Z Erziehungswiss (2010) 13:121–148
DOI 10.1007/s11618-010-0113-z

ALLGEMEINER TEIL

ZfE

# Evaluation von Unterrichtsstandards

**Marcus Pietsch**

**Zusammenfassung:** Das Kernstück deutscher Schulinspektionen ist die Evaluation von Unterrichtsqualität mithilfe fragebogengestützter Expertenratings. Die für die Leistungsmessung zugrunde gelegten Qualitätsmerkmale rekurrieren zumeist auf bekannte Kriterienmatrizen effektiven Unterrichts und orientieren sich entsprechend am Prozess-Produkt-Paradigma der Schuleffektivitätsforschung. Bislang fehlen jedoch noch sowohl komplexe Unterrichtsqualitätsmodelle als auch Maßstäbe für die Leistungsbeurteilung und Standards für die Leistungsbewertung, die es Schulverantwortlichen und Bildungsadministration erlauben, kriteriale Fragestellungen auf Basis der Evaluationsergebnisse zu beantworten und die Schul- und Unterrichtsentwicklung anhand transparenter Kriterien wissensbasiert zu steuern. Im vorliegenden Beitrag wird die Idee aufgegriffen, auf Grundlage vergleichender empirischer Forschung ein gestuftes Modell von Unterrichtsqualität für die Einordnung von Evaluationsergebnissen zu erstellen. Das Modell wird mithilfe der probabilistischen Testtheorie auf Basis von Daten einer Normierungsstichprobe (N = 2240) der Schulinspektion Hamburg generiert. Die Befunde zeigen, dass es mit Daten aus Schulinspektionsverfahren grundsätzlich möglich ist, ein Modell zu erstellen, das die Überführung quantitativer Messwerte in qualitative Aussagen zur Qualität von Unterricht ermöglicht, das sich zwischen Schulformen invariant verhält und sowohl in den Randbereichen der Skala als auch in deren Mittelbereich gut diskriminiert. Die Datenstruktur ist jedoch mehrdimensional angelegt, sodass bei einer eindimensionalen Modellierung von Unterrichtsqualität ein Informationsverlust und Ungenauigkeiten bei der Bestimmung empirischer Kennwerte zu erwarten sind. Gleichwohl sind die beobachteten Verzerrungen gering und die einzelnen Subdimensionen von Unterrichtsqualität korrelieren teilweise hoch miteinander, sodass davon auszugehen ist, dass ein eindimensionales Stufenmodell zur Beschreibung von Unterrichtsqualität eine sinnvolle, empirisch haltbare Approximation des mehrdimensionalen Modells darstellt. Ein solch abgestuftes Modell wird abschließend mithilfe eines Proficiency Scaling aus den Daten heraus entwickelt und inhaltlich vorgestellt.

**Schlüsselwörter:** Qualitätsstufen · Schulinspektion · Schulrückmeldungen · Unterrichtsqualität · Unterrichtsstandards

## Evaluation of classroom teaching standards

**Abstract:** The core element in German school inspections is the evaluation of the quality of classroom teaching using questionnaire-supported expert ratings. The criteria for performance measurement are in most cases based upon research on effective teaching and are, therefore,

**Online publiziert:** 07.04.2010


Dipl.-Päd. M. Pietsch (✉)
Institut für Bildungsmonitoring Hamburg, Beltgens Garten 25,
20537 Hamburg, Deutschland
E-Mail: marcus.pietsch@ifbm.hamburg.de

oriented on the process-product paradigm of school effectiveness research. Complex models for describing the quality of classroom teaching are missing as well as benchmarks for the assessment of and standards for the appraisal of performance, which allow leading personnel in school and education administration to answer questions based on the results of evaluations and to facilitate evidence-based governance and teaching development. This article suggests the development of a multi-level model for classifying the performance in classroom teaching using comparative empirical research. The data (N = 2240) is derived from a sample of the Hamburg School Inspection and the model has been developed using Item Response Theory. The results show that it is possible to develop a model which allows the conversion of quantitative measurements into performance levels with a narrative description of the content which is typical at each level. Furthermore, it can be demonstrated that the model is valid for different types of schools and that single lesson sequences can be discriminated at the margins of the scale as well as in the middle. Nevertheless, the structure of data is found to be multi-dimensional, so that a uni-dimensional scaling procedure may lead to a loss of information and inaccurate estimates. It can be shown, however, that the expected bias is of little significance and that strong correlations between the sub-dimensions of the model can be found. Thus, it can be assumed that using a uni-dimensional model of performance levels to describe the quality of classroom teaching may be a reasonable and empirically tenable approximation of the multidimensional model. Finally, a multi-level model is presented, which is developed using a proficiency scaling.

**Keywords:** Classroom Teaching Standards · Performance Feedback · Performance Levels · Quality of Classroom Teaching · School Inspection

## 1 Problemstellung und Forschungsfragen

Schulinspektionen in Deutschland sollen Schulen, Öffentlichkeit, Bildungsadministration und -politik Rückmeldungen zum Stand schulischer Prozessqualitäten geben. Dabei ist die externe Evaluation von Prozessen in Schule und Unterricht ein Baustein im Gesamtkonzept der Qualitätssicherung und Qualitätsentwicklung im deutschen Bildungssystem (vgl. Böttcher u. Kotthoff 2007a, b). Ziel ist es, durch die einzelschulische Prozessevaluation Wissen für die verschiedenen Akteure im Bildungssystem bereitzustellen, die Einhaltung prozessualer Mindeststandards zu gewährleisten und eine qualitätssteigernde Wirkung innerhalb von Schulen und im Schulsystem zu entfalten (vgl. Pietsch et al. 2009a). Ein zentraler methodischer Baustein, mit dessen Hilfe diese Ziele erreicht werden sollen, ist in allen Schulinspektionen der deutschen Länder die Beurteilung der Unterrichtsqualität mithilfe standardisierter Beobachtungsverfahren (vgl. Bos et al. 2006; Döbert et al. 2008; Stralla 2009). Ausgewählte Unterrichtssequenzen werden mithilfe fragebogengestützter Expertenratings in ihrer Qualität beurteilt, um anschließend auf Basis dieser Beurteilungen Aussagen zur Qualität von Lehr- und Lernbedingungen auf Ebene der Einzelschule treffen zu können.

