Clausen, Marten; Reusser, Kurt; Klieme, Eckhard

Unterrichtsqualität auf der Basis hoch-inferenter Unterrichtsbeurteilungen. Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz

Unterrichtswissenschaft 31 (2003) 2, S. 122-141

Unterrichtswissenschaft

Zeitschrift für Lernforschung
31. Jahrgang / 2003 / Heft 2

Thema
Analyse von Unterrichtsvideos

Verantwortliche Herausgeber
Klaus Peter Wild, Alexander Renkl

Marten Clausen, Kurt Reusser, Eckhard Klieme

Unterrichtsqualität auf der Basis hoch-inferenter Unterrichtsbeurteilungen

Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz[1]

Using high-inference ratings to assess quality of instruction A comparison between Germany and the German-speaking part of Switzerland

Der Beitrag fokussiert die Frage, inwieweit sich ein hoch-inferenter Beurteilungsansatz zum systematischen Vergleich von Unterrichtskulturen eignet und inwieweit sich anhand der hoch-inferenten Urteile Unterschiede zwischen Unterrichtsstunden aus der deutschsprachigen Schweiz und Deutschland feststellen lassen. Anhand eines Rating-Fragebogens wurden je 30 videografierte Unterrichtsstunden im Fach Mathematik aus Deutschland (TIMSS-Video-Stichprobe) und der deutschsprachigen Schweiz (TIMSS-Repeat-Video-Stichprobe) durch vier trainierte Raterinnen und Rater beurteilt. Anhand von Generalisierbarkeitsanalysen konnte für die Mehrzahl der beurteilten Unterrichtsaspekte eine gute Messqualität nachgewiesen werden, so dass insgesamt die Brauchbarkeit des hoch-inferenten Beurteilungsansatzes als belegt angesehen wird. Im Vergleich der beiden Unterrichtskulturen zeigen sich für die deutschschweizer Unterrichtsstichprobe höhere Ausprägungen hinsichtlich der Merkmalsbereiche Instruktionseffizienz und Schülerorientierung. Die Befunde werden hinsichtlich ihrer theoretischen und methodologischen Bedeutung für die Unterrichtsforschung diskutiert.

The paper focuses on the question, whether a high-inference rating approach can be used for cross-cultural comparison of instructional environments and whether systematic differences between lessons from Germany and lessons from the German-speaking part of Switzerland can be identified

1 Ohne die im Rahmen der TIMSS-Video-Studie geleisteten Arbeiten wäre das hier dargestellte Projekt nicht möglich gewesen. Wir bedanken uns bei Jürgen Baumert, Direktor am Max-Planck-Institut für Bildungsforschung in Berlin, für die freundliche Genehmigung zur Nutzung der Unterrichtsaufzeichnungen. Darüber hinaus sind wir der gesamten TIMSS-Video-Gruppe am MPI für Bildungsforschung in Berlin und am Leibniz Institut für die Pädagogik der Naturwissenschaften in Kiel für die exzellente Aufbereitung und Dokumentation des deutschen Videomaterials zu Dank verpflichtet.

by means of these ratings. Using a high inference questionnaire, 30 mathematics lessons from Germany (TIMSS-Video sample) and 30 mathematics lessons from the German-speaking part of Switzerland (TIMSS-Repeat-Video sample) were rated by four trained observers. Generalizability analysis yielded a good quality of measurement for most of the assessed aspects of instruction. Therefore, the high inference approach was found appropriate for cultural comparison. In the comparison of the two instructional cultures, the Swiss sample showed higher levels of effectiveness and student orientation. Results are discussed regarding their theoretical and methodological relevance for instructional research.

1. Einführung

Im Jahr 1997 bereitete die Dritte Internationale Mathematik- und Naturwissenschaftsstudie (Third International Mathematics and Science Study (TIMSS); Beaton, Mullis et al., 1996; Baumert, Lehmann et al., 1997) allen Illusionen über den Leistungsstand deutscher Schülerinnen und Schüler im internationalen Vergleich ein Ende. Der TIMSS-Leistungswert von Population 2 (in Deutschland die 8. Klassenstufe) liegt für Deutschland im Fach Mathematik mit 509 Punkten am internationalen Mittelwert (Mittelwert 500, Standardabweichung 100) etwa gleichauf mit den USA (500). Als mögliche Erklärung für das mäßige Abschneiden der deutschen Schülerinnen und Schüler wird häufig auf Defizite in der Unterrichtsgestaltung verwiesen. Insbesondere im Vergleich zur japanischen Unterrichtskultur erweist sich der deutsche Mathematikunterricht gemäß der Befunde aus der TIMSS-Video-Studie als nur in geringem Maße kognitiv aktivierend (Klieme & Bos, 2000). Einschränkend muss jedoch auf die beträchtlichen kulturellen Unterschiede in Lebensumständen sowie gesellschaftlichen Norm- und Wertvorstellungen verwiesen werden, die bei einem solchen direkten Vergleich zu asiatischen Ländern in Betracht gezogen werden müssen.

Die schweizerische Schülerschaft der achten Klassenstufe erreichte im Fach Mathematik 545 Punkte, wobei eine deutliche Variation innerhalb der verschiedenen Landesteile zu Tage trat. Die deutschsprachige Schweiz weist mit 590 Punkten einen Leistungswert auf, der an die asiatischen Spitzenländer heranreicht und nimmt damit im innereuropäischen Vergleich den ersten Rang ein. Die europäische Spitzenposition schweizerischer Schülerinnen und Schüler hinsichtlich ihrer mathematischen Grundbildung wurde durch die PISA-Studie erneut belegt (Deutsches PISA-Konsortium, 2001). Die Teilnahme der Schweiz an der TIMSS-Repeat-Unterrichtsstudie (TIMSS-R-Video) eröffnet die Möglichkeit zu einem Vergleich zwischen zwei Unterrichtskulturen, die sich nur gering in den außerschulischen Lebensumständen unterscheiden. In der hier dargestellten explorativen Studie werden je 30 videografierte Unterrichtsstunden im Fach Mathematik aus Deutschland (TIMSS-Video-Stichprobe) und der deutschsprachigen Schweiz (TIMSS-Repeat-Video-Stichprobe) anhand eines Ratingfragebogens durch

trainierte Raterinnen und Rater beurteilt. In Anlehnung an Forschungsansätze zur Unterrichtsqualität (vgl. u.a. Einsiedler, 2002; Ditton, 2002) wird untersucht, (1) inwieweit sich ein hoch-inferenter Beurteilungsansatz zum systematischen Vergleich von Unterrichtskulturen eignet und (2) inwieweit sich anhand der hoch-inferenten Urteile systematische Unterschiede zwischen den Unterrichtsstunden aus der deutschsprachigen Schweiz und Deutschland feststellen lassen.

1.1 Unterrichtsqualität

Der Begriff *Unterrichtsqualität* hat sich in jüngerer Zeit als Oberbegriff für Forschungsansätze etabliert, in denen die Wirksamkeit verschiedener Unterrichtsaspekte im Hinblick auf multiple Zielkriterien von Unterricht untersucht wird (vgl. u.a. Einsiedler, 1997). Einsiedler (2002) definiert Unterrichtsqualität als „Bündel von Unterrichtsmerkmalen, die sich als ‚Bedingungsseite' (oder Prozessqualität) auf Unterrichts- und Erziehungsziele (‚Kriterienseite' oder Produktqualität) positiv auswirken, wobei die Kriterienseite überwiegend von normativen Festlegungen bestimmt ist und der Zusammenhang von Unterrichtsmerkmalen und Zielerreichung von empirischen Aussagen geleitet ist" (S. 195). Die untersuchten Unterrichtsmerkmale entstammen verschiedenen Forschungstraditionen, der am Prozess-Produkt-Modell der amerikanischen „Teacher Effectiveness"-Forschung orientierten Instruktionspsychologie sowie der eher europäischen Unterrichtsklimaforschung.

