Lotz, Miriam; Gabriel, Katrin; Lipowsky, Frank

Niedrig und hoch inferente Verfahren der Unterrichtsbeobachtung. Analysen zu deren gegenseitiger Validierung

Zeitschrift für Pädagogik 59 (2013) 3, S. 357-380

D 7484

ZEITSCHRIFT FÜR PÄDAGOGIK

Heft 3 Mai/Juni 2013

■ *Thementeil*

Quantitative und qualitative Unterrichtsforschung – Gemeinsamkeiten und Differenzen

■ *Allgemeiner Teil*

Das Publikationsaufkommen der Zeitschrift für Pädagogik im deutsch-englischen Vergleich

Schulen im Umgang mit Schulinspektion und deren Ergebnissen

BELTZ JUVENTA

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 59 – Heft 3
Mai/Juni 2013

Inhaltsverzeichnis

Thementeil: Quantitative und qualitative Unterrichtsforschung – Gemeinsamkeiten und Differenzen

Allgemeiner Teil

Besprechung

Dokumentation

Table of Contents

Beilagenhinweis: Dieser Ausgabe der Z.f.Päd. liegt ein Prospekt des Beltz Verlags, Weinheim und des Juventa Verlags, Weinheim, bei.

Miriam Lotz/Katrin Gabriel/Frank Lipowsky

Niedrig und hoch inferente Verfahren der Unterrichtsbeobachtung

Analysen zu deren gegenseitiger Validierung

Zusammenfassung: Für die systematische Unterrichtsbeobachtung werden häufig hoch inferente Schätzverfahren zur Beurteilung der Unterrichtsqualität eingesetzt. Deren Validität wird aber vielfach in Frage gestellt, da sie ein hohes Ausmaß an Schlussfolgerungen aufseiten der Beobachter erfordern. Das Ziel der vorliegenden Studie bestand daher darin, die Güte hoch inferenter Beobachtungssysteme im Rahmen der PERLE-Studie exemplarisch zu überprüfen, indem Zusammenhänge zu niedrig inferent erfassten Beobachtungsdaten analysiert wurden. Anhand der Beispiele „Einsatz von Lob" und „Störungsfreiheit" kann gezeigt werden, dass grundsätzlich Zusammenhänge zwischen den auf unterschiedliche Art erfassten Daten bestehen. Allerdings fällt bei hoch inferenten Ratings die Zuordnung zu den einzelnen Ratingstufen nicht immer eindeutig aus. Zudem deutet sich an, dass die Höhe des Zusammenhangs zwischen niedrig und hoch inferent erfassten Beobachtungsdaten von der Art der Definition der hoch inferent erfassten Merkmale mit bedingt wird.[1]

Schlagworte: Unterrichtsbeobachtung, Unterrichtsqualität, Videoanalyse, niedrig inferente Kodierungen, hoch inferente Ratings

1. Einleitung

Unterricht ist ein soziales Geschehen, in dem es vorrangig um Lernen und um die damit zusammenhängende affektiv-motivationale Entwicklung der Lernenden geht. Vor allem in der quantitativen, aber auch in Teilen der qualitativen Unterrichtsforschung rücken Fragen der Beeinflussbarkeit dieser Zielvariablen in den Blickpunkt. Hier bieten Videostudien die Möglichkeit, unterrichtliche Prozesse nicht nur beobachtbar zu machen und verschiedene Unterrichtspraktiken zu beschreiben, sondern auch zur Erklärung von Bildungswirkungen beizutragen (Pauli & Reusser, 2006). Eine wesentliche methodische Voraussetzung solcher Wirkungsanalysen ist die reliable und valide Erfassung der Ausprägung relevanter Unterrichtsmerkmale in den beobachteten Klassen. Quantitative Erklärungsmodelle setzen voraus, dass das Unterrichtsgeschehen auf wichtige, unterrichtstheoretisch und lehr-lern-theoretisch bedeutsame Aspekte hin verdichtet werden kann, die empirisch in unterschiedlicher Ausprägung – im Sinne von Häufigkeit, Intensität, Breite des Geltungsbereichs – auftreten. Die Überprüfung, inwiefern verschiedene

1 Wir danken den Gutachtern und Herausgebern für die wertvollen Anregungen zu unserem Beitrag.

Methoden der Unterrichtsbeobachtung zu inhaltlich ähnlichen Ergebnissen kommen, sich also wechselseitig validieren können, steht im Fokus dieses Beitrags.

Das komplexe unterrichtliche Geschehen konstituiert sich auf mehreren Ebenen, die miteinander in Beziehung stehen und sich wechselseitig beeinflussen. Zum einen setzt es sich aus einer Vielzahl einzelner Interaktionen der Lehrperson[2] mit der Klasse als Ganzes, aber auch zwischen der Lehrperson und einzelnen Schülern sowie der Schüler untereinander zusammen. Zum anderen zeichnet sich Unterricht durch übergeordnete Strukturen und Qualitätsmerkmale aus, die über diese einzelnen Interaktionen hinausgehen.

Je nach interessierender Fragestellung wird entschieden, welche Methoden der Unterrichtsbeobachtung eingesetzt werden: Geht es eher um die Beschreibung einzelner Interaktionen und unterrichtlicher Abläufe sowie um die genaue Charakterisierung von spezifischen Prozessen, die für Beteiligte und Beobachter direkt wahrnehmbar sind, also eher um die „Sichtstruktur" des Unterrichts im Sinne von Pauli und Reusser (2006), so kommen primär niedrig inferente Beobachtungsverfahren zum Einsatz, bei denen einzelne Ereignisse oder kurze Zeitabschnitte vorgegebenen Kategorien zugeordnet werden. Soll hingegen der Unterricht in seiner pädagogischen Tiefenstruktur eingeschätzt werden, werden meist hoch inferente Verfahren verwendet, wie beispielsweise Ratings, die auf einer umfassenden Einschätzung des Unterrichtsgeschehens durch trainierte Beobachter beruhen. Diese Vorgehensweise wird in der deutschsprachigen Forschung vor allem angewandt, um Basisdimensionen der Unterrichtsqualität (Klieme & Rakoczy, 2008) messbar zu machen: Classroom Management (Kounin, 1976), Lernunterstützung (Rakoczy, 2008) sowie kognitive Aktivierung und fachdidaktische Strukturierung (Reusser & Pauli, im vorliegenden Heft).

In diesem Beitrag wird analysiert, inwiefern sich Beobachtungsdaten, die mittels hoch inferenter Ratings gewonnen wurden, durch niedrig inferent erhobene Daten validieren lassen. Dazu werden ausgewählte Beobachtungsdaten aus der Videostudie im Fach Deutsch des PERLE-Projekts (Lipowsky, Faust & Greb, 2009) verwendet.

2. Verschiedene Methoden der quantitativen Unterrichtsbeobachtung – Chancen und Grenzen

Bei der systematischen, quantitativen Unterrichtsbeobachtung lassen sich niedrig bis hoch inferente Verfahren unterscheiden, die jeweils spezifische Vor- und Nachteile mit sich bringen. Die Inferenz meint den Grad an Schlussfolgerungen, der bei der Beobachtung erforderlich ist, wobei hoch und niedrig inferente Verfahren jeweils die beiden äußeren Pole eines Kontinuums darstellen (z.B. Rosenshine, 1970). Bevor niedrig und hoch inferente Verfahren in den beiden folgenden Abschnitten gegenübergestellt werden, verdeutlicht Tabelle 1 die grundlegenden Unterscheidungsaspekte (z.B. Clausen, Reusser & Klieme, 2003; Hugener, 2006; Lipowsky & Rakoczy, 2006; Seidel, 2003).

2 In diesem Beitrag werden zur besseren Lesbarkeit die Begriffe „Schüler" und „Lehrperson" verwendet. Selbstverständlich sind damit immer beide Geschlechter gemeint.

	Niedrig inferente Verfahren	Hoch inferente Verfahren
Bezeichnung	Kodierung/Kategoriensystem	Rating/Schätzverfahren
Art der Datengewinnung	Erfassen der Häufigkeit und Dauer leicht beobachtbarer Unterrichtsereignisse	Schätzverfahren zum Erfassen der Ausprägung eines Merkmals auf einer vorab definierten Skala
Ziel	Beschreibung der Unterrichtsgestaltung	Bewertung der Unterrichtsqualität
Analyseeinheit	in der Regel kurze Abschnitte, z.B. 10-Sekunden-Intervalle oder kurze Ereignisse	in der Regel längere Unterrichtssequenzen oder ganze Unterrichtsstunden
Grad der Interpretation	Verfahren orientieren sich fast ausschließlich an direkt beobachtbarem Verhalten; geringe Spielräume für die Beobachter	Verfahren orientieren sich nur teilweise an direkt beobachtbarem Verhalten; Schlussfolgerungen der Beobachter nötig
Beispiele	Kodierung der Sozialformen	Einschätzung des Unterrichtsklimas

Tab. 1: Niedrig und hoch inferente Beobachtungssysteme

2.1 Niedrig inferente Verfahren

Bei niedrig inferenten Verfahren wird der Grad der Schlussfolgerungen durch den Beobachter möglichst gering gehalten. Die Unterrichtsbeobachtung erfolgt auf der Basis disjunkter Kategoriensysteme, deren einzelne Kategorien durch genaue Definitionen und Ankerbeispiele möglichst präzise unterscheidbar sind. Ein klassisches niedrig inferentes Verfahren stellt beispielsweise die Kodierung der Sozialformen dar. Die einzelnen Sozialformen wie Einzelarbeit, Partnerarbeit oder Gruppenarbeit lassen sich anhand eindeutiger Regeln abgrenzen. Der Kodierer kann diese Regeln anwenden, ohne viele verschiedene Informationen integrieren und bewerten zu müssen. Meist werden niedrig inferente Verfahren mit Zeit- oder Ereignisstichprobenplänen kombiniert, d.h. es werden entweder vorab bestimmte Zeitintervalle (z.B. 10 Sekunden) festgelegt und für jedes dieser Intervalle wird eine Kategorie vergeben (= Zeitstichprobenplan/Time-Sampling), oder ein bedeutsames Ereignis wird definiert und dessen Auftreten sekundengenau identifiziert und kategorisiert (= Ereignisstichprobenplan/Event-Sampling). Als Ergebnis niedrig inferenter Kodierungen erhält man genaue Informationen über die Unterrichtsgestaltung, über die Häufigkeit und Zeitanteile einzelner Phasen oder Ereignisse sowie die zeitliche Aufeinanderfolge und Sequenzierung des Unterrichts. Zusammenfassend liegt der Vorteil niedrig inferenter Kodierungen – durch die detaillierte Aufgliederung einzelner Unterrichtsereignisse und deren Zuordnung zu Kategorien (Seidel, 2005) – vor allem in ihrer Genauigkeit und der hohen Reliabilität der Ergebnisse. Aussagen zur Unterrichtsqualität müssen hierbei aber theoretisch begründet werden und sollten durch Zusammenhänge mit hoch inferenten Ratings, die Qualitätsmerkmale di-

rekt operationalisieren, empirisch belegt werden. Ein eher pragmatischer Anlass für die Suche nach methodischen Alternativen ist der sehr hohe Zeitaufwand, den detaillierte niedrig inferente Kodierungen erfordern.