Für diese Leistungsmessung legen die Schulinspektionen in den Ländern unterschiedlich ausdifferenzierte Kriterienkataloge zugrunde (vgl. Stralla 2009), wobei sich die Dimensionalität im Aufbau dieser Kataloge grundsätzlich aus den länderspezifischen Qualitätsrahmen für Schulqualität herleitet (vgl. Döbert et al. 2008; Kiper 2008; Maritzen 2007; Stralla 2009) und die konkreten Beobachtungsinstrumente – die Operationalisierung von zu bewertenden Merkmalen der Unterrichtsqualität – wiederum häufig auf Vorarbei-

ten und Annahmen der empirisch-psychologischen und schulpädagogischen Forschung zur Wirkung von Unterricht auf Lernerfolge rekurrieren (vgl. z. B. Dobbelstein 2008; Helmke 2009; Pietsch et al. 2009a; Stralla 2009). Insofern folgt die Messung von Unterrichtsqualität im Rahmen von Schulinspektionen einem Prozess-Produkt-Paradigma, wie es durch die Schuleffektivitätsforschung vertreten wird (vgl. z. B. Ditton 2000; Scheerens u. Bosker 1997; Sammons et al. 1995; Seidel 2008). Die dem Modell zugrunde liegende Annahme lautet: Schulen transformieren Inputs durch innerschulische Prozesse in Outputs. Je höher die Qualität der Prozesse, desto größer die Wahrscheinlichkeit, dass die Ergebnisse besser sind.

Schulinspektionen evaluieren der populären Unterscheidung von Ravitch (1995) folgend vor allem *Opportunity to Learn Standards* – Prozess- und Ausstattungsstandards, die Anforderungen an Ausstattung und Gestaltung von Lernumgebung, an das Vorhandensein spezifischer Programme und Ansprüche an die Gestaltung von Unterricht beschreiben –, wobei diese Standards generell die Gestaltung von Lerngelegenheiten definieren, die es Schülerinnen und Schülern mit hoher Wahrscheinlichkeit ermöglichen, definierte Inhalte (*Content Standards*) zu lernen – und dies möglichst effektiv, sodass zu bestimmten bildungsbiografischen Zeitpunkten klar definierte Kompetenzausprägungen (*Performance Standards)* erreicht werden können. Einen wichtigen Orientierungspunkt für die Messung von Unterrichtsqualität durch Schulinspektionen bilden entsprechend die länderübergreifenden Bildungsstandards, die in den letzten Jahren durch die Kultusministerkonferenz der Länder (vgl. z. B. KMK 2005) verabschiedet wurden. Diese *Output-* oder *Performance Standards* sind, wie Oelkers u. Reusser (2008, S. 406) betonen, „immer auch Prozessstandards (…) und damit als Lehr-Lernstandards zu begreifen".

Geben die länderübergreifenden Bildungsstandards das Ziel vor, dann muss eine an ihnen orientierte Unterrichtsgestaltung inhaltlich gehaltvoll, handlungsorientiert und kognitiv aktivierend sein (vgl. Köller 2008), damit auf Schülerseite eine vielfältige Grundbildung, Strategien zur praktischen Nutzung von Wissen, Fähigkeiten zum lebenslangen selbständigen Lernen sowie eine reflexive Handlungskompetenz aufgebaut werden können (vgl. Klieme et al. 2007; Weinert 2001). Wie eine aktuelle Arbeit von Stralla (2009) zeigt, liegen die Schwerpunkte in den Unterrichtsbeobachtungsbögen der deutschen Schulinspektionen deshalb konsequenterweise auf den Bereichen der Individualisierung, der kognitiven Aktivierung von Schülerinnen und Schülern sowie auf dem Aufbau von Kompetenzen. Die einzelnen Maßnahmen zur Qualitätsentwicklung im deutschen Bildungssystem scheinen somit augenscheinlich ineinander zu greifen.

Maßstäbe für die Bewertung eines kompetenzorientierten Unterrichtens, die eine standardisierte Einordnung von Evaluationsergebnissen zur Unterrichtsqualität erlauben, sind im Bereich der externen Einzelschulevaluation derzeit jedoch nur wenig ausgearbeitet. Einerseits ist der Auflösungsgrad, mit dem Inspektionen die Qualität von Unterricht bestimmen, relativ hoch und Rückmeldungen sind entsprechend differenziert (vgl. Pietsch 2009b; Pietsch et al. 2009a) – in Inspektionsberichten werden zumeist alle Merkmale eines Kriterienkataloges nebeneinander gestellt, einzeln ausgewiesen und spezifische Merkmale ggf. normativ gewichtet, dies jedoch zumeist ohne zugrundeliegende empirische Evidenz (vgl. Meyer 2006). Andererseits erfolgt die kriteriumsorientierte Einordnung von Unterrichtsmerkmalen für die Leistungsbewertung durch Schulinspektionen streng normativ und – aus Sicht der Forschung zur Unterrichtsqualität – unter