Die in den Forschungsansätzen zur Unterrichtsqualität eingesetzten Methoden zur Datengewinnung lassen sich auf einem Kontinuum von „niedrig-inferent" bis „hoch-inferent" (vgl. u.a. Rosenshine, 1970) anordnen: *Niedrig-inferente Beobachtungen* beschränken sich auf Aspekte des spezifischen, beobachtbaren Verhaltens, die einfach und „objektiv" zu kodieren sind. Sie erfordern so gut wie keine schlussfolgernden Kognitionen beim Beobachter. Auch stellen sie verhältnismäßig geringe Anforderungen an ein pädagogisch-didaktisches Verständnis der Instruktionsabläufe. Die Orientierung solcher Beurteilungen an der objektiven Beobachtbarkeit erzeugt in den meisten derartigen Studien eine Fokussierung auf kurze, im Unterricht vergleichsweise häufig auftretende Verhaltensäußerungen. Beispiele sind etwa Kodierungen der Klassenorganisation, von Sozial- und Interaktionsformen sowie die Identifizierung der Anteile aufgabenbezogener versus inhaltsfremder Tätigkeiten im Unterricht. In der Regel werden bei niedrig-inferenten Kodierungen Auftretenshäufigkeit und Zeitdauer eines Merkmals oder einer Tätigkeit direkt quantitativ festgestellt. *Hoch-inferente Beurteilungen* erfordern demgegenüber Schlussfolgerungen bzw. interpretative Prozesse seitens der Beobachter, die über das konkret beobachtbare Verhalten hinausgehen und sich auf abstraktere Sachverhalte bzw. globalere Verhaltensmerkmale beziehen. Viele Schüler- und Lehrerbefragungen anhand von unterrichts- bzw. unterrichtsklimabezogenen Fragebogeninventaren lassen sich in diesem Sinne als hoch-inferente Beurteilungen begreifen, bei denen das Ausmaß eines Merkmals oder einer Verhaltenstendenz relativ

ganzheitlich eingeschätzt wird. Bei der methodologischen Entscheidung zwischen einer hoch- oder einer niedrig-inferenten Vorgehensweise lässt sich von einem Übereinstimmungs-Bedeutsamkeits-Dilemma sprechen (Clausen, 2002): Als die objektivere Variante wird zumeist der niedrig-inferente Zugang aufgefasst, da bei solchen Kodierungen die Person des Beurteilenden in geringerem Ausmaß als Fehlerquelle anzusehen ist. Allerdings zeigen Prozess-Produkt-Studien, dass hoch-inferente Beurteilungen meist höhere Zusammenhänge zu schulischen Erfolgs- und Entwicklungskriterien aufweisen (Rosenshine, 1970; Fraser & Walberg, 1981).

1.2 Die TIMSS-Video-Studie

Als Teilelement von TIMSS wurde von einer Forschungsgruppe der Universität Los Angeles unter der Leitung von James Stigler eine Unterrichtsstudie konzipiert. Diese sogenannte TIMSS-Video-Studie zielte auf den Vergleich der Unterrichtskulturen in Japan, den USA und Deutschland ab. Der in Studien mehrfach beobachtete beträchtliche Leistungsvorsprung und die günstigeren motivationalen Lernhaltungen der asiatischen Schülerinnen und Schüler sind seit längerer Zeit Gegenstand kulturvergleichender Schul- und Unterrichtsforschung (vgl. Stevenson & Stigler, 1992). Im Rahmen des internationalen Ansatzes der TIMSS-Video-Mathematik-Studie wurden anhand eines eher niedrig-inferenten Vorgehens Stundenstrukturen und Abläufe sowie Sozialformen und Unterrichtsgespräche erfasst, dokumentiert und verglichen (Stigler, Gonzales, Kawanaka, Knoll & Serrano, 1999; Stigler & Hiebert, 1999). Es fanden sich deutliche Unterschiede zwischen den drei Unterrichtskulturen, die an dieser Stelle nur ausschnittsweise dargestellt werden können: Innerhalb deutscher und US-amerikanischer Unterrichtsstunden richtet sich der Fokus primär auf den Ewerb mathematischer Fertigkeiten und Routinen, während japanische Stunden eher die Schulung von mathematischem Verständnis zum Ziel haben. Die Entwicklung und Darstellung von alternativen Lösungswegen durch die Schüler ist im japanischen Mathematikunterricht etwa dreimal häufiger zu beobachten als in den Vergleichsländern. In Schülerarbeitsphasen werden in Deutschland und den USA fast durchgehend Routineprozeduren trainiert - in Japan sind solche Unterrichtsphasen zu einem größeren Anteil durch anspruchsvolle Problem- und Denkaufgaben gekennzeichnet (vgl. u.a. Baumert, Lehmann et al., 1997, Klieme & Bos, 2000). Die Befunde aus der TIMSS-Video-Studie haben fachspezifisch und fachübergreifend die Diskussion über Formen und Abläufe des Unterrichtens in mathematisch-naturwissenschaftlichen Fächern angeregt. Dabei werden die beschriebenen Merkmale des japanischen Mathematikunterrichts von deutschen Unterrichtsforschern und Fachdidaktikern als wünschenswerte Zielvorstellung eines modernen, auf kognitive Aktivierung ausgerichteten Mathematikunterrichts angesehen (vgl. u.a. Blum & Neubrand, 1998). Die gefundenen Unterschiede wurden als kulturelle Skripts interpretiert - d.h. als typische länderspezifische Abläufe i.S. von Unterrichtskulturen. Vertiefende Videoanalysen, die von der deutschen TIMSS-Video-Gruppe am Max-

Planck-Institut für Bildungsforschung in Berlin vorgenommen wurden, verankern diese Unterschiede u.a. in typischen Aufgabenstellungen, Diskursmustern und Gruppenarbeitsformen (Klieme, Schümer & Knoll, 2001; BMBF, 2001). Darüber hinaus wurde in Berlin ein erweiterter Zugang zum deutschsprachigen Unterrichtsmaterial entwickelt (Clausen, 2002), der sich stärker als der internationale Zugang an europäischen Forschungstraditionen zu Unterrichtsqualität und Unterrichtsklima orientiert. Zusätzlich zu den niedrig-inferenten internationalen Unterrichtskodierungen wurden für die deutschen Unterrichtsstunden hoch-inferente Beurteilungen durchgeführt. Unter Rückgriff auf den Multitrait-Multimethod Ansatz[2] wurden diese hoch-inferenten Urteile den Urteilen der Schülerinnen und Schüler sowie den Urteilen der Lehrkräfte der videographierten Klassen gegenübergestellt und zu den multiplen Kriterien der schulischen Entwicklung gemäß den Fragebogen und Leistungstests der TIMSS-Schulleistungsstudie in Beziehung gesetzt.

1.3 TIMSS-R, TIMSS-R-Video und die vorliegende Studie

Als Folgestudie zu TIMSS wurde im Jahr 1999 die TIMSS-Repeat-Studie (TIMSS-R) initiiert, an der Deutschland allerdings nicht teilnimmt. Auch diese Schulleistungsstudie wird durch eine Videostudie ergänzt, an der sich Australien, Hong Kong, Luxemburg, die Niederlande, die Schweiz, die Tschechische Republik und die USA beteiligen. Zusätzlich werden die Unterrichtsstunden aus Japan, die bereits im Rahmen der ersten TIMSS-Video-Studie analysiert wurden, erneut in den Vergleich mit einbezogen. Die schweizerische TIMSS-R-Video-Gruppe erweitert den internationalen Zugang, indem sie bei der Untersuchung der gesamtschweizerischen Stichprobe gezielt (1) verschiedene Ebenen (Schüler, Klassen, Schulen etc.), (2) verschiedene methodische Zugänge (niedrig-inferent, hoch-inferent, freie Äußerungen) (3) verschiedene Perspektiven (Innensicht der Akteure, außenstehende Experten) (4) verschiedene Beurteilungsmethoden (TIMSS-R-Video und die vorliegende Studie) und (5) verschiedene schulische Entwicklungskriterien berücksichtigt (Reusser, Pauli & Zollinger, 1998; Reusser 2001).