2.2 Hoch inferente Verfahren

Im Gegensatz zu niedrig inferenten Kodierungen geht es bei hoch inferenten Ratings im Kern darum, die Qualität von Unterricht bzw. die Qualität einzelner Unterrichtsabschnitte oder -ereignisse auf einer abgestuften Skala einzuschätzen und dadurch komplexe, miteinander interagierende Merkmale des Unterrichts zu bewerten (Hugener, Rakoczy, Pauli & Reusser, 2006; Petko, Waldis, Pauli & Reusser, 2003). Die Analyseeinheit ist in der Regel die gesamte Unterrichtsstunde, da einzelne Anhaltspunkte häufig über die gesamte Stunde verteilt sind und sich nicht an spezifischen Ereignissen festmachen lassen (Waldis, Grob, Pauli & Reusser, 2010).

Ein hoch inferentes Rating besteht nach Rakoczy und Pauli (2006) zumeist aus einem Gesamturteil, das mehrere verhaltensnah definierte Indikatoren berücksichtigt. In das Gesamturteil können die Häufigkeit, die Intensität und/oder die Verteilung dieser Verhaltensweisen innerhalb der Klasse einhergehen. Mit Letzterem ist gemeint, ob alle Schüler das entsprechende Verhalten zeigen oder nur einzelne. Die Mischung dieser Einzelfacetten im Gesamturteil des Raters verlangt ein hohes Ausmaß an Interpretationen und Schlussfolgerungen, die im Einzelnen nicht standardisierbar sind. Intersubjektiv übereinstimmende Urteile kommen daher erst nach einem Training zustande, bei dem die Rater ihre Interpretationen ausführlich diskutieren – ähnlich dem Vorgehen in der qualitativen Unterrichtsforschung.

Auch im Ratingsystem von Rakoczy und Pauli (2006) kommen allerdings vereinzelt Merkmale vor, die nicht über ein vielschichtiges „Gesamturteil", sondern mittels spezifischer Alternativen einzuschätzen sind: Beim Merkmal „sachlich-konstruktive Rückmeldung" wird im Kern eine Häufigkeitseinschätzung verlangt, von „keine Rückmeldung ist sachlich-konstruktiv" bis „alle Rückmeldungen sind sachlich-konstruktiv" (S. 218). Beim Merkmal „Disziplinprobleme/Unterrichtsstörungen" müssen die Häufigkeit von störendem Schülerverhalten einerseits und die Stärke der Störungen andererseits im Urteil der Beobachter kombiniert werden; die Abstufungen reichen hier von „keine Störungen" bis „hohe Disziplinprobleme" (S. 230-231).

Gerade der interpretative Gehalt könnte aber der Grund für die Aussagekraft hoch inferenter Urteile sein, wenn es beispielsweise um die Erklärung der Leistungs- und Motivationsentwicklung im und durch den Unterricht geht. Bezüglich der Validität haben hoch inferente Urteile gegenüber niedrig inferenten Beobachtungen den Vorteil, dass sie oft einen stärkeren Bezug zur Theorie haben, etwa wenn Basisdimensionen der Unterrichtsqualität unmittelbar durch die Ratings operationalisiert werden (vgl. die eingangs erwähnten Beispiele). Nach Clausen (2002; vgl. auch Reyer, 2004) haben daher hoch inferente Urteile eine stärkere Aussagekraft als niedrig inferente Beobachtungen. Auch Clausen, Reusser und Klieme (2003) konnten zeigen, dass die Zusammenhänge

niedrig inferenter Verfahren mit dem Lernerfolg der Schüler schwächer ausfallen als bei hoch inferenten Einschätzungen.

Obwohl Beobachtungen durch Externe bislang als beste Methode gelten, um Aspekte der Unterrichtsqualität zu erfassen (Clausen, 2002; Praetorius, Lenske & Helmke, 2012), und Rater – da sie nicht direkt im Geschehen involviert sind – im Vergleich zu Schülern oder den Lehrpersonen weniger subjektiven Verzerrungen bei der Beurteilung unterliegen (Hoyt, 2000; Petko et al., 2003), haben hoch inferente Ratings ihrerseits Nachteile, die an dieser Stelle diskutiert werden müssen. In diesem Zusammenhang wird bei der Einschätzung von Unterrichtsqualität zunehmend die Rolle der Rater (Rater Bias) kritisiert (zsf. Pietsch & Tosana, 2008; Praetorius et al., 2012). Insbesondere der Milde-Strenge-Fehler kann als konstanter Messfehler bzw. robuste Determinante des Urteilverhaltens immer wieder bestätigt werden (zsf. Eckes, 2004). Auch weitere mögliche Verzerrungseffekte wie beispielsweise der Primacy- oder der Halo-Effekt können eine Rolle spielen (Waldis et al., 2010). Aus diesem Grund werden Unterschiede in den Einschätzungen der Rater als systematische Fehler angesehen und untersucht. Ein weiteres Problem wird deutlich, wenn man sich die Frage stellt, welche Unterrichtssituationen der einzelne Rater für sein Urteil heranzieht bzw. wie der Rater die Abstufungen einer Skala interpretiert. Zwar sollte nach einem intensiven Training ein gemeinsames theoretisches Verständnis über den zu beurteilenden Unterrichtsaspekt existieren, jedoch kann trotz übereinstimmender Beurteilung nicht nachvollzogen werden, ob die Rater einzelne Situationen unterschiedlich gewichten bzw. unterschiedliche Situationen für ihr Urteil berücksichtigen.

3. Fragestellungen

Üblicherweise wird die Güte von Beobachtungsdaten lediglich über die Berechnung von Beobachterübereinstimmungen oder -reliabilitäten geprüft. Dabei ist eine ausreichend hohe Objektivität bzw. Reliabilität der Daten zwar eine wichtige Grundlage, sie sichert aber noch nicht die Validität der Beobachtungsdaten. Bislang existieren nur wenige Studien, die die Zuverlässigkeit dieser Ratereinschätzungen überprüfen (z.B. Pietsch & Tosana, 2008; Praetorius et al., 2012).

Ausgehend von der Kritik an hoch inferenten Ratings soll daher anhand von Daten aus dem Projekt PERLE überprüft werden, ob sich niedrig und hoch inferent gewonnene Beobachtungsdaten wechselseitig validieren lassen. Dazu wird für zwei ausgewählte hoch inferent erfasste Merkmale untersucht, ob sich Zusammenhänge zu niedrig inferent erfassten Aspekten des Verhaltens der Lehrpersonen ergeben.

Gezielt wurden für diesen Vergleich Unterrichtsmerkmale ausgewählt, die – ähnlich wie die oben erwähnten Merkmale des Ratingsystems von Rakoczy und Pauli (2006) – mittels spezifischer Abstufungen definiert sind und sich dadurch unterschiedlich gut für einen Vergleich mit niedrig inferenten Kodierungen eignen. Die ausgewählten Merkmale stehen zudem unterrichtstheoretisch für zwei verschiedene Basisdimensionen der Unterrichtsqualität: „Einsatz von Lob“ kann als wichtige Komponente einer unterstüt-

zenden Unterrichts- und Beziehungsgestaltung angesehen werden. Bei dem ausgewählten Merkmal geht es speziell um die unterstützende Qualität von Lehrerreaktionen auf Äußerungen und Verhaltensweisen von Schülern. Die Bewertung erfolgt durch ein vierstufiges Rating, bei dem Häufigkeiten einzuschätzen sind: 1 = „Die Lehrperson lobt nicht häufig bis gar nicht"; 2 = „Die Lehrperson lobt weniger häufig"; 3 = „Die Lehrperson lobt häufig"; 4 = „Die Lehrperson lobt sehr häufig". Es liegt auf der Hand, dass eine solche Häufigkeitseinschätzung vergleichsweise gut mit niedrig inferenten Kodierungen übereinstimmen kann, wenn es gelingt, die einzelnen Ereignisse (d.h. lobende, also affektiv positiv getönte Reaktionen der Lehrperson) zu identifizieren.