teilweise nicht-begründbaren Annahmen (vgl. Meyer 2006). Und wenn soziale Vergleiche abgeboten werden, ignorieren diese zumeist das Problem, dass die zugrunde gelegten Vergleichsgruppen häufig Stichproben sind, die die empirisch-sozialwissenschaftlichen Gütekriterien der Datenerhebung nicht erfüllen und somit im schlimmsten Fall ein ‚schiefes Bild' als Vergleichsmaßstab und Orientierungspunkt bieten (vgl. Bos et al. 2006). Ergänzend kommt hinzu, dass empirische Ergebnisse zum Zusammenspiel der einzelnen Merkmale der Unterrichtsqualität ebenso wie empirisch abgesicherte Modelle zur Unterrichtsqualität, wie sie im Rahmen von Schulinspektionsverfahren gemessen werden soll, aktuell nicht vorliegen (vgl. Dobbelstein 2008).

Dabei scheint aus theoretischer Perspektive vor allem die Darstellung isolierter Einzelmerkmale von Unterrichtsqualität problematisch, besteht doch hinreichend Evidenz, dass einzelne Unterrichtsmerkmale in der Regel zusammenhängen und nicht isoliert betrachtet werden sollten, da ihre Wirksamkeit begrenzt ist (vgl. Brophy 2000; Fraser et al. 1987; Helmke 2003, 2006; Meyer 2004; Seidel u. Shavelson 2007). Hinzu kommt, dass alle Merkmalslisten, die zur variablenzentrierten Messung von Unterrichtsqualität herangezogen werden, in gewisser Weise arbiträr sind und hinsichtlich ihres Auflösungsgrades beliebig ausdifferenziert werden können (vgl. Helmke 2006). Helmke (2003), auf dessen Arbeiten viele Schulinspektionen verweisen, empfiehlt daher explizit, sich bei der Analyse von Daten zur Unterrichtsqualität das Gesamtprofil von Unterrichtsmerkmalen anzusehen, um so auf die Qualität von Unterricht zu schließen. Die rezente empirisch-pädagogische Befundlage weist diesbezüglich darauf hin, dass sich eine allgemeine Unterrichtsqualität, wie sie auch im Rahmen von Schulinspektionsverfahren gemessen werden soll, bereits anhand einer kleinen Anzahl von Basisdimensionen beschreiben lässt. So wurden u. a. in einer Videostudie im Rahmen der *Third International Mathematics and Science Study* (TIMSS) drei Faktoren zweiter Ordnung ermittelt (vgl. Klieme et al. 2001), die sich in ähnlicher Art und Weise ebenfalls im Rahmen des *Programme for International Student Assessment* (PISA) finden ließen (Klieme u. Rakoczy 2003) und mit deren Hilfe sich Lernzuwächse durch effektives Unterrichten erklären lassen. Klieme et al. (2006) schlagen daher vor, diese Dimensionen als empirisch nachweisbare Grunddimensionen guten Unterrichts zu nutzen: a) Strukturierte, klare und störungspräventive Unterrichtsführung, b) unterstützendes, schülerorientiertes Unterrichtsklima und c) kognitive Aktivierung.

Die durch Klieme et al. (2001, S. 54) vorgelegten Befunde der TIMS-Videostudie machen jedoch auch deutlich, dass die drei postulierten Grunddimensionen nicht als unabhängig voneinander zu betrachten sind, sondern vielmehr einen hierarchischen und kumulativen Charakter haben:

> Es ist deutlich, dass guter Unterricht in allen drei Grunddimensionen ausgewiesen sein muss. Jede Dimension erfüllt im Hinblick auf die Leistungs- und Motivationsentwicklung der Schüler und die Sicherung der Arbeitsbedingungen im Klassenverband wesentliche Funktionen. Es ist daher falsch, beispielsweise Schülerorientierung und direktes, störungspräventives Verhalten des Lehrers gegeneinander auszuspielen. (…) Das eine ist die Grundvoraussetzung, auf der kognitiv aktivierende Instruktionsprozesse aufbauen müssen, um erfolgreiches fachliches Verstehen zu ermöglichen.

Mit Blick auf die gängigen Kriterienmatrizen effektiven Unterrichts bedeutet eine Modellierung von Unterrichtsqualität, wie sie im Rahmen von Schulinspektionsverfahren gemessen werden soll, entsprechend, dass zu erwarten ist, dass Unterricht umso erfolgreicher – im Sinne von kompetenz- bzw. lernförderlich – ist, je mehr Merkmale positive Ausprägungen aufweisen, dass Stärken in einigen Merkmalen Schwächen in anderen Merkmalen gegebenenfalls kompensieren können und dass einige Merkmale die Voraussetzung für das Gelingen anderer Merkmale darstellen (vgl. hierzu auch Helmke 2003; Helmke u. Weinert 1997; Meyer 2004).