Die vorliegende Studie ist eingebettet in eine umfassendere Forschungskooperation zwischen der schweizerischen TIMSS-R-Video-Gruppe (Universität Zürich) und den deutschen Unterrichtsforschungsprojekten, die aus der TIMSS-Video-Gruppe hervorgegangen sind (DIPF Frankfurt, Universität Mannheim, MPI für Bildungsforschung Berlin, Projekt „Pythagoras" im Schwerpunktprogramm „Bildungsqualität von Schulen" der DFG). Für die vorliegende Studie bildet der hoch-inferente Zugang anhand von Globalbeurteilungen durch trainierte Raterinnen und Rater die Schnittstelle zum An-

2 Grundgedanke im Multitrait-Multimethod-Ansatz ist es, die Validität von mehreren zu messenden Konstrukten (Traits) zu prüfen, indem man verschiedene Operationalisierungen dieser Konstrukte (Methoden oder Datenquellen) miteinander kontrastiert (vgl. u.a. Campbell & Fiske, 1959).

satz der deutschen TIMSS-Video-Gruppe: Die Studie zielt darauf ab, mögliche Unterschiede zwischen den Stunden aus Deutschland und der deutschsprachigen Schweiz zu explorieren und dabei gleichzeitig die Brauchbarkeit eines hoch-inferenten Beurteilungsansatzes für den Kulturvergleich auf den Prüfstand zu stellen.

2. Fragestellungen und Untersuchungsdesign

Ziel der vorliegenden Studie ist es, anhand eines explorativen Vorgehens zu prüfen, inwieweit sich anhand von hoch-inferenten Beurteilungen *Unterschiede zwischen den in Deutschland und der deutschsprachigen Schweiz beobachteten Unterrichtsstunden* identifizieren lassen. Einschränkend muss hinzugefügt werden, dass die relativ geringe Stichprobengröße von je 30 Stunden pro Land kaum umfassende Aussagen erlaubt. Allerdings lassen sich potentiell anhand dieser Stichproben Tendenzen erkennen, die in Folgestudien besser herausgearbeitet werden können. Unmittelbar verbunden mit der inhaltlichen Frage nach Länderdifferenzen im Vergleich zweier Unterrichtskulturen ist die methodische Frage nach der Objektivität, Reliabilität und Validität der Erfassung derartiger Unterschiede. Neben der Betrachtung der Länderunterschiede wird auch der *Effekt des Herkunftslandes der Beurteilenden* anhand von Varianzanalysen geprüft, um eventuelle Parteilichkeit im Sinne eines kulturellen Bias ausschließen zu können.

Da hoch-inferente Beurteilungen erfordern, dass die Beurteilenden Schlussfolgerungen und Interpretationen über das konkret Beobachtbare hinaus anstellen, sind sie im Vergleich zu niedrig-inferenten Beobachtungen potentiell anfälliger für systematische und unsystematische Beurteilungsfehler. Schon aus diesem Grunde gilt es bei der Anwendung hoch-inferenter Beurteilungsansätze, die *Qualität der Beurteilungsergebnisse* zu prüfen - bei einer Anwendung zum Vergleich zweier Unterrichtskulturen ist eine solche Prüfung unerlässlich. Um die erhobenen Beurteilungen auf ihre Qualität hin zu untersuchen, wird auf den Ansatz der *Generalisierbarkeitstheorie* zurückgegriffen. Ergebnisse von hoch-inferenten Unterrichtsbeurteilungen sind zumeist Skalenwerte für die untersuchten Konstrukte, die wiederum als Mittelwerte über die verschiedenen von den Beurteilenden eingeschätzten Items berechnet werden. Die resultierenden Skalenwerte erlauben keine Berechnung der klassischen Übereinstimmungsmaße (Prozent Übereinstimmung, Cohens Kappa etc.), wie sie bei niedrig-inferenten Beobachtungen vorzunehmen sind. Eine komplexere Reliabilitätsanalyse unter Rückgriff auf die Generalisierbarkeitstheorie erscheint für diese Art Daten angemessener (Cronbach, Gleser, Nanda & Rajaratnam, 1972) und wird u.a. von Helmke (2002) unter Verweis auf Renkl und Helmke (1993) angemahnt. Die methodische Herangehensweise wird im Ergebnisteil erläutert. Ausführliche Darstellungen des Generalisierbarkeitsansatzes finden sich bei Shavelson und Webb (1991) sowie bei Brennan (2001).

Die Stichprobe: In der hier vorgestellten Studie wurden hoch-inferente Unterrichtsbeurteilungen für 60 Unterrichtsstunden vorgenommen, davon 30 aus Deutschland und 30 aus der deutschsprachigen Schweiz. Für jeden Lehrer bzw. jede Klasse war jeweils eine Unterrichtsstunde im Fach Mathematik der 8. Jahrgangsstufe Gegenstand der Beurteilung. Beide Teilstichproben sind jeweils als Zufallsstichproben aus den Stichproben der TIMSS-Video-Studie (Deutschland) und der TIMSS-R-Studie (Schweiz) gezogen worden. Die verschiedenen Schulformen sind entsprechend den Hauptstichproben vertreten. Die für die deutsche Teilstichprobe vorliegenden Leistungswerte weichen nicht bedeutsam von der Hauptstichprobe ab. Natürlich wird eine einzige Unterrichtsstunde der komplexen Verhaltensinteraktion der jeweiligen Lehrer und Schüler kaum gerecht und erzeugt allenfalls ein unscharfes Bild des Unterrichtsgeschehens. Wie die in TIMSS-Video herausgearbeiteten Kulturunterschiede belegen, verdeutlichen derartige Unterrichtsstichproben jedoch in der Zusammenschau von verschiedenen Unterrichten einer Kultur und in deren interkulturellem Vergleich zentrale Charakteristika von Unterrichtskulturen als Lernumwelten. Die im Rahmen des deutschen TIMSS-Video-Ansatzes aufgezeigten Zusammenhänge zur Entwicklung von Leistung und Fachinteresse können im Sinne der prädiktiven Validität als ein weiterer Beleg für die Brauchbarkeit solch eingeschränkter Unterrichtsstichproben gewertet werden.

Das Beurteilungsinstrument: Das in dieser Studie eingesetzte hoch-inferente Beurteilungsinventar basiert auf einem Unterrichtsbeurteilungsinstrument, das bereits im Rahmen der TIMSS-Video-Studie zum Einsatz kam (vgl. u.a. Clausen, 2002). In seiner ursprünglichen Form basiert das Instrument auf dem Schülerfragebogen zum Unterricht, der in den Studien BIJU und TIMSS eingesetzt wurde. Es umfasst einerseits Skalen, die Schülerfragebogen zum Unterrichtsklima (LASSO, v. Saldern & Littig, 1987; ICEQ, Fraser, 1980) entnommen wurden, und andererseits Skalen, die im Rahmen der Studien BIJU und TIMSS zur Erfassung von Unterrichtskonstrukten aus der Lehr-Lern-Forschung am Max-Planck-Institut für Bildungsforschung in Berlin entwickelt wurden (Gruehn, 2000). Hinzu kommen drei Aspekte des „Problemlösenden Mathematikunterrichts" (Fokussierung, Lehrer als Mediator, Einbettung in multiple authentische Kontexte). Diese wurden mit Blick auf die Kernelemente des kognitiv aktivierenden japanischen Unterrichts, wie sie in der TIMSS-Video-Studie herausgearbeitet wurden, von Klieme und Clausen (1999) entwickelt. Speziell für die vorliegende Studie wurde das Instrument orientiert an klassischen Verfahren ergänzt um die Unterrichtsaspekte „Positive Fehlerkultur", „Mathematische Produktivität" und „Aggressionen" (Schüler gegen den Lehrer, Schüler untereinander, Lehrer gegen Schüler).