Das Merkmal „Störungsfreiheit" steht für erfolgreiches Classroom Management. Die vier vorgegebenen Abstufungen variieren von 1 = „häufige massive Unterrichtsstörungen bzw. große Disziplinprobleme" über 2 = „einzelne massive Störungen oder häufigere kleinere Störungen" und 3 = „mehrere kleinere Störungen, die den Unterrichtsablauf aber nicht beeinflussen" bis 4 = „weitgehend störungsfreier Unterricht" (vgl. Tabelle 2). Dieses Merkmal erfordert von den Ratern, sowohl die Häufigkeit als auch Stärke sowie die vermuteten Konsequenzen von Unterrichtsstörungen einzuschätzen und diese Informationen in einem Gesamturteil zu kombinieren und zu integrieren. Das Urteil ist also facettenreicher als im ersten Beispiel, sodass schwächere Zusammenhänge mit niedrig inferenten Kodierungen zu erwarten sind.

Die so vorgenommenen hoch inferenten Einschätzungen der beiden Unterrichtsmerkmale „Einsatz von Lob" durch die Lehrperson und „Störungsfreiheit" werden im Folgenden mit niedrig inferenten Messungen verglichen, die auf Kategorisierungen aller Reaktionen der Lehrkraft auf Verhaltensweisen und Äußerungen der Schüler beruhen. Zum einen werden Zusammenhänge zwischen dem hoch inferent erfassten „Einsatz von Lob" und der niedrig inferent erfassten Häufigkeit affektiv positiver Lehrerreaktionen betrachtet; zum anderen wird analysiert, ob die hoch inferent eingeschätzte „Störungsfreiheit" des Unterrichts mit der niedrig inferent erfassten Häufigkeit von Lehrerreaktionen auf Störverhalten zusammenhängt. Für beide Merkmale stellt sich zudem die Frage, wie gut die einzelnen Stufen voneinander abgrenzbar sind und wie die Rater die einzelnen Ratingstufen interpretieren. Was verstehen die Rater beispielsweise unter „sehr häufigem Lob"?

Zusammenfassend werden folgende Fragen beantwortet:

1. Wie hoch sind die Zusammenhänge zwischen ausgewählten niedrig und hoch inferent erfassten Beobachtungsdaten?

1.1 Wie hoch sind die Zusammenhänge zwischen dem „Einsatz von Lob" (hoch inferent) und der Häufigkeit affektiv positiver Lehrerreaktionen (niedrig inferent)?

1.2 Wie hoch sind die Zusammenhänge zwischen der Einschätzung der „Störungsfreiheit" (hoch inferent) und der Häufigkeit von Reaktionen auf Unterrichtsstörungen (niedrig inferent)?

1.3 Sind die Zusammenhänge bei „Einsatz von Lob" – wie oben vermutet – enger als bei dem Merkmal „Störungsfreiheit"?

2. Wie werden die Ratingstufen der beiden hoch inferenten Ratingdimensionen „Einsatz von Lob“ und „Störungsfreiheit“ von den Beobachtern interpretiert? Sind die einzelnen Stufen gut voneinander abgrenzbar?

4. Methodisches Vorgehen

Bevor auf die niedrig und hoch inferenten Beobachtungssysteme eingegangen wird, welche im Beitrag miteinander in Beziehung gesetzt werden, wird zunächst die Datengrundlage, die Videostudie Deutsch der PERLE-Studie, beschrieben. Abschließend wird dargestellt, welche Analysen zur Beantwortung der im vorherigen Abschnitt genannten Fragestellungen angewendet wurden.

4.1 Datengrundlage: Die Videostudie Deutsch der PERLE-Studie

Die Datengrundlage bildet die Videostudie Deutsch des längsschnittlich angelegten Forschungsprojekts PERLE zur Persönlichkeits- und Lernentwicklung von Grundschulkindern (Lipowsky et al., 2009). Für die Videostudie wurden 50 Lehrpersonen gebeten, eine circa 90-minütige Unterrichtseinheit zu planen und darin ein bestimmtes Bilderbuch einzuführen, die Schüler einen Brief aus der Perspektive der Hauptfigur des Bilderbuchs schreiben zu lassen sowie eine Leseübung durchzuführen (vgl. Lotz et al., 2011). Nach der Transkription der Unterrichtsstunden fand zunächst eine Basiskodierung statt, innerhalb derer die Lektionsdauer, die Sozialformen und die inhaltsbezogenen Aktivitäten (z.B. Auseinandersetzung mit dem Bilderbuch, Briefschreiben, Leseübung, Organisatorisches etc.) niedrig inferent kodiert wurden. Auf dieser Grundlage fanden weitere vertiefende Analysen zur Unterrichtsgestaltung und -qualität statt, von denen zwei Beobachtungssysteme in den folgenden beiden Abschnitten skizziert werden, da diese exemplarisch zur Beantwortung der Fragestellungen ausgewählt wurden.

4.2 Die Erfassung von „Einsatz von Lob“ und „Störungsfreiheit“ über ein hoch inferentes Rating

Die Merkmale „Einsatz von Lob“ und „Störungsfreiheit“ wurden mit hoch inferenten Ratings erfasst. Das gesamte hoch inferente Ratingsystem umfasste acht Merkmale zum Classroom Management und zehn zum Unterrichtsklima (vgl. Gabriel, in Vorb.). Alle Merkmale dieses hoch inferenten Ratingsystems wurden fachunspezifisch operationalisiert, d.h. die Indikatoren nahmen weder auf fachspezifische Besonderheiten Bezug, noch erforderte das Rating spezifische fachwissenschaftliche oder fachdidaktische Kenntnisse der Rater. Die Rater wurden in der Anwendung des Manuals umfassend trainiert. Die Entwicklung des Ratingsystems basiert auf einem Verfahren, bei dem sowohl deduktive als auch induktive Herangehensweisen kombiniert werden (Hugener et al.,

	„Einsatz von Lob“	„Störungsfreiheit“
Quellen	Eigenentwicklung in Anlehnung an Hofer (1985)	adaptiert nach Clausen (2002) und Rakoczy & Pauli (2006)
Grundidee (Auszug)	Diese Dimension erfasst, ob die Lehrperson ihre Schüler lobt. […]	Diese Dimension erfasst, inwieweit der Unterricht störungsfrei abläuft und nicht immer wieder durch Störungen unterbrochen oder beeinträchtigt wird. […]
Indikatoren (Beispiele)	▪ Die Lehrperson äußert Freude über die Leistungen bzw. das Verhalten der Klasse bzw. einzelner Schüler („Super!“, „Toll“ etc.).	▪ Der Unterricht läuft ohne Störungen ab. ▪ Der Unterricht ist so geplant und organisiert, dass Disziplinstörungen nicht vorkommen.
Negativ-indikatoren (Beispiel)	▪ [Es wurden keine Negativindikatoren formuliert.]	▪ Die Lehrperson muss wiederholt zur Ruhe mahnen (z.B. wiederholtes „Psst…“).
Antwort-format (4-stufige Skala)	▪ „4“ = Die Lehrperson lobt sehr häufig. ▪ „3“ = Die Lehrperson lobt häufig. ▪ „2“ = Die Lehrperson lobt weniger häufig. ▪ „1“ = Die Lehrperson lobt nicht häufig bis gar nicht.	▪ Eine „4“ wird vergeben, wenn der Unterricht weitgehend störungsfrei abläuft. ▪ Eine „3“ wird vergeben, wenn mehrere kleinere Störungen auftreten, die den Unterrichtsablauf aber nicht beeinflussen. ▪ Eine „2“ wird vergeben, wenn einzelne massive Störungen oder häufigere kleinere Störungen erkennbar sind, die den Unterrichtsablauf beeinflussen. ▪ Eine „1“ wird vergeben, wenn häufig massive Unterrichtsstörungen bzw. große Disziplinprobleme auftreten. Dies äußert sich besonders in unübersichtlich-chaotischen und lauten Situationen, in denen die Lehrperson Mühe hat sich durchzusetzen.
Reliabilität		
Relativer G-Koeffizient	.86	.90
„wahre“ Varianz	74%	79%
Systema-tischer Fehler	1%	4%
Unsystema-tischer Fehler	25%	17%

Tab. 2: Die hoch inferenten Dimensionen „Einsatz von Lob“ und „Störungsfreiheit“

2006; Jacobs, Kawanaka & Stigler, 1999; Seidel, 2003). Gemäß dem hoch inferenten Ansatz wurden im Manual für jedes Merkmal jeweils die Grundidee, die Indikatoren und Negativindikatoren sowie das Antwortformat (4-stufig) ausführlich beschrieben (vgl. Tabelle 2). Als Analyseeinheit wurde die gesamte Unterrichtsstunde gewählt.

Um die Qualität der hoch inferenten Daten zu überprüfen, wurde in der vorliegenden Studie auf den Ansatz der Generalisierbarkeitstheorie zurückgegriffen (vgl. Clausen et al., 2003; Cronbach, Gleser, Nanda & Rajaratnam, 1972). Mit Hilfe dieses Ansatzes kann eine beobachtete Variation in den Urteilen der Rater auf verschiedene potenzielle Varianzquellen zurückgeführt und deren relativer Anteil bestimmt werden. Dadurch wird ausdifferenziert, welcher Anteil der insgesamt beobachteten Variation zwischen den Ratingurteilen (a) auf Unterschiede zwischen den Unterrichtsstunden zurückgeführt werden kann (konstruktbezogene oder „wahre" Varianz), (b) durch Urteilstendenzen der einzelnen Rater zustande kommt (systematische Fehlervarianz) und (c) wie viel unsystematische Variation in die Beurteilung des betreffenden Merkmals einfließt (unsystematische Fehlervarianz, hier konfundiert mit der Wechselwirkung zwischen beobachteter Stunde und Rater). Die Berechnungen erfolgten mit Hilfe des Programms GT (Ysewijn, 1997). In Anlehnung an Rakoczy und Pauli (2006) entspricht ein relativer G-Koeffizient von größer .65 einer zufriedenstellenden Qualität der hoch inferenten Daten. Für das Merkmal „Störungsfreiheit" liegt der relative Generalisierbarkeitskoeffizient bei .86, für „Einsatz von Lob" bei .90. Die prozentuale Verteilung der Varianzkomponenten kann der Tabelle 2 entnommen werden.