Das Fehlen von robusten, empirisch haltbaren Bezugsnormen hingegen scheint vor allem insofern problematisch, als ein breiter Konsens dahingehend besteht, dass eine evaluationsbasierte Schulentwicklung vor allem dann erfolgreich sein kann, wenn sich Schulen im Vergleich zu anderen Schulen und/oder anhand kriterialer Maßstäbe einschätzen können (vgl. z. B. Rolff 2007; Visscher u. Coe 2002, 2003). Insbesondere elaborierte Rückmeldeformate, die über ein reines Vermitteln von Evaluationsbefunden im Sinne eines *Knowledge of Results* hinausgehen, haben dabei empirisch nachweislich ein erhöhtes Potenzial, Lern- und Entwicklungsprozesse zu stimulieren (vgl. Bangert-Drowns et al. 1991; Kluger u. deNisi 1996; Kulhavy u. Stock 1989). Rückmeldungen sollten, so Hattie u. Timperley (2007), daher immer Informationen darüber enthalten, welches Ziel grundsätzlich angestrebt wird *(Feed-up)*, wie weit man auf dem Weg zur Zielerreichung bereits vorangekommen ist *(Feed-Back)* und welche Schritte als nächstes auf dem Weg zur Zielerreichung vollzogen werden sollten *(Feed-Forward)*. Dies gilt, wie Ehren u. Visscher (2006, 2008) im Rahmen ihrer Theorie zum Einfluss von Schulinspektionen auf Schulentwicklungsprozesse aufzeigen, umso mehr für Schulen mit geringer Innovationskapazität, für die es im Rahmen von Schulinspektionsverfahren besonders wichtig ist, mit den Rückmeldungen Hinweise zu potenziellen Weiterentwicklungen und ein klares, anhand transparenter Kriterien gezeichnetes Bild eigener Stärken und Schwächen zu erhalten, um auf Basis dieser Informationen zielgerichtete Entwicklungen wissensbasiert angehen zu können.

Von solch elaborierten Rückmeldeformaten sind deutsche Schulinspektionen derzeit jedoch ebenso weit entfernt wie von der Nutzung komplexer Modelle zur Beschreibung von Unterrichtsqualität. Daher müssen diesbezüglich derzeit noch Grundlagenarbeiten geleistet werden. Einen Vorschlag, wie die Modellierung eines Konstrukts der Unterrichtsqualität für den Bereich von Evaluationen aussehen kann, haben jüngst Meyer u. Klapper (2006) im Rahmen der Forderung nach Unterrichtsstandards gemacht, wobei Meyer (2008, S. 78) konkretisiert, dass Unterrichtsstandards den gleichen Ansprüchen wie die länderübergreifenden Bildungsstandards genügen und dabei die folgenden Prämissen erfüllen müssen:

1. Sie müssen an ein theoretisches Modell der Unterrichtsqualität angedockt werden.
2. Sie sollten in sich gestuft dargestellt werden.
3. Und sie sollten standardisiert sein, d. h. in geeichte regional, national oder international gültige Messskalen übertragen worden sein.

Leitende Idee dieses Ansatzes ist die Annahme, dass Unterrichtstandards eine eindimensionale Struktur aufweisen und sich als abgestuftes Modell mit unterschiedlich hohen Anforderungen an das Lernen und Lehren im Unterricht, vergleichbar den Kompetenz-

stufenmodellen der länderübergreifenden Bildungsstandards (vgl. z. B. Bremerich-Vos u. Böhme 2009), darstellen lassen, sich an den gängigen Kriterienmatrizen zur Bestimmung von Unterrichtsqualität orientieren und auf diesem Wege „Lehrern, Schülern, Eltern, Schulleitungen und Inspektoren helfen, die Qualität des Unterrichts verlässlich und nachprüfbar zu bestimmen“ (Meyer 2008, S. 79).

Insbesondere der Vorschlag, Unterrichtsqualität in Form eines abgestuften Modells darzustellen, in dem Informationen durch die Überführung quantitativer Messwerte in qualitative Aussagen zur Qualität von Unterricht derart verdichtet werden, dass sie eine Einordnung von Evaluationsbefunden in eine kriteriale Bezugsnorm mit Best-Practice-Charakter ermöglichen, ist für Inspektionsverfahren interessant. Denn wichtigstes Ziel bei der Modellierung solcher Abstufungen ist es, die Evaluationsergebnisse kriterial interpretier- und somit praktisch nutzbar zu machen (vgl. Pietsch et al. 2009a). Dies ist insofern von Bedeutung, als bekannt ist, dass Schulpraktiker nach wie vor häufig Probleme haben, empirische Befunde zu lesen und zu interpretieren, wenn diese auf komplexen Datenmodellierungen beruhen und als empirische Kennziffern dargestellt werden (vgl. Rolff 2007). Entsprechend empfinden sie vor allem ein kriteriales Rückmeldeformat, das ein eher geringes Abstraktionsniveau zur Beschreibung der Evaluationsergebnisse nutzt, als sinnvoll und gewinnbringend für die Schul- und Unterrichtsentwicklung (vgl. Bonsen et al. 2006). Ein weiterer Vorteil eines solchen Modells liegt darin, dass es unterstellt, dass die Qualität von Unterricht kumulativ-hierarchisch beschrieben werden kann und einzelne Teilbereiche von Unterrichtsqualität entsprechend systematisch aufeinander aufbauen. Auf Basis eines solchen Modells wäre es Schulinspektionen möglich, Schulen im Rahmen von Schulrückmeldungen sowohl Informationen zum Ist-Stand der Unterrichtsqualität als auch zu potenziellen Weiterentwicklungsmöglichkeiten derselben zu geben und ihnen somit transparente Informationen für eine wissensbasierte Schul- und Unterrichtsentwicklung anhand eines empirisch gültigen Modells bereitzustellen.