Das Beurteilungsinstrument umfasst insgesamt 94 Items, die im Rahmen der Auswertung zu Kurzskalen der verschiedenen Unterrichtsaspekte zusammengefasst werden. Die erfassten Merkmale können inhaltlich in vier

Merkmalsbereiche unterteilt werden, die sich faktorenanalytisch replizieren lassen (negative Merkmale sind kursiv gekennzeichnet):

(1) Instruktionseffizienz [Klassenführung, Regelklarheit, Time-on-Task, *Zeitverschwendung, Disziplinprobleme, Aggressionen (Schüler gegen den Lehrer, Schüler untereinander, Lehrer gegen Schüler)*]

(2) Schülerorientierung [Positive Fehlerkultur, Positive Schülerorientierung, Diagnostische Kompetenz (Sozialbereich), Individuelle Lernunterstützung, Individuelle Bezugsnormorientierung, Individualisierung, Multiple authentische Kontexte, *Überforderndes Tempo*]

(3) Kognitive Aktivierung (Mathematische Produktivität, Anspruchsvolles Üben, Lehrer als Mediator, Pacing, Motivierungsfähigkeit, *Repetitives Üben, Sprunghaftigkeit*)

(4) Klarheit und Strukturiertheit [Strukturierungshilfen, Klarheit, Diagnostische Kompetenz (Leistungsbereich, Fokussierung)]

Der Beurteilungsprozess: Um einen eventuellen kulturellen Bias der Beurteilenden (bspw. eine „Heimatparteilichkeit") prüfen zu können, wurden alle 60 Unterrichtsstunden von vier trainierten Ratern und Raterinnen unabhängig voneinander beurteilt, wobei jeweils zwei aus Deutschland und zwei aus der deutschsprachigen Schweiz stammen. Das Training fand in Form eines einwöchigen Workshops statt, innerhalb dessen die Beurteilenden mit dem Beurteilungsinstrument vertraut gemacht wurden und unter direkter Betrachtung der Übereinstimmung anhand von Übungsstunden die Unterrichtsbeurteilung erlernten. Vor der Bewertung wurde der Unterrichtsablauf einer Stunde jeweils vollständig beobachtet. Dabei war es den Beurteilenden freigestellt, einzelne Passagen eingehender zu betrachten. Das hoch-inferente Beurteilungsinventar wurde als Online-Fragebogen realisiert, um eine zentrale Datenerfassung für die räumlich getrennten Rater und Raterinnen zu ermöglichen. Eine solche internetbasierte Lösung erlaubt bereits während des Beurteilungsprozesses einen direkten Zugriff auf alle Beurteilungen.

3. Ergebnisse

Bei der Darstellung der Ergebnisse wird zunächst auf die *Analysen zur Generalisierbarkeit* eingegangen, dann werden *Unterschiede zwischen deutschen und schweizerischen Beurteilenden* betrachtet und anschließend wird auf *Unterschiede zwischen der deutschen und deutschschweizerischen Unterrichtsstichprobe eingegangen.*

3.1 Prüfung der Qualität der Beurteilungen anhand des Generalisierbarkeitsansatzes

Bei der Anwendung der Generalisierbarkeitstheorie (Cronbach, Gleser, Nanda & Rajaratnam, 1972) geht es im Kern darum, eine gemessene Variation auf verschiedene potentielle Varianzquellen (Facetten) zurückzuführen und deren relativen Anteil zu bestimmen. Für die verschiedenen Unterrichtsas-

pekte wird jeweils betrachtet, welcher Anteil der vorliegenden Variation *tatsächliche Unterschiede zwischen Unterrichtsstunden* (die „wahre" Varianz) abbildet, welcher Anteil auf *charakteristische Unterschiede in der Beurteilung der Unterrichtsstunden* durch mehrere Rater (systematische Fehlervarianz, Strenge-, Milde-, Extrem- und Mittetendenzen, vgl. u.a. Mummendey, 1995) zurückzuführen ist und wie viel *unsystematische Variation* (unsystematische Fehlervarianz) in das erhobene Maß mit einfließt. Anhand einer Varianzzerlegung lassen sich Varianzkomponenten für alle drei Anteile bestimmen, anhand deren relativer Bedeutung sich die Güte der Messung abschätzen lässt. In der Terminologie des Generalisierbarkeitsansatzes handelt es sich bei der vorliegenden Analyse um eine 2-Facetten-G-Studie mit den Facetten „Unterricht" und „Rater". Weiter liefert der Generalisierbarkeitsansatz sog. Generalisierbarkeitskoeffizienten, die das Pendant zu den Reliabilitätskoeffizienten der klassischen Testtheorie darstellen: Der absolute G-Koeffizient (auch *index of dependability* genannt, Brennan & Kane, 1977) hat für absolute Entscheidungen, bei denen nicht nur die Rangreihe, sondern auch die absolute Höhe eine Rolle spielt, Gültigkeit. Für Entscheidungen, bei denen nur die Rangreihe der Beurteilungsgegenstände und nicht die absolute Höhe eine Rolle spielt, wird der relative Generalisierbarkeitskoeffizient betrachtet. Auch innerhalb des hoch-inferenten Beurteilungsansatzes ist die Höhe der Inferenz nicht für alle Unterrichtsmerkmale gleich hoch. Eine schlechte Generalisierbarkeit ist speziell für solche Merkmale zu erwarten, die nur selten zu beobachten sind und deren Sichtbarkeit im Sinne einer Verhaltensnähe gering ist (siehe auch Clausen, 2002).

Die Berechnungen zur Generalisierbarkeit wurden mit dem Programm GT (Ysewijn, 1997) vorgenommen. Die Ergebnisse der Generalisierbarkeitsstudie sind in Tabelle 1 und Abbildung 1 wiedergegeben. Allgemein lässt sich feststellen, dass sich auf der Basis von vier Beurteilenden für die Mehrzahl der beurteilten Unterrichtsaspekte gute bis sehr gute Generalisierbarkeitskoeffizienten ergeben. Für die vier verschiedenen Merkmalsbereiche zeigen sich allerdings charakteristische Unterschiede.

Tab. 1: Varianzkomponenten der beobachteten Unterrichtsmerkmale für Video, Rater, Video*Rater und entsprechende absolute und relative Generalisierbarkeitskoeffizienten

	Varianzkomponenten (in Klammern relative Bedeutung in%)[++]			Generalisierbarkeitskoeffizienten	
Instruktionseffizienz	VK Video (%)	VK Rater (%)	VK V*R (%)	absolut	relativ
Klassenführung	.30 (60)	.04 (8)	.16 (32)	.85	.88
Regelklarheit	.41 (71)	.03 (5)	.14 (25)	.91	.92
Time-on-Task	.34 (57)	.07 (11)	.19 (32)	.84	.88
Zeitverschwendung	.33 (66)	.01 (1)	.16 (33)	.89	.89
Disziplinprobleme	.51 (77)	.02 (3)	.13 (20)	.93	.94

Aggressionen (Schüler gegen d. Lehrer)	.45 (74)	.03 (5)	.13 (21)	.92	.93
Aggressionen (Schüler gegen Schüler)	.30 (69)	.01 (2)	.12 (28)	.90	.91
Aggressionen (Lehrer gegen Schüler)	.08 (45)	.02 (9)	.09 (47)	.76	.79
Schülerorientierung	VK Video (%)	VK Rater (%)	VK V*R (%)	absolut	relativ
Positive Fehlerkultur	.18 (30)	.02 (4)	.38 (66)	.64	.65
Schülerorientierung	.16 (29)	.14 (26)	.25 (45)	.62	.72
Diagn. Kompetenz (Sozialbereich)	.11 (27)	.02 (6)	.28 (67)	.60	.62
Individuelle Lernunterstützung	.31 (55)	.03 (5)	.23 (40)	.83	.85
Individuelle Bezugsnormorientierung	.19 (32)	.04 (7)	.35 (60)	.66	.68
Individualisierung	.34 (75)	.00 (1)	.11 (25)	.92	.92
Probleml. U.: Multiple authent. Kontexte	.43 (55)	.03 (4)	.33 (42)	.83	.84
Überforderndes Tempo	.22 (33)	.06 (10)	.38 (57)	.66	.70
Kognitive Aktivierung	VK Video (%)	VK Rater (%)	VK V*R (%)	absolut	relativ
Mathematische Produktivität	.26 (30)	.36 (41)	.25 (28)	.64	.81
Anspruchsvolles Üben	.34 (48)	.04 (5)	.34 (47)	.78	.80
Probleml. U.: Lehrer als Mediator	.30 (51)	.08 (13)	.21 (35)	.81	.85
Pacing	.19 (34)	.02 (3)	.34 (63)	.68	.69
Motivierungsfähigk.	.37 (46)	.06 (7)	.37 (46)	.78	.80
Repetitives Üben	.55 (49)	.06 (5)	.52 (46)	.79	.81
Sprunghaftigkeit	.25 (49)	.00 (0)	.27 (52)	.79	.79
Klarheit und Strukturiertheit	VK Video (%)	VK Rater (%)	VK V*R (%)	absolut	relativ
Strukturierungshilfen	.16 (20)	.28 (36)	.34 (44)	.50	.65
Klarheit	.09 (16)	.24 (41)	.25 (43)	.43	.59
Diagn. Kompetenz (Leistungsbereich)	.05 (10)	.16 (32)	.29 (58)	.31	.41
Probleml. U.: Fokussierung	.27 (42)	.09 (14)	.28 (44)	.74	.79

