x2 x3 x4, weight=wgt);
RUN;
```

Die Ergebnisse werden im Datensatz *results* (im „*work*"-Verzeichnis) gespeichert. Der Datensatz enthält zeilenweise für jede Variable folgende Angaben (die maximale Anzahl der Kategorien der Items wird im Folgenden mit n[164] bezeichnet):

1. Die Werte der arbiträren *scores*: a1-a[n]
2. Die prozentualen Häufigkeiten der Kategorien: percent1-percent[n]
3. Die Kategorienwahrscheinlichkeiten: p1-p[n]
4. Die kumulierten Kategorienwahrscheinlichkeiten: pcum1-pcum[n]
5. Die Grenzen zwischen jeweils benachbarten Kategorien: g1-g[n-1]
6. Die entsprechenden Ordinaten (Wahrscheinlichkeitsdichten) an den Grenzen: y1-y[n]; Anmerkung: die zur Berechnung der optimalen *scores* erforderlichen Ordinaten y0 und y[n+1] (vgl. (9)) werden jeweils auf Null gesetzt.

---

[161] Das Macro verwendet folgende temporären Datensätze im „*work*"-Verzeichnis: results, freq, ncat, trans1, trans2, trans. Diese Datensätze werden durch das Macro erzeugt bzw. (wenn bereits vorhanden) überschrieben. Der zu analysierende Datensatz muss deshalb anders benannt sein als die Temporärdatensätze.

[162] Die Variablen müssen einzeln – getrennt durch Leerzeichen – aufgeführt werden. Die in SAS üblichen Listenfunktionen wie z.B. x1-x10 oder a--x werden nicht unterstützt.

[163] Zuvor muss das Macro eingelesen werden, indem die Macro-Syntax mit *submit* ausgeführt wird.

[164] Die Zahl der Kategorien der Variablen kann sich natürlich auch unterscheiden. Die Anzahl der Kategorien n bezieht sich dann auf die jeweilige Variable.

7. Der arithmetische Mittelwert sowie die Varianz bzw. Standardabweichung der arbiträren *scores*: M_a bzw. Var_a bzw. s_a
8. Die Varianz bzw. Standardabweichung der optimalen *scores*: Var_o bzw. s_o
9. Die Korrelation der optimalen *scores* mit der zugrunde liegenden intervallskalierten Variable: r_ku
10. Der relative Fehleranteil, den O'Brien als *pure categorization error* bezeichnet: e_ku
11. Die Kovarianz bzw. Korrelation zwischen den optimalen Kategorien*scores* o_i und den arbiträren *scores*: Cov_ka bzw. r_ka
12. Der von O'Brien als *pure transformation error* bezeichnete relative Varianzanteil: e_t
13. Die Korrelation der arbiträren *scores* mit der zugrunde liegenden intervallskalierten Variable: r_g
14. Der Gesamtanteil an Fehlervarianz, die auf die Umwandlung einer intervallskalierten, normalverteilten Variable in eine ordinale Variable mit n Kategorien, den Kategorienwahrscheinlichkeiten p_i und arbiträrem *scoring* zurückführbar ist: e_g

```
%MACRO groupingerr(datset, varlist, weight=off);

  %LET maxv=1;
  %LET var&maxv=%QSCAN(&varlist,&maxv,%STR( ));
  %DO %WHILE(&&var&maxv ne);
    %LET maxv=%eval(&maxv+1);
    %LET var&maxv=%QSCAN(&varlist,&maxv,%STR( ));
  %END;
  %LET maxv=%EVAL(&maxv-1);

  DATA results; SET _NULL_;

  %DO loop=1 %TO &maxv;

    DATA freq; SET _NULL_;
    DATA ncat; SET _NULL_;
    DATA trans1; SET _NULL_;
    DATA trans2; SET _NULL_;
    DATA trans; SET _NULL_;

    PROC FREQ DATA=&datset (WHERE=(&&var&loop NE .)); TABLES &&var&loop / OUT=freq NOPRINT;
    %IF &weight NE off %THEN %DO; WEIGHT &weight; %END;

    PROC MEANS DATA=freq NOPRINT; OUTPUT OUT=ncat N(&&var&loop)=n;