4.3 Die niedrig inferente Kodierung von Lehrerreaktionen während der Leseübung

Die niedrig inferenten Codes, die hier zur Validierung herangezogen werden, beziehen sich auf die Art und Weise, in der eine Lehrkraft auf Verhaltensweisen oder Äußerungen von Schülern reagiert. Hierzu wurden in einem ersten Schritt alle verbalen und/oder nonverbalen Reaktionen der Lehrperson auf Schüleräußerungen oder Schülerverhaltensweisen identifiziert. In einem anschließenden Auswertungsschritt wurden darauf aufbauend unterschiedliche Arten von Lehrerreaktionen differenziert betrachtet. Es wurde unter anderem kodiert, welches Schülerverhalten der Reaktion der Lehrperson jeweils vorausging. Dabei wurden beispielsweise lesebezogene Reaktionen (z.B. Reaktionen auf Vorleseverhalten oder Antworten auf inhaltliche Fragen zu dem Lesetext) von den Reaktionen unterschieden, die auf Unterrichtsstörungen durch Schüler erfolgten. Auch das Timing (z.B. unmittelbare vs. verzögerte Reaktion der Lehrkraft) sowie die Form der Reaktion (verbal vs. nonverbal) und deren „Öffentlichkeit" (nur den jeweiligen Schüler adressierend vs. öffentlich) wurden kategorisiert. Diese Aspekte sind zentral für die Analyse der Lehrerreaktionen auf Unterrichtsstörungen, da anhand der Kategorien „Timing", „Form" und „Öffentlichkeit" bestimmt werden kann, inwiefern die Lehrperson Störungen durch minimalen Aufwand unterbindet und den Unterricht dadurch nicht (lange) unterbricht (vgl. z.B. den Low-Profile-Ansatz nach Borich, 2008). Schließlich wurde die affektive Tönung der

Lehrerreaktion untersucht, wobei im vorliegenden Kategoriensystem neutrale von positiv und negativ getönten Reaktionen unterschieden wurden. Affektiv positiv getönte Reaktionen entsprechen inhaltlich dem hoch inferent erfassten „Einsatz von Lob".

Im Gegensatz zum Rating der Klassenführung und des Unterrichtsklimas wurden die Reaktionen der Lehrpersonen nicht für die gesamte Lektionsdauer, sondern nur für die Phase der Leseübung kodiert, die vorab durch eine Kodierung der inhaltsbezogenen Aktivitäten abgegrenzt wurde.

Für die Auswertung wurden zwei Beobachter geschult, die jeweils die Hälfte der Videos kodierten. Für die niedrig inferente Kodierung wurden andere Beobachter trainiert als für die hoch inferenten Ratings. Zu fünf Zeitpunkten wurden anhand von insgesamt sieben Unterrichtsvideos die Beobachterübereinstimmungen für die niedrig inferente Kodierung der Lehrerreaktionen berechnet. Tabelle 3 zeigt, dass die prozentuale Übereinstimmung der Kodierer mit der Master-Kodierung (der Kodierung durch den Entwickler des Beobachtungssystems) ausreichend hoch ist, um von einer reliablen Auswertung sprechen zu können.

Kategorie	Prozentuale Übereinstimmung
Identifikation der Lehrerreaktion (Event-Sampling)	≥ 91.43%
Art des vorausgehenden Schülerverhaltens	≥ 98.47%
Affektive Tönung	≥ 98.98%
Form der Reaktion	≥ 97.27%
Öffentlichkeit der Reaktion	≥ 97.79%
Timing der Reaktion	≥ 98.98%

Tab. 3: Prozentuale Übereinstimmung der Kodierer mit der Master-Kodierung bei der niedrig inferenten Kodierung der Lehrerreaktionen

Für die weitere Auswertung wurden die Kodierungen quantifiziert, und zwar in der Form, dass die Dauer der Leseübung bzw. die Gesamtanzahl der darin beobachteten Lehrerreaktionen in Rechnung gestellt wurde. Als niedrig inferente Kennziffern wurden demnach die Anzahl der auf Störungen bezogenen bzw. der affektiv getönten Reaktionen pro Minute Leseübung einerseits und deren Anteil an der Gesamtzahl aller Lehrerreaktionen andererseits verwendet.

4.4 Stichprobe und Analysemethoden

Für die Analysen konnten die Daten von 42 Lehrpersonen herangezogen werden, da für diese sowohl die hoch inferenten Einschätzungen als auch die niedrig inferenten Kodierungen der Leseübungsphase vorlagen. Diese Anzahl resultiert daraus, dass in zwei Klassen keine Leseübung durchgeführt wurde und in anderen Klassen die Leseübungen in der Unterrichtseinheit zeitlich nur einen sehr geringen Stellenwert von weniger als zehn Minuten einnahmen. Diese wurden von den weiteren Analysen ausgeschlos-

sen, da hier nicht mehr gewährleistet ist, dass die niedrig inferente Kodierung der Leseübung ausreichende Generalisierbarkeit für die Gesamtstunde – für die das hoch inferente Rating angewendet wurde – besitzt. Bei einer durchschnittlichen Unterrichtsdauer von 87 Minuten haben die verbleibenden Leseübungen eine Dauer von 11 bis 56 Minuten, im Durchschnitt etwa 29 Minuten.

Neben Korrelationen zur Überprüfung des Zusammenhangs zwischen niedrig und hoch inferent erfassten Beobachtungsdaten kommen deskriptive Statistiken zum Einsatz, die anhand von Boxplot-Diagrammen dargestellt werden. Mit diesen werden Gruppen, die über das hoch inferente Rating gebildet wurden, anhand der niedrig inferenten Daten verglichen, um die Frage nach der Interpretation der hoch inferenten Ratingstufen zu beantworten.

5. Ergebnisse

Die Darstellung der Ergebnisse erfolgt in zwei Schritten. Zunächst werden die Ergebnisse der Analysen zum Zusammenhang zwischen dem „Einsatz von Lob“ (hoch inferent) und der Häufigkeit affektiv positiver Lehrerreaktionen (niedrig inferent) berichtet (Fragestellung 1.1). Anschließend wird auf die Zusammenhänge zwischen der Einschätzung der „Störungsfreiheit“ (hoch inferent) und der Häufigkeit von Reaktionen auf Unterrichtsstörungen (niedrig inferent) eingegangen (Fragestellung 1.2), und es wird geprüft, ob diese zweite Korrelation tatsächlich niedriger ausfällt als die erstgenannte (Fragestellung 1.3). Ergänzend wird jeweils die Bedeutung der Abstufungen in den hoch inferenten Ratings diskutiert (Fragestellung 2).

5.1 Zusammenhänge zwischen dem „Einsatz von Lob“ (hoch inferent) und der Häufigkeit affektiv positiver Lehrerreaktionen (niedrig inferent)

Im ersten Abschnitt werden zunächst die deskriptiven Ergebnisse zum „Einsatz von Lob“ anhand des hoch inferenten Beobachtungssystems sowie zur Häufigkeit affektiv positiver Lehrerreaktionen anhand der Ergebnisse der niedrig inferenten Kodierung beschrieben. Nach dem Bericht über die Korrelation zwischen den mit den beiden unterschiedlichen Verfahren gewonnenen Daten wird analysiert, wie die Rater die hoch inferenten Ratingstufen interpretieren.

Wie häufig loben die Lehrpersonen?

Betrachtet man zunächst die deskriptiven Ergebnisse der niedrig inferenten Kodierungen (vgl. Tabelle 4), so lässt sich daraus ablesen, dass durchschnittlich 13.9% aller Lehrerreaktionen während der Leseübung affektiv positiv sind, wobei die Spannweite von 2.8% bis 44.9% reicht. Im Durchschnitt reagieren die Lehrpersonen ca. alle zwei bis drei Minuten affektiv positiv auf ein Schülerverhalten.

		Min	Max	M	SD
niedrig inferent	prozentualer Anteil affektiv positiver Reaktionen an der Gesamtanzahl der Lehrerreaktionen	2.82	44.90	13.89	8.78
	Anzahl affektiv positiver Reaktionen pro Minute Leseübungszeit	0.09	1.65	0.61	0.35
hoch inferent	„Einsatz von Lob“	1.00	4.00	2.13	0.78

Tab. 4: Deskriptive Ergebnisse zur Häufigkeit affektiv positiver Lehrerreaktionen und zum „Einsatz von Lob“

Im hoch inferenten Rating liegt der Mittelwert bei 2.1, wobei die Ratingstufe „2“ bedeutet, dass die Lehrperson „weniger häufig“ lobt.

Lässt sich das hoch inferente Rating mit der niedrig inferenten Kodierung validieren?

Die Korrelationskoeffizienten nach Pearson[3] zeigen relativ hohe Zusammenhänge von $r = .69$ bzw. $r = .66$ zwischen den hoch inferent erfassten Daten einerseits und der Häufigkeit affektiv positiv gefärbter Reaktionen pro Minute Leseübung bzw. deren Anteil an der Gesamtzahl aller Lehrerreaktionen andererseits. Die Koeffizienten sind statistisch hoch signifikant (jeweils $p < .001$, zweiseitig). Das bedeutet: Je häufiger die Lehrpersonen laut der niedrig inferenten Kodierung mit positiver affektiver Tönung auf Schülerverhalten reagieren, desto intensiver wird auch der „Einsatz von Lob“ hoch inferent eingeschätzt.

Wie interpretieren die Rater die einzelnen Ratingstufen?

Zunächst wurden anhand der Ergebnisse des hoch inferenten Ratings vier Gruppen von Wertebereichen gebildet, welche anschließend mithilfe der niedrig inferenten Kodierungen deskriptiv analysiert wurden (vgl. Tabelle 5).