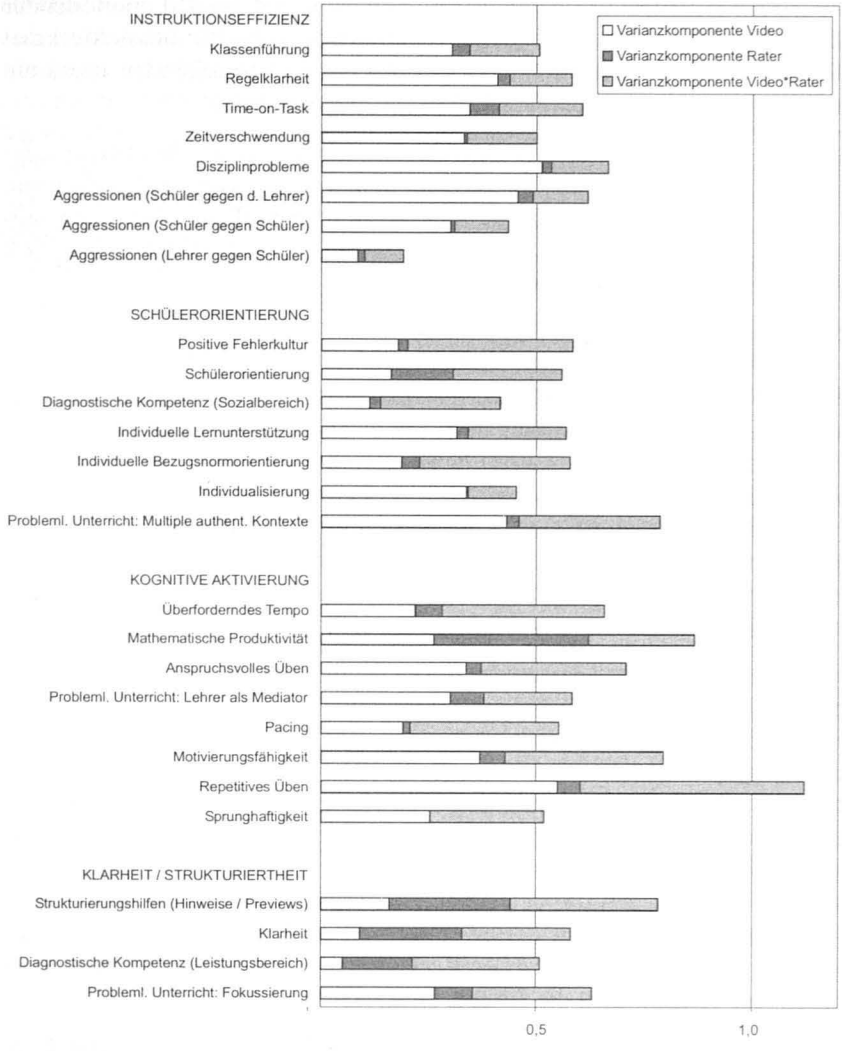
++ Durch Rundung addieren sich die in den Klammern dargestellten prozentualen Anteile nicht in jedem Fall zu 100.

(1) Der Bereich *Instruktionseffizienz* umfasst Unterrichtsmerkmale, die für Beobachter relativ deutlich sichtbar sind. Daher ist es nicht überraschend, dass für diese Unterrichtsaspekte die höchsten Generaliserbarkeitskoeffizienten als Reliabilitätsschätzungen resultieren. Ein Generalisierbarkeitswert größer 0.90 für vier Rater zeigt an, dass aus einer Befragung von nur einem einzelnen Beurteiler für die entsprechenden Unterrichtsaspekte Messwerte mit einer beachtlichen Reliabilität von mehr als 0.70 resultieren würden. Mit jedem weiteren Rater sollte dieser Wert ansteigen. Die Varianzkomponente Video, die im Sinne der klassischen Testtheorie die „wahre" Variation der deutschen und schweizerischen Stunden widerspiegelt, steht durchweg in einem günstigen Verhältnis zu den systematischen Raterunterschieden (Varianzkomponente Rater) und zur unsystematischen Fehlervarianz (Varianzkomponente Video*Rater). Lediglich für das Merkmal „Aggressionen vom Lehrer gegen die Schüler", das Verhaltensweisen wie Bevorzugung bzw. Benachteiligung und zynische Bemerkungen thematisiert, zeigt sich ein ungünstigeres Verhältnis. Aus Abbildung 1, die neben der Betrachtung der relativen Anteile der Varianzkomponenten den Vergleich der Gesamtvarianz für die verschiedenen Beurteilungsmerkmale erlaubt, wird deutlich, dass dieses ungünstigere Verhältnis darauf zurückzuführen ist, dass diese Formen aggressiven Lehrerverhaltens - glücklicherweise - fast nie zu beobachten sind. Wenn ein Merkmal eine sehr geringe „wahre" Variabilität aufweist, so steht diese zwangsläufig in einem ungünstigeren Verhältnis zu systematischen und unsystematischen Fehleranteilen.

(2) Für den Merkmalsbereich *Schülerorientierung* zeigen sich z.T. niedrigere Generalisierbarkeitskoeffizienten zwischen 0.60 und 0.70. Dies dürfte damit zu erklären sein, dass dieser Bereich Unterrichtsaspekte thematisiert, die für Beobachter weniger deutlich sichtbar und damit weniger zugänglich sind und ein höheres Ausmaß an Interpretation mit sich bringen. Ausnahmen stellen die Unterrichtsaspekte „Individualisierung" und „Individuelle Lernunterstützung" dar, die am Lehrerverhalten und am Unterrichtsdiskurs einfacher festzumachen sind.

(3) Der Merkmalsbereich *Kognitive Aktivierung* weist überwiegend gute bis sehr gute Generalisierbarkeitskoeffizienten zwischen 0.70 und 0.80 auf. Auffällig ist das Merkmal „Mathematische Produktivität", das die deutlichsten systematischen Unterschiede zwischen den Beurteilenden zeigt (Varianzkomponente Rater). Für dieses Merkmal unterscheiden sich alle vier Beurteilenden bezüglich ihres mittleren Urteilsniveaus - die Raterurteile sind demnach nicht gut genug „adjustiert". Da sich ein solcher Effekt zwar auf Mittelwerte, nicht aber auf die relative Rangordnung der beurteilten Unterrichtsstunden auswirken kann, findet sich für den relativen Generalisierbarkeitskoeffizienten ein guter Wert von 0.81.

Abb. 1: Beobachtete Varianz der Ratings und Zerlegung in Varianzkomponenten für Video, Rater sowie für die Interaktion Video*Rater



(4) Der vierte Merkmalsbereich *Klarheit und Strukturiertheit* liefert die niedrigsten Generalisierbarkeitskoeffizienten. Für diesen Bereich lässt sich eine systematische Abweichung eines der deutschen Beurteilenden feststellen, der deutsche wie schweizerische Stunden hinsichtlich der Klarheit und Strukturiertheit im Vergleich zu den anderen drei Beurteilenden positiver einschätzt. Dies führt zu einer deutlichen Varianzkomponente Rater und bei dem insgesamt relativ geringen Varianzanteil Video zu einer geringeren Generalisierbarkeit. Problematisch ist insbesondere der Unterrichtsaspekt „Diagnostische Kompetenz des Lehrers im Leistungsbereich", der u.a. darauf abhebt, inwieweit der Lehrer es bemerkt, wenn die Schüler etwas nicht

verstehen. 90% der beobachteten Variabilität dieses Merkmals geht zu Lasten von systematischen und unsystematischen Fehleranteilen. Eine Verbesserung der Messgenauigkeit durch eine Erhöhung der Anzahl beurteilender Personen ist zwar theoretisch möglich - praktisch sollte für dieses Merkmal jedoch eher an einer besseren Verankerung der zu beurteilenden Items am konkret beobachtbaren Lehrerverhalten gearbeitet werden.

Insgesamt lässt sich festhalten, dass die Beurteilungen aus den verschiedenen Bereichen mit wenigen Ausnahmen eine gute Messqualität aufweisen. Der Bereich *Instruktionseffizienz*, der Merkmale umfasst, die einen störungsfreien Unterrichtsablauf mit guter Zeitnutzung charakterisieren, fällt durch besonders gute Messqualität auf.