Nachfolgend wird dieser Ansatz aufgenommen. Behandelt wird dabei insbesondere die Frage, inwieweit es möglich ist, mithilfe von Daten aus Schulinspektionsverfahren ein abgestuftes Modell der Unterrichtsqualität zu erstellen, das empirisch-statistischen Gütekriterien genügt und für den Einsatz durch Schulinspektionen geeignet ist. Hierfür werden Befunde aus Analysen von Daten aus einer Normierungsstichprobe der Schulinspektion Hamburg zur Qualität von Unterricht an Hamburger Schulen dargestellt. Das Modell selber wird, aktuellen empirischen Verfahrensstandards folgend, mithilfe der probabilistischen Testtheorie erstellt. Im Folgenden wird zuerst die Datengrundlage beschrieben. Anschließend wird die grundlegende statistische Vorgehensweise dargestellt und über Analysen zur Modellwahl, Dimensionalität der Daten, zur Item- und Skalenqualität und zu potenziellen differenziellen Itemfunktionen berichtet. Auf diese Weise soll geklärt werden, inwieweit es möglich ist, ein empirisch tragfähiges Stufenmodell der Unterrichtsqualität aus vergleichender Perspektive zu modellieren. Im darauf folgenden Teil des Beitrags wird die Abstufung des Modells inhaltlich behandelt. In einem ersten Schritt wird dargestellt, wie und unter welchen Annahmen und Maßgaben diskrete Abstufungen in der metrischen Skala „Unterrichtsqualität“ vorgenommen wurden. Im zweiten Schritt werden die Abstufungen inhaltlich beschrieben und für die jeweiligen Abstufungen charakteristische Merkmale von Unterrichtsqualität dargestellt.

## 2 Erstellung und Prüfung eines metrischen Modells allgemeiner Unterrichtsqualität

### 2.1 Datengrundlage

Die Grundlage für die Analysen bilden Daten, die die Schulinspektion Hamburg im Zeitraum von Januar bis Juni 2008 im Rahmen einer Normierungsstichprobe an 32 Hamburger Schulen erhoben hat. Die Schulinspektion Hamburg setzt zur Ermittlung dieser Schulstichprobe eine mehrstufige Zufallsauswahl ein, die sich an den Merkmalen Schulform und soziale Zusammensetzung der Schülerschaft der Schule – indiziert über die Hamburger KESS-Indices zu sozialen Eingangsvoraussetzungen von Schülerinnen und Schüler auf Schulebene (vgl. Pietsch et al. 2007) – orientiert. Für die Schulstichprobenziehung werden Schulen in einem ersten Schritt nach Schulform und in einem zweiten Schritt nach den sozialen Eingangsvoraussetzungen ihrer Schülerschaften innerhalb dieser Schulform gruppiert. Anschließend wird aus diesen Gruppierungen eine Anzahl von Schulen zufällig gezogen, die der Verteilung der Schulform- und sozialen Schülerschaftsmerkmale innerhalb dieser Schulformen folgen. Die 32 Schulen bilden das Allgemeinbildende Hamburger Schulsystem, in einem Verhältnis von etwa eins zu 13, ab.

Die Daten zur Messung von Unterrichtsqualität selber wurden mittels Beobachtung von Unterrichtssequenzen erhoben. Für diese Einsichtnahme in den Unterricht standen an zwei bis drei Tagen pro Schule je 20 Minuten pro Beobachtung zur Verfügung. Die Auswahl der zu besuchenden Unterrichtssequenzen erfolgte jeweils vor dem eigentlichen Schulbesuch in Form einer stratifizierten Zufallsstichprobe, wobei hier Unterrichtseinheiten je Schulstunde gezogen wurden. Als Grundgesamtheit wurde die Anzahl von potenziellen Unterrichtsstunden pro Woche an einer Schule zugrunde gelegt. Dies geschah vor dem Hintergrund, dass die Schulinspektion Hamburg Unterricht nicht als ausschließlich von der Lehrerperson abhängig, sondern als Angebot-Nutzungs-Beziehung betrachtet, sodass davon ausgegangen wird, dass beispielsweise auch die Altersspezifität der unterrichteten Schülerschaft sowie der Klassenkontext zu berücksichtigen sind, da diese Determinanten ebenfalls mitentscheiden, in welchem Umfang Schülerinnen und Schüler das Angebot „Unterricht“, das ihnen durch Lehrkräfte unterbreitet wird, überhaupt nutzen können. Die Qualität von Unterricht gilt hier, in Anlehnung an Fend (1998), als eine Ko-Produktion von Lehrkräften und Schülerinnen und Schülern. Es wird also davon ausgegangen, dass es durchaus möglich ist, weniger guten Unterricht bei einer fähigen Lehrkraft zu sehen, wenn lehrkraftunabhängige Merkmale des Unterrichts die Qualität beschränken. Das bedeutet praktisch für die Stichprobenziehung, dass Lehrkräfte im Rahmen einer Schulinspektion ggf. häufiger, jedoch in verschiedensten Kontexten und auch von verschiedenen Inspektionsmitgliedern gesehen werden sollten. Die Zuweisung der Beobachter zu den zu beobachtenden Unterrichtssequenzen erfolgte deshalb ebenso wie die Ziehung der Unterrichtssequenzen randomisiert. Darüber hinaus wurden in rund 10 % aller Fälle Doppelbeobachtungen durchgeführt, um so die Qualität der Bewertungen zu sichern. An reinen Grundschulen wurden – je Schule – 40, an allen anderen Schulformen mindestens 80 Unterrichtssequenzen beobachtet.