3 Um die unterschiedlichen Analyseeinheiten zu berücksichtigen (gesamte Unterrichtsstunde für das hoch inferente Rating; Leseübung für die niedrig inferente Kodierung), wurden zusätzlich Partialkorrelationen berechnet, welche die Zusammenhänge zwischen niedrig inferenter Kodierung und hoch inferentem Rating von einem möglichen Einfluss der Dauer der Leseübung befreien. Da sich hierdurch aber keine bedeutsamen Veränderungen in der Höhe der Korrelationen zeigten, werden nur die Ergebnisse der bivariaten Korrelationsanalysen berichtet.

„Einsatz von Lob“ (hoch inferent)		Häufigkeit von affektiv positivem Feedback (niedrig inferent)	*Min*	*Max*	*M*	*SD*
Werte-bereiche	Anzahl der Klassen					
1.00 - 1.50	12	prozentualer Anteil affektiv positiver Reaktionen an der Gesamtanzahl der Lehrerreaktionen	2.82	23.33	7.99	6.00
1.75 - 2.50	19		3.92	27.73	13.51	5.90
2.75 - 3.50	9		8.07	31.03	17.85	7.77
3.75 - 4.00	2		25.34	44.90	35.12	13.83
1.00 - 1.50	12	Anzahl affektiv positiver Reaktionen pro Minute Leseübungszeit	0.09	1.12	0.38	0.30
1.75 - 2.50	19		0.15	0.93	0.55	0.20
2.75 - 3.50	9		0.36	1.27	0.83	0.28
3.75 – 4.00	2		1.25	1.65	1.45	0.28

Tab. 5: Häufigkeit affektiv positiver Lehrerreaktionen in Abhängigkeit von den hoch inferenten Wertebereichen für das Merkmal „Einsatz von Lob“

In denjenigen Klassen, die im hoch inferenten Rating sehr positive Werte erhalten (3.75 bis 4.00), wurden 35.1% der Reaktionen als affektiv positiv (= jede 3. Reaktion) kodiert. In diesen Klassen erfolgt mehr als einmal pro Minute eine derartige Reaktion. Lehrpersonen, die im Mittel einen Wert um die „3“ erhalten haben (2.75 bis 3.50), reagieren im Durchschnitt alle 1.4 Minuten in dieser Weise (0.8 affektiv positive Reaktionen pro Minute), jede 5. bis 6. Reaktion ist affektiv positiv. Lehrpersonen mit einem Wert um „2“ im hoch inferenten Rating reagieren etwa alle zwei Minuten mit positivem Affekt (0.6 affektiv positive Reaktionen pro Minute), jede 7. Reaktion ist derartig getönt. Bei Lehrpersonen mit Werten um „1“ sind hingegen nur 8.0% der Reaktionen (jede zwölfte) affektiv positiv; eine solche Reaktion der Lehrperson kann in diesen Klassen durchschnittlich nur alle 4.2 Minuten festgestellt werden (0.4 affektiv positive Reaktionen pro Minute). Anhand dieser gruppenweisen Betrachtung lässt sich also im Nachhinein nachvollziehen, was die Rater unter „häufigem“ Lob verstanden haben.

Abbildung 1 macht jedoch deutlich, dass innerhalb der Gruppen die Streuung relativ hoch ist und es dadurch zu Überschneidungen zwischen den einzelnen Ratingstufen kommt. In diesem sogenannten Boxplot werden die verschiedenen Streu- und Lagemaße gemeinsam abgebildet: Innerhalb der „Box“ liegen jeweils 50% der Werte aus der niedrig inferenten Kodierung, der waagerechte Strich bildet den Median ab, und die sogenannten Antennen veranschaulichen den gesamten Streubereich mit Ausnahme einzelner „Outlier“, die durch Punkte gekennzeichnet sind.

Positiv hervorzuheben ist zunächst, dass sich die Boxen zwischen den einzelnen Ratingstufen kaum überschneiden. Überschneidungen zeigen sich fast nur in den „Antennen“. Dies bedeutet, dass den hoch inferenten Ratingstufen grundsätzlich gut unterscheidbare niedrig inferente Kennziffern entsprechen, beide Verfahren sich also wechselseitig validieren.

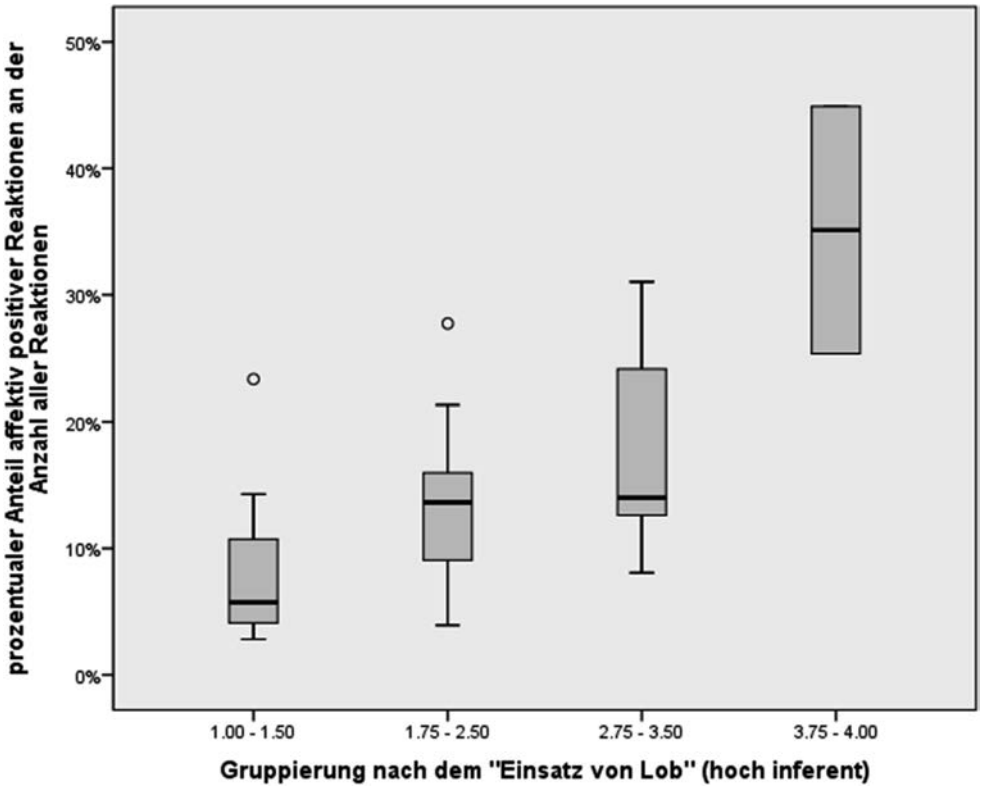

Abb. 1: Gruppenweiser Vergleich der niedrig und hoch inferenten Beobachtung zum „Einsatz von Lob“ und zur Häufigkeit affektiv positiver Lehrerreaktionen

Es fällt allerdings auch auf, dass zwischen den Werten um „2“ und „3“ kaum Unterschiede im Median bestehen. Aus der Abbildung lässt sich erkennen, dass die Abgrenzung der Ratingstufen „1“ und „3“, „1“ und „4“, „3“ und „4“ sowie „2“ und „4“ deutlich ausfällt. Kleinere Überschneidungen zeigen sich hingegen zwischen „1“ und „2“, etwas größere zwischen „2“ und „3“. Die Herstellung einer Gleichabständigkeit zwischen den einzelnen Ratingstufen gelingt nicht. Während „1“, „2“ und „3“ relativ nahe beieinander liegen, ist die „4“ weiter entfernt von den restlichen Ratingstufen.

5.2 Zusammenhänge zwischen der Einschätzung der „Störungsfreiheit“ (hoch inferent) und der Häufigkeit von Reaktionen auf Unterrichtsstörungen (niedrig inferent)

Wie bereits im vorherigen Abschnitt werden auch hier zunächst deskriptive Ergebnisse beider Beobachtungsverfahren aufgeführt, anschließend Zusammenhänge zwischen den Beobachtungsdaten analysiert und abschließend wird der Frage nachgegangen, wie die

Beurteiler die hoch inferenten Ratingstufen interpretieren. Zusätzlich wird exemplarisch aufgezeigt, wie sich hoch inferent gewonnene Beobachtungsdaten durch das zusätzliche Heranziehen niedrig inferent gewonnener Daten konkretisieren lassen.

Inwiefern ist der Unterricht störungsfrei? Wie häufig erfolgen Reaktionen auf Unterrichtstörungen?

Die niedrig inferente Kodierung ergibt, dass während der Leseübung im Durchschnitt 9.7% aller Reaktionen auf Schüleräußerungen oder -verhalten infolge von Störverhalten auftreten und 0.5 Reaktionen auf Unterrichtsstörungen pro Minute auftreten. Dabei zeigen sich aber deutliche Unterschiede zwischen den Lehrpersonen. Während einige Lehrpersonen kaum auf Störungen reagieren (müssen) – die Lehrperson mit der geringsten relativen Häufigkeit reagiert beispielsweise innerhalb einer etwa 55-minütigen Leseübung nur zweimal auf ein Störverhalten –, gibt es andere Lehrpersonen, die bis zu zweimal pro Minute auf ein Störverhalten reagieren und bei denen sich bis zu 28.7% aller Reaktionen auf Unterrichtsstörungen beziehen (vgl. Tabelle 6).

		Min	*Max*	*M*	*SD*
niedrig inferent	prozentualer Anteil von Reaktionen auf Unterrichtsstörungen an der Gesamtanzahl der Lehrerreaktionen	1.05	28.74	9.66	6.35
	Anzahl von Reaktionen auf Unterrichtsstörungen pro Minute Leseübungszeit	0.04	1.74	0.47	0.37
hoch inferent	„Störungsfreiheit“	1.50	4.00	3.01	0.64

Tab. 6: Deskriptive Ergebnisse zu Reaktionen der Lehrpersonen auf Unterrichtsstörungen und zur „Störungsfreiheit“

Die „Störungsfreiheit“, welche hoch inferent eingeschätzt wurde, wird über alle Klassen hinweg mit Werten von 1.50 bis 4.00 bewertet, wobei der Mittelwert mit 3.01 relativ hoch liegt. Laut Definition der Ratingstufe wird die „3“ vergeben, „wenn mehrere kleinere Störungen auftreten, die den Unterrichtsablauf aber nicht beeinflussen“.