3.2 Unterschiede zwischen deutschen und schweizerischen Beurteilenden

Für alle 27 Unterrichtsaspekte wurden univariate 2*2 Varianzanalysen mit den Faktoren „Raternationalität" und „Land" als Faktoren durchgeführt. Für diese Analysen ergibt sich bei 60 beurteilten Stunden und vier Beurteilenden eine Stichprobengröße von 240. Eine Parteilichkeit der Beurteilenden würde sich in einer signifikanten Interaktion zwischen den beiden Faktoren ausdrücken. Da für die vorliegenden Beurteilungen die Nullhypothese postuliert wird, d.h. keine bedeutsamen Unterschiede angenommen werden, wird für die 27 Varianzanalysen keine Alpha-Adjustierung vorgenommen.

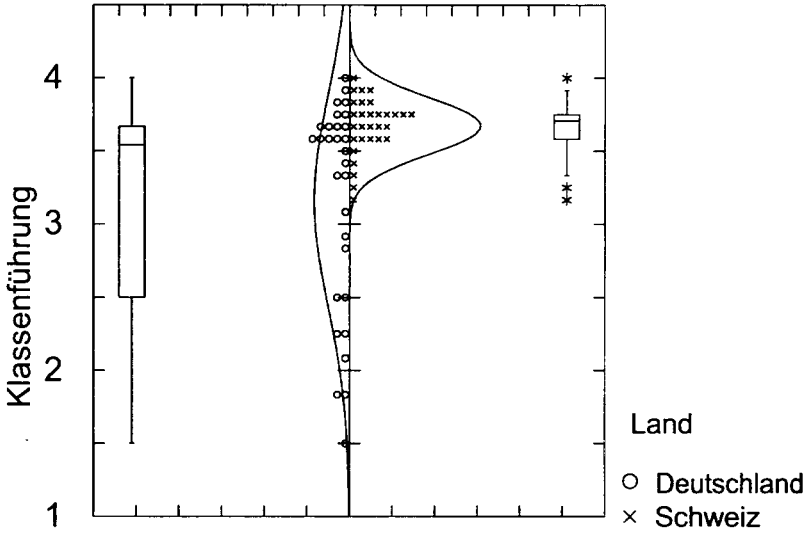
Bei 27 vorgenommenen Varianzanalysen ergibt sich für lediglich fünf Unterrichtsmerkmale eine signifikante Interaktion. Dies sind aus dem Merkmalsbereich *Schülerorientierung* die Aspekte „Positive Schülerorientierung", „Motivierungsfähigkeit", „Individuelle Lernunterstützung" und „Individuelle Bezugsnormorientierung" sowie aus dem Merkmalsbereich *Klarheit und Strukturiertheit* der Aspekt „Strukturierungshilfen". Alle diese Wechselwirkungen sind nicht in Richtung einer „Heimatlandparteilichkeit" gerichtet - eher neigen die deutschen Beurteilenden dazu, die schweizerischen Stunden hinsichtlich dieser Unterrichtsmerkmale noch positiver zu beurteilen als ihre schweizerischen Kolleginnen und Kollegen. Weiter muss hinzugefügt werden, dass die Effekte vergleichsweise schwach sind.

3.3 Unterschiede zwischen der deutschen und deutschschweizerischen Unterrichtsstichprobe

Im Folgenden wird betrachtet, in welcher Hinsicht sich für die beiden untersuchten Unterrichtsstichproben im Kulturvergleich Differenzen ergeben. Analysiert werden für diesen Vergleich jeweils die über die vier Beurteilenden aggregierten Skalenwerte für die betrachteten Unterrichtsmerkmale. Auch wenn die unten aufgeführten Unterschiede jeweils inferenzstatistisch abgesichert sind, soll hervorgehoben werden, dass diese Befunde auf der Basis einer eingeschränkten Stichprobe einer weiteren empirischen Absicherung bedürfen.

(1) Im Merkmalsbereich *Instruktionseffizienz* finden sich deutliche Unterschiede: Die Stunden aus der deutschsprachigen Schweiz weisen eine durchgehend gute Klassenführung mit störungsfreien Unterrichtsabläufen auf. Innerhalb dieser Stunden gibt es hinsichtlich der entsprechenden Merkmale auf hohem Niveau kaum Variation. Im Gegensatz dazu finden sich für die deutsche Stichprobe deutlich ungünstigere Ergebnisse. Etwa ein Viertel der deutschen Stunden liegt in einem Bereich, in dem nicht von einem reibungslosen Unterrichtsablauf gesprochen werden kann. Die in Abbildung 2 gegenübergestellten Verteilungen verdeutlichen die Unterschiede am Beispiel des Unterrichtsaspekts „Klassenführung". Für die Mehrzahl der Unterrichtsaspekte aus diesem Merkmalsbereich finden sich ähnliche Ergebnisse, die an dieser Stelle nicht ausführlich dargestellt werden können. So weisen die schweizerischen Stunden ein höheres Ausmaß an „Regelklarheit" und genutzter Unterrichtszeit („Time-on-Task") auf, während für die deutschen Stunden höhere Werte bei „Disziplinproblemen" und „Zeitverschwendung" resultieren.

Abb. 2: Verteilungen des Unterrichtsmerkmals Klassenführung in deutschen Klassen und Klassen aus der deutschsprachigen Schweiz



(2) Auch innerhalb des Merkmalsbereichs *Schülerorientierung* lassen sich auf verschiedenen Dimensionen deutliche Differenzen zwischen den beiden Stichproben erkennen. Abbildung 3 zeigt die Befunde für das Unterrichtsmerkmal „Individualisierung". Auch wenn allgemein für diese Merkmale nur selten sehr hohe Werte vergeben wurden, so lässt sich doch feststellen, dass schweizerische Stunden einen deutlich höheren Anteil individualisierender Unterrichtsformen aufweisen. In Deutschland sind derartige Unterrichtsformen nur selten nachzuweisen. Weitere, etwas schwächere Unterschiede zeigen sich hinsichtlich der „Positiven Schülerorientierung" und

des positiven Umgangs mit Fehlern („Positive Fehlerkultur"). Wiederum zeichnen sich hier die schweizerischen Stunden durch ein höheres Niveau aus.

Abb. 3: Verteilungen des Unterrichtsmerkmals Individualisierung in deutschen Klassen und Klassen aus der deutschsprachigen Schweiz

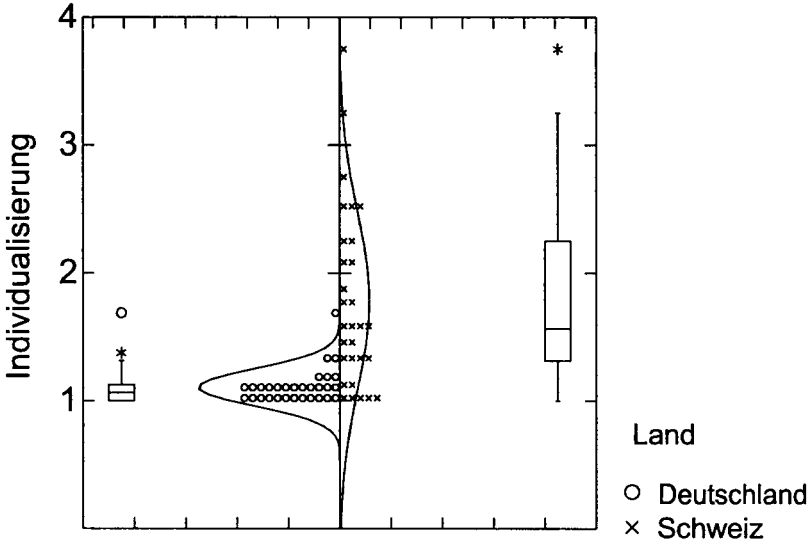
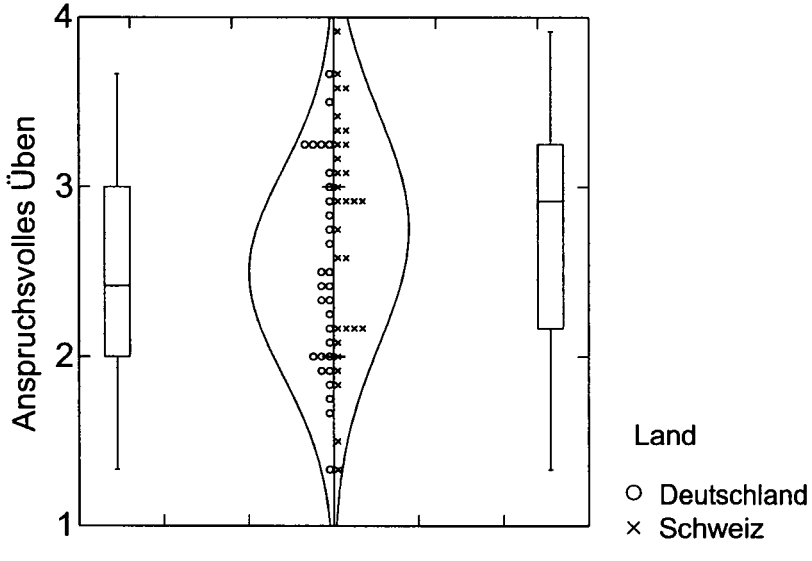