    DATA _NULL_; SET ncat; CALL SYMPUT("ncat", n);

    PROC TRANSPOSE DATA=freq OUT=trans1; VAR &&var&loop;

    PROC TRANSPOSE DATA=freq OUT=trans2; VAR percent;

    DATA trans;
      MERGE
        trans1(RENAME=(%DO i=1 %TO &ncat; COL&i=a&i %END;))
        trans2(DROP=_NAME_ _LABEL_ RENAME=(%DO i=1 %TO &ncat; COL&i=percent&i %END;))
      ;
      %DO i=1 %TO &ncat; p&i=percent&i/100; %END;

      pcum1=p1;
      %DO i=2 %TO &ncat; pcum&i=pcum%EVAL(&i-1)+p&i; %END;

      %DO i=1 %TO &ncat-1; g&i=QUANTILE('NORMAL',pcum&i); %END;

      y0=0;
      %DO i=1 %TO &ncat-1; y&i=PDF('NORMAL', g&i); %END;
```

```
      y%EVAL(&ncat)=0;

      * optimale scores;
      %DO i=1 %TO &ncat; o&i=(y%EVAL(&i-1)-y&i)/p&i; %END;

      * M(arbiträre scores);
      M_a=%DO i=1 %TO &ncat; p&i*a&i+ %END; 0;

      * Varianz(arbiträre scores);
      Var_a=%DO i=1 %TO &ncat; p&i*(a&i-M_a)**2+ %END; 0;
      s_a=SQRT(Var_a);

      * Varianz(optimale scores);
      Var_o=%DO i=1 %TO &ncat; p&i*o&i**2+ %END; 0;
      s_o=SQRT(Var_o);

      * Korrelation(optimale scores, latente intervallskalierte Variable);
      r_ku=s_o;

      * relativer Fehlervarianzanteil aufgrund Kategorisierung (categorization error);
      e_ku=1-r_ku**2;

      * Kovarianz(optimale, arbiträre scores);
      Cov_ka=%DO i=1 %TO &ncat; p&i*a&i*o&i+ %END; 0;

      * Korrelation(optimale, arbiträre scores);
      r_ka=Cov_ka/(s_a*s_o);

      * relativer Fehlervarianzanteil aufgrund Transformation (transformation error);
      e_t=r_ku**2*(1-r_ka**2);

      * Korrelation(arbiträre scores, latente intervallskalierte Variable);
      r_g=r_ku*r_ka;

      * relativer Fehlervarianzanteil gesamt (grouping error);
      e_g=1-r_g**2;

    RUN;

    DATA results; SET results trans;
    RUN;

  %END;
%MEND;
```

Lebenslauf

Persönliche Angaben

Name	Wolfgang Gerold Wagner
Geburtsdatum und -ort	08.06.1971 in Pirmasens

Schule

1977 bis 1990	Grundschule Ruhbank-Erlenbrunn (Pirmasens) Hugo-Ball-Gymnasium (Pirmasens) Abschluss: Abitur

Studium und Beruf

09 / 1992 bis 03 / 2000	Studium der Psychologie an der Universität Koblenz-Landau Abschluss: Diplom-Psychologe
04 / 2000 bis 03 / 2002	Zusatzstudium: Kommunikationspsychologie und Medienpädagogik
11 / 2001 bis 09 / 2005	Wissenschaftlicher Mitarbeiter im Projekt DESI (Deutsch Englisch Schülerleistungen International) bei Prof. Dr. A. Helmke (Universität Koblenz-Landau, Campus Landau, Fachbereich Psychologie, Abteilung Entwicklungspsychologie). Mitarbeit in den Projekten MARKUS (Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext) und PEPP (Projekt: Evaluation der Publikationen des Pädagogischen Zentrums und ihres Ertrages)
Seit 10 / 2005	Wissenschaftlicher Mitarbeiter (Projektkoordinator) im Kooperationsprojekt „Empirische Unterrichtsforschung“ bei Prof. Dr. A. Helmke (Universität Koblenz-Landau, Campus Landau, Fachbereich Psychologie, Abteilung Entwicklungspsychologie)