Lässt sich das hoch inferente Rating mit der niedrig inferenten Kodierung validieren?

Die Korrelation zwischen der hoch inferent eingeschätzten „Störungsfreiheit“ des Unterrichts und dem Anteil der Reaktionen auf Unterrichtsstörungen an der Gesamtanzahl aller Lehrerreaktionen beträgt $r = –.49$ ($p < .001$, zweiseitig). Dies bedeutet, dass in einem Unterricht, der als relativ störungsfrei eingeschätzt wird, die Lehrperson auch

seltener auf Unterrichtsstörungen reagiert. Betrachtet man die Zahl der Reaktionen auf Unterrichtsstörungen pro Minute Leseübung, so findet sich ebenfalls eine signifikant negative Korrelation mit der „Störungsfreiheit“ ($r = –.46$; $p < .01$, zweiseitig): Je störungsfreier der Unterricht beurteilt wird, desto größer sind auch die Abstände zwischen einzelnen Reaktionen auf Unterrichtsstörungen.

Wie vermutet (Fragestellung 1.3), fallen diese Zusammenhänge deutlich schwächer aus als in Abschnitt 5.1. Das als Häufigkeitseinschätzung definierte Merkmal „Einsatz von Lob“ hängt also stärker mit der niedrig inferenten Kodierung der Häufigkeit affektiv positiver Lehrerreaktionen zusammen als die niedrig inferent erfasste Häufigkeit von Reaktionen auf Störungen mit der hoch inferent erfassten „Störungsfreiheit“, die Häufigkeitseinschätzungen mit einer Beurteilung der Stärke und der Konsequenzen bestimmter Handlungen im Unterrichtsverlauf kombiniert.

Wie interpretieren die Rater die einzelnen Ratingstufen?

In Tabelle 7 sind für die vier anhand des hoch inferenten Ratings der „Störungsfreiheit“ gebildeten Gruppen die Ergebnisse der niedrig inferenten Kodierung dokumentiert.

„Störungsfreiheit“ (hoch inferent)		Häufigkeit von Reaktionen auf Unterrichtsstörungen (niedrig inferent)	*Min*	*Max*	*M*	*SD*
Werte-bereiche	**Anzahl der Klassen**					
1.00 - 1.50	1	prozentualer Anteil von Reaktionen auf Unterrichtsstörungen an der Gesamtanzahl der Lehrerreaktionen	nur eine Klasse		12.09	-
1.75 - 2.50	12		5.04	28.74	14.84	7.96
2.75 - 3.50	20		1.05	17.36	8.41	4.30
3.75 – 4.00	9		1.79	11.49	5.27	3.02
1.00 - 1.50	1	Anzahl von Reaktionen auf Unterrichtsstörungen pro Minute Leseübungszeit	nur eine Klasse		0.53	-
1.75 - 2.50	12		0.32	1.74	0.81	0.45
2.75 - 3.50	20		0.04	1.10	0.36	0.26
3.75 – 4.00	9		0.08	0.56	0.24	0.14

Tab. 7: Häufigkeit von Reaktionen auf Unterrichtsstörungen in Abhängigkeit von den hoch inferenten Wertebereichen für das Merkmal „Störungsfreiheit“

Im Wertebereich um „1“ befindet sich nur eine Klasse, weshalb diese Ergebnisse statistisch nicht interpretierbar sind. In den Wertebereichen um „2“, „3“ und „4“ verhalten sich die Mittelwerte im Vergleich erwartungsgemäß: Je besser der Wert im hoch inferenten Rating, desto seltener erfolgen Reaktionen auf Unterrichtsstörungen. Während sich bei Lehrpersonen, die im hoch inferenten Rating Werte um „2“ erhalten, durchschnittlich 14.8% aller Reaktionen auf Unterrichtsstörungen beziehen (alle 1.7 Minuten lässt sich eine solche Reaktion beobachten; 0.8 Reaktionen auf Störungen pro Minute),

zeichnen sich Lehrpersonen mit sehr guten Werten im hoch inferenten Rating (= 4) auch in der niedrig inferenten Kodierung durch einen relativ geringen Anteil von Reaktionen auf Störungen aus (nur 5.3% der Reaktionen beziehen sich auf Störungen; nur alle 5.7 Minuten erfolgt eine derartige Reaktion bzw. erfolgen nur 0.2 Reaktionen auf Störungen pro Minute).

Die grafische Darstellung zeigt aber auch hier, dass sich zwischen den einzelnen Ratingstufen größere Überschneidungen ergeben (vgl. Abbildung 2).

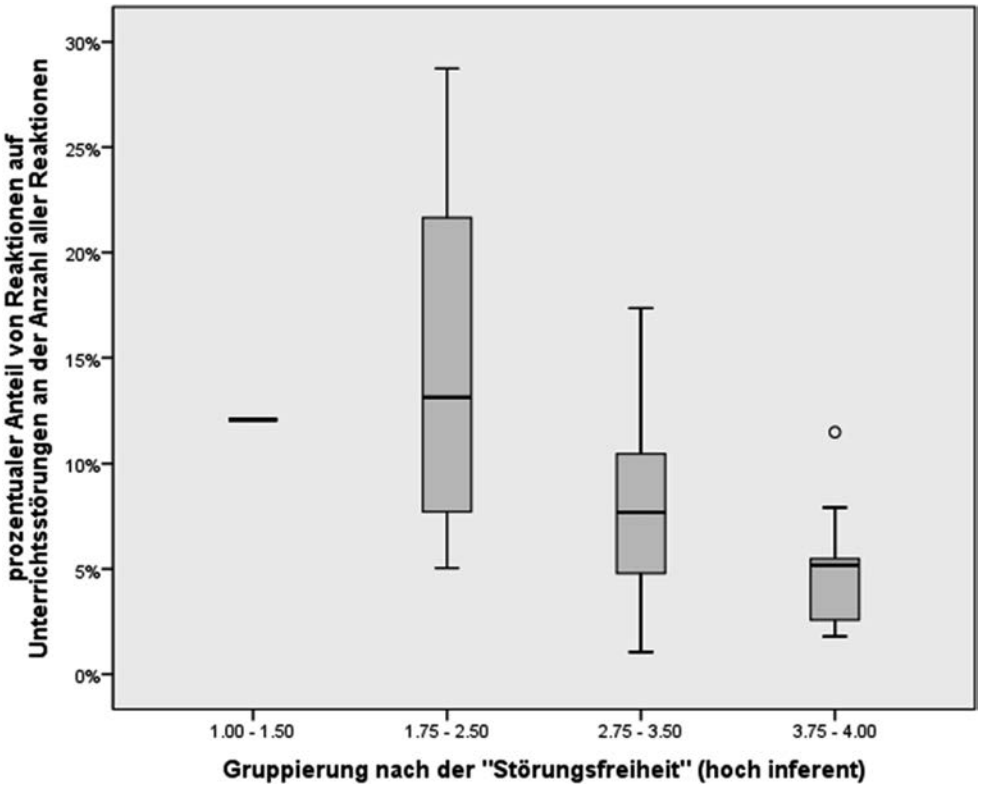

Abb. 2: Gruppenweiser Vergleich der niedrig und hoch inferenten Beobachtung zur „Störungsfreiheit" und zur Häufigkeit der Reaktionen auf Unterrichtsstörungen

Die Boxplots verdeutlichen noch einmal, dass zwar die Mediane der niedrig inferenten Kodierung (prozentualer Anteil von Reaktionen auf Unterrichtsstörungen) erwartungsgemäß mit steigenden Werten im hoch inferenten Rating zurückgehen, dass sich aber insbesondere die Ratingstufen um die Werte „2" und „3" überschneiden. Dies ist unter anderem darauf zurückzuführen, dass die „1" mit der Bedeutung „häufige massive Unterrichtsstörungen bzw. große Disziplinprobleme" nahezu gar nicht vergeben wurde und dadurch die Spannweite eingeschränkt ist. So haben viele Klassen den Wert

„2“ erhalten, die sich laut der niedrig inferenten Kodierung in der Häufigkeit der Reaktionen auf Unterrichtsstörungen deutlich unterscheiden. Auffallend sind insbesondere zwei Klassen, deren Lehrpersonen laut niedrig inferenter Kodierung extrem häufig auf Unterrichtsstörungen reagieren (konkret: in denen fast 30 Prozent aller Lehrerreaktionen nach störendem Schülerverhalten erfolgen), obwohl das hoch inferente Rating dieser beiden Stunden mit 2.50 genau auf dem neutralen Punkt der vierstufigen Skala liegt. Warum wurden diese beiden Stunden nicht mit „1“ („häufige massive Unterrichtsstörungen“) oder zumindest einer glatten „2“ („häufigere kleinere Störungen“) bewertet?

In beiden Klassen dauert die Leseübung etwa 15 Minuten. In Abbildung 3 wird die Art der Reaktionen auf Unterrichtsstörungen anhand der niedrig inferenten Kodierung genauer beschrieben und im Vergleich zum Mittelwert der Gesamtstichprobe dargestellt. Aus theoretischer Sicht erscheint es im Sinne eines optimalen Classroom Managements (z.B. Borich, 2008; Kounin, 1976) sinnvoll, (1) Störungen möglichst unmittelbar zu unterbrechen sowie (2) derart unauffällig auf Störungen durch Einzelschüler zu reagieren, dass die anderen Schüler nicht gestört werden. Dazu kann die Lehrperson beispielsweise (3) nonverbal auf Störverhalten reagieren.