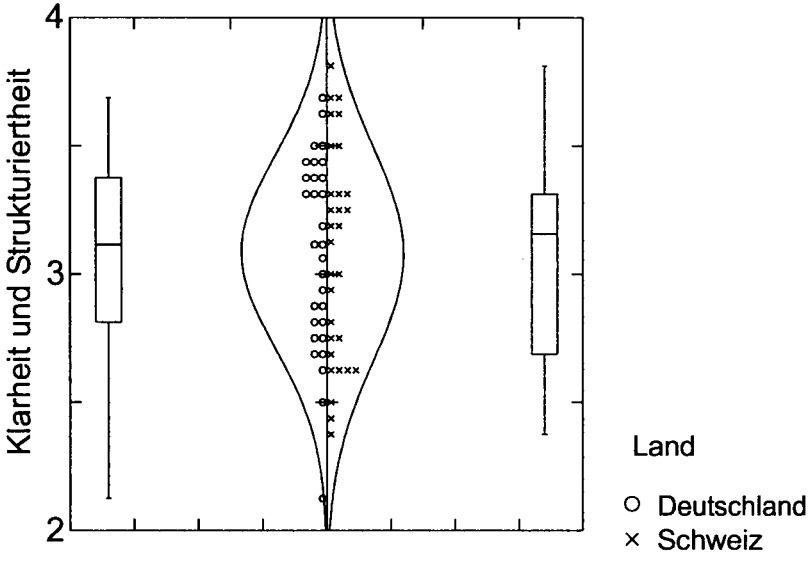
Abb. 4: Verteilungen des Unterrichtsmerkmals Anspruchsvolles Üben in deutschen Klassen und Klassen aus der deutschsprachigen Schweiz



(3) Im Merkmalsbereich *Kognitive Aktivierung* lassen sich nur wenige bedeutsame Unterschiede feststellen. In Abbildung 4 sind beispielhaft die Verteilungen für das Merkmal „Anspruchsvolles Üben" wiedergegeben. Auch wenn die schweizerischen Stunden im Mittel leicht höher liegen, zeigen die beiden Verteilungen eine breite Überlappung. Für die übrigen in diesen Bereich gruppierten Unterrichtsaspekte zeigen sich kaum bedeutsame Unterschiede.

(4) Auch für den Bereich *Klarheit und Strukturiertheit* finden sich allenfalls geringe Unterschiede zwischen den Stichproben. Wie Abbildung 5 entnommen werden kann, liegen deutsche wie schweizerische Stunden auf ähnlich hohem Niveau.

Abb. 5: Verteilungen des Unterrichtsmerkmals Klarheit und Strukturiertheit in deutschen Klassen und Klassen aus der deutschsprachigen Schweiz



4. Diskussion

Im Vergleich der beiden in der vorliegenden Studie betrachteten Unterrichtsstichproben aus Deutschland und der Deutschschweiz lassen sich charakteristische Unterschiede nachweisen - höhere *Instruktionseffizienz* und *Schülerorientierung* in der deutschsprachigen Schweiz, auf der anderen Seite allenfalls geringe Unterschiede in Bezug auf *Kognitive Aktivierung* und *Klarheit/Strukturiertheit.* Aus deutscher Sicht wird deutlich, dass Effizienz und Schülerorientierung - auch im Rahmen eines vergleichbaren kulturellen Kontextes - in wesentlich stärkerem Ausmaß realisiert werden können, als es in Deutschland der Fall zu sein scheint. Während innerhalb der deutschschweizerischen Stichprobe ein reibungsloser Unterrichtsablauf mit guter Zeitnutzung als gesichert angesehen werden kann, wird in etwa einem Vier-

tel der Unterrichtsstunden aus Deutschland darum gerungen, dass ein Unterricht in geordneter Form stattfindet. Die Befunde für die Stichprobe aus der deutschsprachigen Schweiz zeigen, dass Instruktionseffizienz nicht zwangsläufig stark variieren muss. Die sehr positiven Ausprägungen auf den entsprechenden Dimensionen bei gleichzeitiger geringer Variationsbreite unterstreichen dies eindrucksvoll. Das hohe Ausmaß an Schülerorientierung in der schweizerischen Stichprobe geht einher mit einer gewissen reformpädagogischen Tradition in der öffentlichen Schule der Deutschschweiz und einer stetigen und nachhaltigen Entwicklung der Lehrpraxis und Lehrerausbildung hin zu einer Kultur erweiterter Lehr- und Lernformen in den letzten zehn Jahren (ELF; vgl. u.a. Stebler & Reusser, 2000; Reusser, Pauli, Grob, Waldis, Hugener & Krammer, 2001). Die gleichzeitige hohe Instruktionseffizienz zeigt, dass eine hohe Schülerorientierung mit adaptiv-individualisierenden Unterrichtsformen nicht zwangsläufig Kosten auf der Effizienzseite nach sich ziehen muss.

In den Ergebnissen der Generalisierbarkeitsstudie wird deutlich, dass für die verschiedenen in dieser Studie theoretisch und faktorenanalytisch unterschiedenen Merkmalsbereiche (1) *Instruktionseffizienz*, (2) *Schülerorientierung*, (3) *Kognitive Aktivierung* und (4) *Klarheit und Strukturiertheit* unterschiedlich hohe Generalisierbarkeitskoeffizienten resultieren. Die relativen Schwächen des hoch-inferenten Beurteilungsansatzes liegen in den Bereichen *Klarheit/Strukturiertheit* und *Schülerorientierung*. Hier besteht Verbesserungspotential einerseits hinsichtlich der konzeptuellen Schärfung der betrachteten Konstrukte und andererseits hinsichtlich der Items und Skalen, anhand derer diese Konstrukte eingeschätzt werden. Gemäß den klassischen Befunden der Prozess-Produktforschung zu den Zusammenhängen von Instruktionsmerkmalen und schulischer Leistungsentwicklung ist *Instruktionseffizienz* jener Merkmalsbereich, dessen Wirksamkeit im Hinblick auf das Leistungskriterium durch empirische Studien am besten belegt ist (vgl. Rosenshine, 1970; Rosenshine & Furst, 1971; Rosenshine, 1979; Brophy & Good, 1986; Creemers, 1994; Wang, Haertel & Walberg, 1990). Die in der vorgestellten Studie belegte sehr gute Messqualität der Beurteilungen im Bereich Instruktionseffizienz lässt diesen Hauptbefund der Prozess-Produktforschung in anderem Licht erscheinen: Die Tatsache, dass die Varianz mit Bezug auf das Merkmalsbündel *Instruktionseffizienz* (effectiveness-Varianz) am sichtbaren Verhalten besser zu verankern und damit für Videorater gut zu beurteilen ist, führt zu einem höheren Anteil wahrer Varianz in den entsprechenden Messungen. Das heißt, analog zur klassischen Testtheorie sollte der Anteil „wahrer" Variation die Obergrenze der Korrelation einer Variablen mit einem Kriterium darstellen. Die jeweils geringeren Zusammenhänge, die in der Unterrichtsqualitätsforschung üblicherweise für die *übrigen* Merkmalsbereiche resultieren, sind daher möglicherweise darauf zurückzuführen, dass in diesen Bereichen die Zusammenhänge durch jeweils höhere systematische und unsystematische Messfehleranteile über-

lagert und somit unterdrückt werden. Insgesamt lässt sich festhalten, dass die Brauchbarkeit des hoch-inferenten Beurteilungsansatzes in den vorliegenden Analysen belegt werden konnte. Eine solche Brauchbarkeit kann jedoch nicht pauschal angenommen werden - sie muss mit jeder Studie erneut unter Beweis gestellt werden.