Die einzelnen Unterrichtssequenzen wiederum mussten von den Inspektorinnen und Inspektoren anhand eines Bewertungsbogens beurteilt werden, der 30 Kriterien zur Mes-

sung der Unterrichtsqualität umfasst. Dieser Bogen wird ergänzt durch einen Appendix, der die einzelnen Items inhaltlich detaillierter, jedoch nicht erschöpfend illustriert, um den Inspektorinnen und Inspektoren so Anhaltspunkte für beobachtbare Merkmale zu geben. Die 30 Kriterien dienen als Indikatoren für Qualitätsmerkmale des Hamburger Orientierungsrahmen Schulqualität (Behörde für Bildung und Sport 2006) und orientieren sich primär an den Kategorien guten – im Sinne von effektiven – Unterrichts nach Helmke (2006), sodass die eingesetzten Items die Messung von Unterrichtsgelingensbedingungen auf Basis einer Angebots-Nutzens-Beziehung ermöglichen sollen (vgl. Pietsch u. Tosana 2008). In Folge einer explorativen Faktorenanalyse, die mit Daten einer Pilotuntersuchung durchgeführt wurde, wurden die entwickelten Indikatoren im Unterrichtsbogen gemeinsam unter den sechs Kategorienbeschreibungen „Klassenmanagement und Klassenklima“, „Unterricht strukturieren, Methoden variieren“, „Motivieren, intelligent Üben, aktiv Lernen“, „Schülerorientierung und Unterstützung“, „Individuelle Förderung“ sowie „Lernerfolgssicherung“ gruppiert, um Schulen bei der Rückmeldung von Befunden auf Einzelitemebene die Möglichkeit zu bieten, eine Anschlussmöglichkeit an den aktuellen schulpädagogischen (vgl. z. B. Meyer 2004) und pädagogisch-psychologischen (vgl. z. B. Helmke 2003) Diskurs zum Thema Unterrichtsqualität zu finden. Die 30 Items sind auf einer vierstufigen Ratingskala (Skalenniveau: ‚trifft nicht zu‘ bis ‚trifft zu‘) zu bewerten, wobei eine fünfte Kategorie markiert werden konnte, sofern die Unterrichtsbeobachter ein Merkmal für „nicht beobachtbar“ hielten. Grundsätzlich wird davon ausgegangen, dass im Rahmen der 20-minütigen Unterrichtssequenzen nahezu alle Kriterien beobacht- und einschätzbar sind und die Kategorie „nicht beobachtbar“ nur in Ausnahmefällen genutzt wird.[1] Dabei decken die eingesetzten Items, wie Stralla (2009) im Rahmen einer vergleichenden Untersuchung aufzeigt, die national und international gängigen Kriterienkataloge zur Qualität von Unterricht differenziert ab, sodass zu erwarten ist, dass mithilfe des Erhebungsinstrumentes Unterrichtsqualität im Sinne effektiven Unterrichtens differenziert erfasst werden kann.

Grundlage für die nachfolgenden Analysen bilden 2240 Unterrichtsbeobachtungen, wovon 731 (33 %) Sequenzen auf reine Grundschulen, 592 (26 %) Sequenzen auf Grund-, Haupt- und Realschulen, 313 (14 %) Sequenzen auf Gesamtschulen und 604 (27 %) Sequenzen auf Gymnasien entfallen. Diese wurden durch 41 Inspektorinnen und Inspektoren der Schulinspektion Hamburg bewertet. Eine Analyse der vorliegenden Daten, die analog Pietsch u. Tosana (2008) mithilfe der Generalisierbarkeitstheorie (vgl. Brennan 2001) durchgeführt wurde, zeigt: Die Inter-Beobachter-Reliabilität $G_{rater}$ der vorliegenden Stichprobe liegt bei 0.924 und der Varianzanteil, der durch die Beobachter in die Bewertungen eingebracht wird, beläuft sich auf rund 7,7 % der Gesamtvarianz. Hierbei sind Strengeeffekte nur in geringem Maße nachweisbar (29 % der Beurteilervarianz), wohingegen Beurteiler-Item-Interaktionen mit einem Anteil von 71 % an der gesamten Beurteilervarianz das Gros des Beurteilerbias ausmachen.

## 2.2 Statistische Modellierung

Abgestufte Modelle zur Definition kriterialer Standards lassen sich am einfachsten mithilfe von *Item-Response-Modellen* (IRT-Modellen) erstellen. Da es sich bei dem zur Unterrichtsbeobachtung eingesetzten Instrument im technischen Sinne um einen Frage-

bogen mit abgestuftem Antwortformat handelt, muss auf ein probabilistisches Analysemodell zur Modellierung ordinaler Datenstrukturen zurückgegriffen werden. Deren im internationalen Kontext gebräuchlichste Formen sind das *Partial-Credit-Modell* nach Masters (1982) und das *Rating-Scale-Modell* nach Andrich (1978). Um zu prüfen, welches der beiden Modelle angemessener für den Umgang mit den vorliegenden Daten ist, wurden mithilfe der Software ConQuest (Wu et al. 1998), die auch für die weiteren IRT-Analysen des Beitrages genutzt wurde, sowohl ein *Rating-Scale-* als auch ein *Partial-Credit-Modell* berechnet. Dabei wurde die Skala im Sinne eines *Powertests* modelliert. Das heißt: In die Analyse wurden auf Fallebene nur die tatsächlich beobachteten Items aufgenommen und in den Randsummen der Datenmatrix berücksichtigt.[2] Wie ein Modellvergleich zeigt, ist ein *Partial-Credit-Modell* besser auf die vorhandenen Daten anwendbar als ein *Rating-Scale-Modell* ($\Delta D = 1354$, $df = 57$, $p < 0{,}001$). Entsprechend bildet ein solch allgemeines IRT-Modell – bei dem für jede Ausprägung eines jeden Items eine separate Itemcharakteristikfunktion beschrieben wird – für ordinalskalierte Daten die Grundlage der nachfolgenden Berechnungen.