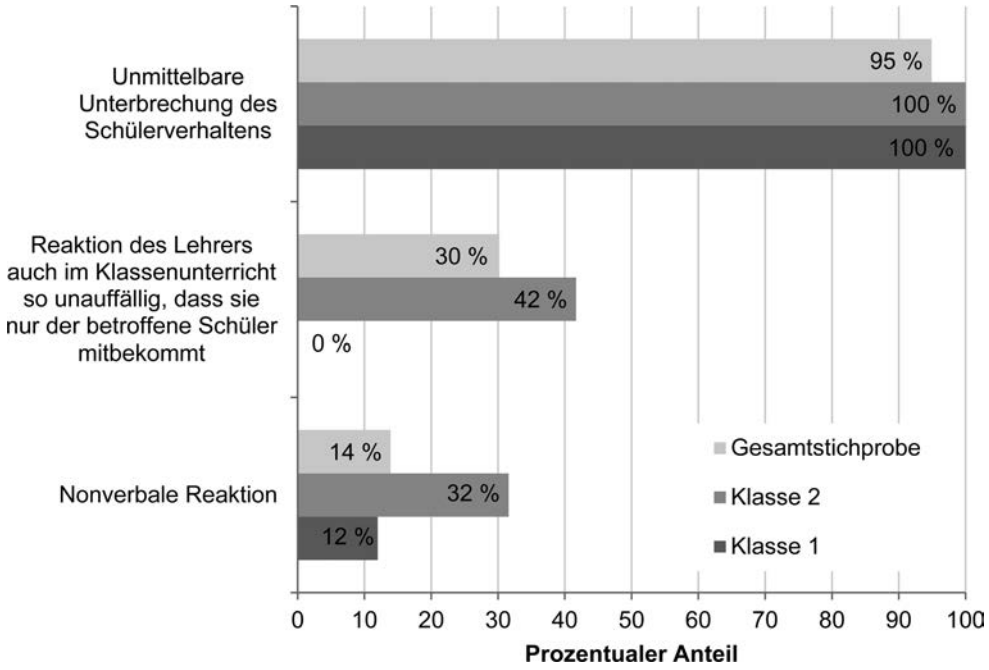

Abb. 3: Analyse zweier ausgewählter Lehrpersonen in ihrem Umgang mit Unterrichtsstörungen

Im Vergleich zur Gesamtstichprobe zeigt sich, dass die Lehrperson der Klasse 2 diesen Kriterien sehr gut entspricht, d.h. Störungen in allen Fällen unmittelbar unterbricht, in fast der Hälfte der Fälle unauffällig reagiert und vergleichsweise häufig nonverbal signalisiert, dass der Schüler das Störverhalten unterlassen soll. Eine erneute Betrachtung des Videos ergab, dass über die gesamte Stunde hinweg eine sehr ruhige Atmosphäre in der Klasse herrscht und die Schüler auch während den Schülerarbeitsphasen sehr leise arbeiten. Die Lehrperson fordert die Schüler dennoch immer wieder auf, leiser zu sein, und scheint eine eher geringe „Störungstoleranz" zu haben. Die häufige Reaktion der Lehrperson auf (vermeintliche) Störungen spiegelt sich also korrekt im hohen Kennwert der niedrig inferenten Kodierung wider, während die Beobachter im hoch inferenten Rating – ebenfalls korrekt – in Rechnung gestellt haben, dass der Unterrichtsverlauf tatsächlich gar nicht durch störendes Verhalten beeinträchtigt wurde.

Für die Lehrperson der Klasse 1 zeigt sich beim nochmaligen Anschauen des Videos, dass im Unterricht viele kleinere Störungen sowie einzelne massivere Störungen vorkommen, die den Unterricht beeinflussen und zu Lehrerreaktionen führen, was sich im niedrig inferenten Kennwert korrekt abbildet. Zudem fällt auf, dass die Lehrperson in Einzelfällen sehr effektiv mit auftretenden Störungen umgeht. So reagiert sie beispielsweise auf eine Störung durch zwei Schüler, die sich unterhalten, indem sie einen der beiden Schüler aufruft, obwohl er sich nicht gemeldet hat, und unterbindet damit unauffällig die Störung. Vermutlich haben die Beobachter im hoch inferenten Gesamturteil solche komplexen pädagogischen Interaktionen berücksichtigt und dementsprechend eine Einstufung gewählt, die trotz allem eine mittlere Störungsfreiheit anzeigt.

Beide Fälle lassen übereinstimmend erkennen, dass die Kombination einer Häufigkeitseinschätzung mit der qualitativen Beurteilung pädagogischer Interaktionen, wie sie beim Merkmal „Störungsfreiheit" verlangt ist, zu Abweichungen der hoch inferenten Einschätzung vom Ergebnis niedrig inferenter Kodierungen führen kann.

6. Diskussion

In der Studie wurde anhand von ausgewählten Daten der Videostudie Deutsch des PERLE-Projekts die Güte hoch inferenter Beobachtungssysteme exemplarisch überprüft, indem Zusammenhänge zu niedrig inferent erfassten Beobachtungsdaten analysiert wurden. Für die hoch inferenten Ratingmerkmale „Einsatz von Lob" und „Störungsfreiheit" konnte gezeigt werden, dass jeweils Zusammenhänge zur niedrig inferent erfassten Häufigkeit bestimmter Lehrerreaktionen bestehen. Somit validieren sich hoch inferente und niedrig inferente Beobachtungen wechselseitig, im Sinne einer konvergenten Validierung. Ergänzende – hier nicht im Detail tabellierte – Befunde sprechen außerdem auch für die divergente Validität: „Einsatz von Lob" hängt nicht mit dem niedrig inferenten Kennwert für Reaktionen auf Störungen zusammen, und „Störungsfreiheit" hängt nicht mit dem Vorkommen positiver affektiver Reaktionen zusammen.

Einschränkend muss jedoch bemerkt werden, dass sich die niedrig und hoch inferenten Beobachtungen nicht auf die gleiche Analyseeinheit beziehen. Während die hoch

inferente Einschätzung über die Gesamtstunde gebildet wurde, bezieht sich die niedrig inferente Kodierung auf einen Stundenausschnitt (die Leseübung), der in den einzelnen Klassen einen zeitlich unterschiedlichen Stellenwert einnimmt. Es ist daher wünschenswert, ähnliche Analysen durchzuführen, bei denen sich beide Formen der Unterrichtsbeobachtung auf dieselbe Analyseeinheit beziehen. Dies ist als einer der nächsten Auswertungsschritte im Rahmen der PERLE-Studie geplant.

Je mehr sich die Definitionen des hoch und niedrig inferent erfassten Verhaltens ähneln, desto höher ist erwartungsgemäß deren Zusammenhang: So zeigen sich in der vorliegenden Studie besonders hohe Zusammenhänge zwischen dem hoch inferent eingeschätzten „Einsatz von Lob“ und der Häufigkeit affektiv positiver Lehrerreaktionen. Betrachtet man die Ratingstufen (Die Lehrperson lobt „sehr häufig“, „häufig“, „weniger häufig“ oder „nicht häufig bis gar nicht“), fällt auf, dass es hier auch im hoch inferenten Rating ausschließlich um eine Erfassung einer Häufigkeit geht. Es werden keine zusätzlichen Qualitätsaspekte berücksichtigt und innerhalb des Items müssen vom Rater auch nicht mehrere inhaltlich unterschiedliche Aspekte zu einem Urteil integriert werden. Gemessen an der Grundidee des hoch inferenten Ratings ist der Grad der interpretativen Schlussfolgerungen für die Rater hier relativ gering.

Im Gegensatz dazu ist die Dimension „Störungsfreiheit“ inhaltlich komplexer: Diese Dimension soll erfassen, inwieweit der Unterricht störungsfrei abläuft und nicht immer wieder durch größere oder kleinere Störungen unterbrochen oder beeinträchtigt wird, sodass er nicht im geordneten Rahmen durchgeführt werden kann. Zur Häufigkeit von Unterrichtsstörungen kommt also auch noch eine Einschätzung dazu, nämlich inwiefern der Unterricht durch diese beeinträchtigt wird. Im Unterschied dazu wurde niedrig inferent lediglich kodiert, wie häufig die Lehrperson auf Unterrichtsstörungen reagiert. Diese beiden Aspekte sind sich inhaltlich zwar nahe, was sich durch die mittelhohe Korrelation auch bestätigt, sie erfassen aber nicht exakt dasselbe, sodass Unterschiede zwischen niedrig und hoch inferent gewonnenen Daten nicht unbedingt auf eine mangelnde Qualität des hoch inferenten Ratings hinweisen müssen. Dies zeigt auch die erneute Betrachtung der Videos aus den zwei Beispielklassen.

Aus den Ergebnissen der vorliegenden Studie ließe sich schlussfolgern, dass die Verbesserung der Validität hoch inferenter Items dann gelingen könnte, wenn die einzelnen Ratingstufen noch klarer und eindeutiger definiert wären. Gerade bei derart formulierten Ratingstufen, die eine Häufigkeit abbilden sollen, wäre es möglich, es nicht dem Ermessen des einzelnen Raters zu überlassen, was er unter „häufig“ versteht, sondern dies in gewisser Weise vorab festzulegen. Dadurch würde die Inferenz verringert werden. Vor dem Hintergrund, dass einmal entwickelte Ratingdimensionen häufig in verschiedenen Studien eingesetzt werden, ist eine genaue Definition der Ratingstufen auch bedeutsam, um Ergebnisse verschiedener Studien besser miteinander vergleichen zu können. Wird ein Ratingsystem für eine Studie adaptiert, könnte auf diese Weise eher gewährleistet werden, dass die Rater in beiden Studien dasselbe Verständnis von beispielsweise „häufigem Lob“ haben.