Die vorliegenden Ergebnisse haben uns ermutigt, die hoch-inferenten Beurteilungen auszuweiten. Zum einen wird der Vergleich zwischen deutschen und deutschschweizerischen Unterrichtsstunden auf die jeweiligen Gesamtstichproben der TIMSS-Video-Studie und der TIMSS-R-Video-Studie erweitert, die wiederum jeweils von zwei deutschen und zwei deutschschweizerischen Beurteilenden betrachtet werden. Weiter werden in der Schweiz auch die Unterrichtsstichproben aus den französisch- und italienischsprachigen Landesteilen eingeschätzt. Mit der Erweiterung der Stichproben wird ein zuverlässigere Prüfung der gefundenen Länderdifferenzen und eine Prüfung von Zusammenhängen zwischen dem Instruktionsverhalten und den schulischen Entwicklungskriterien der Leistung und Lernmotivation möglich. Eine weitere Validierungsmöglichkeit stellen die niedrig-inferenten internationalen Basiskodierungen der TIMSS-R-Video-Studie dar, die in Kürze für die betrachteten Stunden vorliegen werden.

Die relative Bedeutung der verschiedenen Unterrichtsqualitätsdimensionen für inter- und intranationale Unterschiede bedarf der weiteren Untersuchung. Zur Erklärung des Leistungsunterschieds zwischen der Schweiz und Deutschland leisten die vorgestellten Befunde allerdings einen interessanten Beitrag: Soweit Unterrichtsmerkmale für diesen Unterschied verantwortlich zu machen sind, handelt es sich wohl eher um „klassische" Faktoren der Unterrichtsqualität wie Instruktionseffizienz und Schülerorientierung, nicht jedoch um stärker fachliche und didaktische Merkmale wie Strukturiertheit und kognitive Aktivierung. Möglicherweise spiegeln sich in der Instruktionseffizienz und der Schülerorientierung auch soziokulturelle Bedingungen - soziale Regeln, Wertorientierungen, Leistungsorientierungen der Schüler, Ansehen und Status des Lehrerberufs u.a.m. - als Zeichen einer gesamtgesellschaftlichen Wertschätzung von Bildung.

Literatur

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O. & Neubrand, J. (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde.* Opladen: Leske + Budrich.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L. & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics & Science Study.* Boston, MA: Boston College, Center for the Study of Testing, Evaluation Policy.

Blum, W. & Neubrand, M. (Hrsg.) (1998). *TIMSS und der Mathematikunterricht.* Hannover: Schroedel.

BMBF (Hrsg.) (2001). *TIMSS - Impulse für Schule und Unterricht.* Bonn: Bundesministerium für Bildung und Forschung.
Brennan, R. L. (2001). *Generalizability Theory.* New York, NY: Springer.
Brennan, R. L. & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 13,* 119-135.
Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328-377). New York, NY: Macmillan Publishing Co.
Clausen, M. (2002). *Qualität von Unterricht - Eine Frage der Perspektive?* Münster: Waxmann.
Campbell, D. T. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.
Creemers, B. P. M. (1994). *The effective classroom.* London: Cassell.
Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York, NY: John Wiley.
Deutsches PISA-Konsortium (Hrsg.) (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske + Budrich.
Ditton, H. (2002). Unterrichtsqualität. *Unterrichtswissenschaft, 3,* 197-212.
Einsiedler, W. (1997). Unterrichtsqualität und Leistungsentwicklung: Literaturüberblick. In F.E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 225-240). Weinheim: PVU.
Einsiedler, W. (2002). Das Konzept „Unterrichtsqualität". *Unterrichtswissenschaft, 3,* 194-196.
Fraser, B. (1980). *Criterion validity of an individualized classroom environment questionnaire.* Sydney: McQuaire University.
Fraser, B. & Walberg, H. J. (1981). Psychosocial learning environment in science classrooms: A review of research. *Studies in Science Evaluation, 8,* 67-92.
Gruehn, S. (2000). *Unterricht und schulisches Lernen.* Münster: Waxmann.
Helmke, A. (2002). Kommentar: Unterrichtsqualität und Unterrichtsklima: Perspektiven und Sackgassen. *Unterrichtswissenschaft, 3,* 261-277.
Klieme, E. & Bos, W. (2000). Mathematikleistung und mathematischer Unterricht in Deutschland und Japan: Triangulation quantitativer und qualitativer Forschungsansätze im Rahmen der TIMS-Studie. *Zeitschrift für Erziehungswissenschaft, 3,* 359-379.
Klieme, E. & Clausen, M. (1999). Identifying facets of problem solving in Mathematics instruction. *Vortrag auf der Jahrestagung der American Educational Research Association 1999 in Montreal* (ERIC EDRS: ED432608).
Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In BMBF (Hrsg.), *TIMSS - Impulse für Schule und Unterricht* (S. 43-57). Bonn: Bundesministerium für Bildung und Forschung.
Mummendey, H. D. (1995). *Die Fragebogenmethode.* Göttingen: Hogrefe.
Renkl, A. & Helmke, A. (1993). Prinzip, Nutzen und Grenzen der Generalisierungstheorie. *Empirische Pädagogik, 7(1),* 63-85.
Reusser, K. (2001). Bridging instruction to learning - Where we come from and where we need to go. A research strategy and its implementation in a cross-cultural video survey in Switzerland. *Invited address held at the 9th EARLI Conference in Fribourg/Switzerland.*

Reusser, K., Pauli, C., Grob, U., Waldis, M., Hugener, I. & Krammer, K. (2001). Integrating insider's (participant's) and outsider's (researcher's) perspectives on teaching and learning: The case of adaptive instruction. *Paper presented at the 9th EARLI Conference in Fribourg/Switzerland.*

Reusser, K., Pauli, C. & Zollinger, A. (1998). Mathematiklernen in verschiedenen Unterrichtskulturen - Eine Videostudie im Anschluss an TIMSS. *Beiträge zur Lehrerbildung, 16*, 427-438.

Rosenshine, B. (1970). Evaluation of instruction. *Review of Educational Research, 40*, 279-300.

Rosenshine, B. (1979). Content, time and direct instruction. In P. L. Peterson & H. J. Walberg (Hrsg.), *Research on teaching* (pp. 28-56). Berkeley, CA: McCutchan.

Rosenshine, B. & Furst, N. (1971). Research on teacher performance criteria. In B. O. Smith (Ed.), *Research in teacher education: A symposium* (pp. 37-72). Englewood Cliffs, NJ: Prentice-Hall.

Saldern, M. v. & Littig, K. E. (1987). *Landauer Skalen zum Sozialklima.* Weinheim: Beltz.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Stebler, R. & Reusser, K. (2000). Progressive, classical, or balanced - A look at mathematical learning environments in Swiss-German lower-secondary schools. *Zentralblatt für Didaktik der Mathematik (ZDM), 32*(1), 1-10.

Stevenson, H. W. & Stigler, J. W. (1992). *The learning gap.* New York, NY: Summit Books.

Stigler, J., Gonzales, P., Kawanaka, T., Knoll, S. & Serrano, A. (1999) *The TIMSS videotape classroom study: Methods and findings from an exploratory reserach project on eigth grade mathematics instruction in Germany, Japan, and the United States.* Washington, D.C.: National Center for Educational Statistics (www.ed.gov/NCES).

Stigler, J. W. & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York: The Free Press.

Wang, M. C., Haertel, G. D. & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *Journal of Educational Research, 84*, 30-43.

Ysewijn, P. (1997). *GT - Programm für Generalisierbarkeitsstudien.* Neuchatel: Institut de recherche et de documentation pédagogique (http://www.irdp.ch/methodo/generali.htm).

Anschrift des Autors

Dr. Marten Clausen,
Universität Mannheim, Lehrstuhl für Erziehungswissenschaft II,
Kaiserring 14-16, 68131 Mannheim,
E-Mail: marten.clausen@phil.uni-mannheim.de