### 2.3 Dimensionalität

Generell ist es sinnvoll, eine Skala zur Unterrichtsqualität für die Nutzung im Rahmen von Schulinspektionsverfahren als eindimensionales Konstrukt abzubilden, da die Inspektionen in den Ländern aus ökonomischen Gründen in der Regel nur kleine Itemmengen im Rahmen ihrer Unterrichtsbeobachtungsbögen nutzen (vgl. Stralla 2009). Eine reliable mehrdimensionale Skalierung, die es erlaubt, Kennwerte oder gar Zuordnungen zu Abstufungen für einzelne Teilbereiche von Unterricht auf Subskalen differenziert und ohne Boden- und Deckeneffekte auszuweisen und auf diesem Wege im Rahmen von Rückmeldungen auf Einzelschulebene empirisch zuverlässig analytische Detailfragen dazu zu beantworten, *warum* eine bestimmte Qualität von Unterricht nicht erreicht wird, ist aus Gründen der Datenqualität daher nahezu unmöglich. Hinzu kommen die postulierte Annahme, dass einzelne Unterrichtsmerkmale „ein Qualitätsnetzwerk von sich gegenseitig unterstützenden Faktoren“ (Meyer u. Klapper 2006, S. 100) bilden, und der empirische Befund, dass bestimmte Merkmale von Unterrichtsqualität die Voraussetzung dafür darstellen, dass andere Prozesse gelingen können, bestimmte Teilbereiche von Unterricht also kumulativ-hierarchisch aufeinander aufbauen (vgl. Klieme et al. 2001). Folglich spricht eine Vielzahl von Gründen dafür, Unterrichtsqualität für die Nutzung im Rahmen von Schulinspektionsverfahren als eindimensionales Konstrukt zu modellieren.

Gleichwohl kann die statistische Modellierung einer eindimensionalen Skala unter Nutzung mehrdimensionaler Items im Rahmen der Item-Response-Theorie ggf. mit Verzerrungen in der Schätzung von Item- und Personenparametern einhergehen (vgl. Chen u. Thissen 1997; Yen 1984, 1993), denn lokal abhängige Items sind potenziell redundant und enthalten daher weniger Informationen als im IRT-Modell unterstellt (vgl. Sireci et al. 1991). Daher können bei solchen Fehlspezifikationen auch Skalenreliabilitäten überschätzt werden (vgl. Wainer u. Thissen 1996). In der Regel kann als erster Hinweis auf eine mehrdimensionale Modellstruktur die Verletzung der grundlegenden IRT-Annahme der lokalen stochastischen Unabhängigkeit von Items gelten. Wird entsprechend eine lokale Abhängigkeit von Items (LID – *Local Item Dependence*) aufgedeckt, so impliziert

diese auch, dass zusätzliche Dimensionen im Modell vorhanden sind, die in einem eindimensionalen IRT-Modell nicht ausmodelliert wurden (vgl. Reckase et al. 1988; Yen 1984, 1993).

**Prüfung auf lokale Abhängigkeit von Items.** In einem ersten Schritt wurde daher geprüft, ob die Annahme der lokalen stochastischen Unabhängigkeit der eingesetzten Items verletzt wird. Hierzu gibt es verschiedene Verfahren, wobei die gebräuchlichsten die Korrelation von Residuen zwischen Variablen nutzen, um eine Abhängigkeit zwischen Items aufzudecken (vgl. Chen u. Thissen 1997; Ferrara et al. 1997; Huynh et al. 1995; Yen 1984, 1993). Wie Huynh et al. (1995) zeigen konnten, hängen viele dieser Indices in sehr hohem Maße zusammen, wobei die verschiedenen Indices zu nahezu äquivalenten Mittelwerten und Standardabweichungen kommen. Ein mit gängiger Standardsoftware besonders einfach und elegant zu berechnender Residualindex ist der PRT-Index (*PRT steht für partielle Korrelation*). Bei diesem Maß wird eine partielle Inter-Item-Korrelation berechnet, wobei der Rohwert der Gesamtskala herauspartialisiert wird. Hierdurch können die Residuen der Items bestimmt werden. Die gemittelte Korrelation der so ermittelten Residuen aller binären Itemkombinationen eines Instruments ist der PRT-Index und gibt Auskunft darüber, ob und, falls ja, inwieweit die Annahme der lokalen stochastischen Unabhängigkeit verletzt wird.

Für diesen Index gilt, da er Kennwerte vergleichbar dem gängigen Q3-Index (vgl. Yen 1984, 1993) einnimmt: je niedriger die Korrelation der Residuen und je geringer deren Streuung, desto unabhängiger sind die einzelnen Items voneinander (vgl. Huynh et al. 1995). Ein Wert nahe Null dieses Index weist darauf hin, dass eine Eindimensionalität zu erwarten ist; hohe Werte, ebenso wie große Standardabweichungen der Statistiken, deuten hingegen auf eine mehrdimensionale Datenstruktur hin (vgl. Reckase et al. 1988; Yen 1984, 1993), wobei hoch-negative Werte nachweisen, dass Itempaare verschiedene latente Konstrukte messen, während hoch-positive Werte nachweisen, dass Itempaare dasselbe latente Konstrukt messen (vgl. Habing et al. 2005). Als *Benchmark* für eine Auffälligkeit gelten dabei Indexwerte von größer 0.20 (vgl. Yen 1993).