Das Ergebnis des hoch inferenten Ratings zur „Störungsfreiheit“, bei dem der Wert „1“ für kein einziges Video vergeben wurde, wodurch sehr viele Klassen mit einer „2“

bewertet wurden, legt zudem nahe, noch größere Teilstichproben der Videos oder Videos aus Pilotierungsstudien für die Entwicklung der Beobachtungssysteme heranzuziehen, um schon bei der Definition der Ratingstufen eine ausreichend hohe Varianz in den Einschätzungen zu gewährleisten.

Ein weiterer Vorschlag, um Schwierigkeiten bei hoch inferenten Ratings zu minimieren, könnte lauten, dass man auch bei hoch inferenten Ratings versucht, nicht zu viele unterschiedliche inhaltliche Aspekte mit einem Item zu erfassen (wie hier z.B.: Häufigkeit von Störungen und Grad der Beeinträchtigung des Unterrichts durch die Störungen), sondern sich bemüht, die einzelnen Items „rein" zu halten. Durch die Formulierung mehrerer Items zu einem Inhaltsbereich könnten die komplexen Informationen im Anschluss daran wieder generiert werden. Auch dies würde dazu führen, den Grad an Schlussfolgerungen zu verringern. Dem Ansatz, in hoch inferent zu erfassenden Items jeweils nur eine Facette zu bewerten, wird bereits in anderen Beobachtungsstudien nachgegangen (z.B. Pietsch & Tosana, 2008; Praetorius et al., 2012; Trepke, Seidel & Dalehefte, 2003). So naheliegend die Vorschläge zur Eingrenzung hoch inferenter Ratings auf klarer abgegrenzte Häufigkeitsurteile sind, so fraglich ist doch, ob mit der Annäherung an niedrig inferente Kodierungen nicht zugleich andere, wichtige Aspekte der Validität verloren gingen. Mit einer solchen Eingrenzung ginge nämlich auch ein spezifisches Merkmal hoch inferenter Ratingsysteme verloren, dass nämlich mehrere Aspekte zu einem Globalurteil integriert werden und dadurch Tiefenstrukturen besser erfasst werden können.

In weiteren Analysen wird es deshalb auch darum gehen, zu überprüfen, welche Art von Beobachtungsdaten höhere Zusammenhänge zu Drittvariablen – wie etwa der Leistungs- und Motivationsentwicklung – aufweist, da neben der genauen Analyse unterrichtlicher Prozesse ein besonderes Anliegen der Unterrichtsforschung darin besteht, diejenigen Merkmale zu identifizieren, die „den Unterschied machen" und die Schüler tatsächlich in ihrer Lern- und Persönlichkeitsentwicklung zu fördern vermögen. Auch diese Ergebnisse könnten dann dazu dienen, Beobachtungssysteme entsprechend zu entwickeln und zu validieren.

Hoch und niedrig inferente Verfahren der Unterrichtsbeobachtung haben ihre spezifischen Vor- und Nachteile sowie unterschiedliche Aussagekraft. Die Kombination beider Verfahren erlaubt zum einen eine gegenseitige Validierung der Ergebnisse und kann zum anderen dazu beitragen, ein umfassendes Bild des Unterrichts zu erhalten, indem sowohl übergeordnete Qualitätsmerkmale des Unterrichts bewertet als auch genaue Analysen einzelner Interaktionen zwischen den verschiedenen Akteuren im unterrichtlichen Geschehen vorgenommen werden.

Literatur

Borich, G. D. (2008). *Observation skills for effective teaching*. Upper Saddle River: Pearson Merrill Prentice Hall.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität*. Münster: Waxmann.

Clausen, M., Reusser, K., & Klieme, E. (2003). Unterrichtsqualität auf der Basis hoch-inferenter Unterrichtsbeurteilungen: Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz. *Unterrichtswissenschaft, 31*(2), 122-141.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: John Wiley.

Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge. Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache" (TestDaF). *Diagnostica, 50*(2), 65-77.

Gabriel, K. (in Vorb.). *Videobasierte Erfassung von Unterrichtsqualität in der Grundschule – Eine Teilstudie des PERLE-Projekts zur Erfassung der Klassenführung und des Unterrichtsklimas im Anfangsunterricht* (Dissertation, Universität Kassel).

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64-86.

Hugener, I. (2006). Überblick über die Beobachtungsinstrumente. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis", Teil 3: Videoanalysen* (S. 45-54). Frankfurt a.M.: GFPF/DIPF.

Hugener, I., Rakoczy, K., Pauli, C., & Reusser, K. (2006). Videobasierte Unterrichtsforschung: Integration verschiedener Methoden der Videoanalyse für eine differenzierte Sicht auf Lehr-Lernprozesse. In S. Rahm, I. Mammes & M. Schratz (Hrsg.), *Schulpädagogische Forschung, Unterrichtsforschung, Perspektiven innovativer Ansätze* (S. 41-53). Innsbruck: Studien Verlag.

Jacobs, J. K., Kawanaka, T., & Stigler, J. (1999). Integrating qualitative and quantitative approaches to the analysis of video data on classroom teaching. *International Journal of Educational Research, 31*, 717-724.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik, 54*(2), 222-237.

Kounin, J. S. (1976). *Techniken der Klassenführung*. Bern: Huber.

Lipowsky, F., Faust, G., & Greb, K. (Hrsg.) (2009). *Dokumentation der Erhebungsinstrumente des Projekts „Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE)*. Frankfurt a.M.: GFPF.

Lipowsky, F., & Rakoczy, K. (2006). *Videoanalysen in der Unterrichtsforschung* (Workshop auf der DGFE Summerschool, Ludwigsfelde, September 2006).

Lotz, M., Berner, N. E., Gabriel, K., Post, S., Faust, G., & Lipowsky, F. (2011). Unterrichtsbeobachtung im Projekt PERLE. In D. Kucharz, T. Irion & B. Reinhoffer (Hrsg.), *Grundlegende Bildung ohne Brüche* (S. 183-194). Wiesbaden: VS Verlag für Sozialwissenschaften.

Pauli, C., & Reusser, K. (2006). Von international vergleichenden Video Surveys zur videobasierten Unterrichtsforschung und -entwicklung. *Zeitschrift für Pädagogik, 52*(6), 774-797.

Petko, D., Waldis, M., Pauli, C., & Reusser, K. (2003). Methodologische Überlegungen zur videogestützten Forschung in der Mathematikdidaktik. *Zentralblatt für Didaktik der Mathematik, 35*(6), 265-280.

Pietsch, M., & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. Das Multifacetten-Rasch-Modell und die Generalisierbarkeitstheorie als Methoden der Qualitätssicherung in der externen Evaluation von Schulen. *Zeitschrift für Erziehungswissenschaft, 11*(3), 430-452.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction, 22*(6), 387-400.

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht.Unterricht aus der Perspektive von Lernenden und Beobachtern*. Münster: Waxmann.

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis", Teil 3: Videoanalysen* (S. 206-233). Frankfurt a.M.: GFPF/DIPF.

Reyer, T. (2004). *Oberflächenmerkmale und Tiefenstrukturen im Unterricht. Exemplarische Analysen im Physikunterricht der gymnasialen Sekundarstufe*. Berlin: Logos.

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

Seidel, T. (2003). Videobasierte Kodierverfahren in der IPN Videostudie Physik – ein methodischer Überblick. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht"* (S. 99-111). Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN).

Seidel, T. (2005). Video analysis strategies of the IPN Video Study – a methodological overview. In T. Seidel, M. Prenzel & M. Kobarg (Hrsg.), *How to run a video study. Technical report of the IPN Video Study* (S. 70-78). Münster: Waxmann.

Trepke, C., Seidel, T., & Dalehefte, I. M. (2003). Zielorientierung im Physikunterricht. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht"* (S. 201-228). Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN).

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 171-208). Münster: Waxmann.

Ysewijn, P. (1997). *GT-Programm für Generalisierbarkeitsstudien*. Neuchâtel: Institut de recherche et de documentation pédagogique. http://www.irdp.ch/methodo/generali.htm [03.01.2013].

Abstract: Often, highly inferential estimation procedures for the assessment of the quality of teaching are used in the systematic observation of lessons. The validity of these procedures, however, is frequently questioned since they require a high degree of inferences on the part of the observer. The aim of the present study thus is to examine exemplarily the quality of highly inferential systems of observation within the framework of the PERLE-study by analyzing coherences with lowly inferential observation data. Based on the examples of "the use of praise" and "no disturbances", it can be shown that, basically, coherences do exist between data collected through different procedures. However, in the case of highly inferential ratings, the allocation to individual rating levels is not always unequivocal. Furthermore, it appears that the degree of coherence between lowly and highly inferential observation data also depends on how the characteristics collected through highly inferential procedures have been defined.

Keywords: Classroom Observation, Instructional Quality, Video Analysis, Low-inference Coding, High-inference Rating

Anschrift der Autorinnen/des Autors

Miriam Lotz, Universität Bamberg, Fakultät Humanwissenschaften,
Institut für Erziehungswissenschaft, Lehrstuhl für Grundschulpädagogik und
Grundschuldidaktik, Markusstraße 8a, 96047 Bamberg, Deutschland
E-Mail: miriam.lotz@uni-bamberg.de

M.A. Katrin Gabriel, Goethe-Universität Frankfurt, Fachbereich Erziehungswissenschaften,
Institut für Pädagogik der Elementar- und Primarstufe, Senckenberganlage 15,
60054 Frankfurt am Main, Deutschland
E-Mail: K.Gabriel@em.uni-frankfurt.de

Prof. Dr. Frank Lipowsky, Universität Kassel, Fachbereich Humanwissenschaften,
Institut für Erziehungswissenschaft, Professur für Empirische Schul- und Unterrichtsforschung,
Nora-Platiel-Straße 1, 34109 Kassel, Deutschland
E-Mail: lipowsky@uni-kassel.de