Die Kennwerte des hier berechneten PRT-Index machen deutlich, dass eine Mehrdimensionalität zu erwarten sein sollte. Für alle Subdimensionen liegen die PRT-Statistiken im positiven Bereich und die Annahme der lokalen stochastischen Unabhängigkeit wird somit bei Modellierung eines eindimensionalen IRT-Modells verletzt. Betrachtet man die PRT-Statistiken für die inhaltlichen Itemgruppierungen in den Beobachtungsbögen im Detail, dann zeigt sich, dass insbesondere im Bereich „Individuelle Förderung“ (PRT = 0,24) die Items zu starke Abhängigkeiten voneinander aufweisen und dieser Bereich mit einem Indexwert größer 0,20 bedenklich scheint. Am unauffälligsten ist der Bereich „Motivieren, intelligent Üben, aktiv Lernen“ mit einem mittleren Korrelationskoeffizienten in Höhe von PRT = 0,03 (SD = 0,05) sowie der Bereich „Unterricht strukturieren, Methoden variieren“ (PRT = 0,06, SD = 0,09). Die PRT-Statistiken für die weiteren Itemgruppen liegen zwischen diesen Extrempolen und deuten somit ebenfalls auf eine tendenziell mehrdimensionale Datenstruktur hin, wobei die Kennwerte aber keine bedenklichen Ausmaße annehmen, da der PRT-Index kleiner als 0,20 ausfällt. Entsprechend ist somit zwar eine Mehrdimensionalität zu erwarten; gleichwohl sollten even-

tuelle Verzerrungen von Kennwerten aufgrund der konstatierten lokalen Abhängigkeiten nur in äußerst geringem Maße auftreten.

**Prüfung auf Modellgültigkeit.** Um weiterführend zu prüfen, ob auch aus vergleichender Perspektive eine ein- oder mehrdimensionale Struktur die vorhandenen Daten besser beschreibt, wurde nachfolgend ein Modellvergleich durchgeführt. Hierfür wurde in einem ersten Schritt ein eindimensionales IRT-Modell berechnet und in einem zweiten Schritt ein sechsdimensionales IRT-Modell generiert, das den Itemgruppierungen im Hamburger Unterrichtsbeobachtungsbogen folgt. Diese Modelle können mithilfe des informationstheoretischen Index BIC (*Bayesian Information Criterion,* vgl. Schwartz 1978) verglichen werden. Dieser Index ermöglicht es, Auskunft darüber zu geben, welches der getesteten Modelle am besten zu den vorliegenden Daten passt, nicht aber, welches Modell als absolut gut – im Sinne von modellkonform – gelten kann. Je kleiner der Index ausfällt, desto besser passt ein Modell auf die Daten. Das eindimensionale Modell weist einen BIC in Höhe von 131947 auf, wohingegen das mehrdimensionale Modell einen BIC von 125191 aufweist. Entsprechend zeigt sich auch hier, dass die Daten aus den in Hamburg durchgeführten Unterrichtsbeobachtungen eher eine mehrdimensionale als eine eindimensionale Struktur aufweisen.

**Prüfung des Zusammenhangs der Dimensionen und des Generalfaktors.** Nichtsdestotrotz lassen sich hohe Zusammenhänge zwischen den einzelnen Subdimensionen (D1 bis D6) und der eindimensionalen Gesamtskala (G) „Unterrichtsqualität“ sowie moderate bis hohe Zusammenhänge zwischen den jeweiligen Subdimensionen nachweisen. Dies machen messfehlerfreie Korrelationen deutlich, die abschließend mithilfe eines direkten Schätzverfahrens berechnet wurden (vgl. Tab. 1).

In der vorliegenden Studie liegen die latenten Korrelationen zwischen den einzelnen Subdimensionen der Skala und der Gesamtskala im Bereich von ca. $r=0{,}75$ bis $r=0{,}85$. Einzig die aus einem Item bestehende Dimension „Lernerfolgssicherung“ hängt mit einem Korrelationskoeffizienten in Höhe von $r=0{,}67$ nur moderat mit der Gesamtskala zusammen. Bei Betrachtung der Zusammenhänge zwischen den einzelnen Subdimensionen hingegen fällt auf, dass diese teilweise relativ deutlich voneinander diskriminie-

**Tab. 1:** Latente Korrelationen der Subdimensionen (D1 bis D6) sowie der Gesamtskala (G) Unterrichtsqualität (im unteren Triangel), Interne Konsistenz der Skalen (Cronbachs $\alpha$, auf der Hauptdiagonalen)

| | **G** | D1 | D2 | D3 | D4 | D5 | D6 |
|---|---|---|---|---|---|---|---|
| **G** | **0.928** | | | | | | |
| D1 | 0.814 | **0.906** | | | | | |
| D2 | 0.839 | 0.776 | **0.716** | | | | |
| D3 | 0.876 | 0.403 | 0.736 | **0.843** | | | |
| D4 | 0.760 | 0.641 | 0.654 | 0.613 | **0.744** | | |
| D5 | 0.739 | 0.442 | 0.642 | 0.660 | 0.642 | **0.740** | |
| D6 | 0.676 | 0.681 | 0.808 | 0.652 | 0.829 | 0.727 | – |

ren; insbesondere der Bereich „Klassenmanagement und Klassenklima“ zeigt geringe Zusammenhänge mit den anderen Bereichen. Und vor allem die Bereiche „Motivieren, intelligent Üben, aktiv Lernen“ sowie „Individuelle Förderung“ sind von diesem Bereich des Unterrichts vergleichsweise unabhängig ($r=0{,}40$ und $r=0{,}44$). Die weiteren latenten Korrelationen zwischen den einzelnen Subdimensionen bewegen sich in etwa im Bereich von $r=0{,}65$ bis $r=0{,}75$. Entsprechend deuten die vorliegenden Befunde darauf hin, dass grundsätzlich ein Modell mit Subdimensionen Unterrichtsqualität angemessener beschreibt als ein eindimensionales Modell. Nichtsdestotrotz weisen die vorgelegten Analysen aber auch darauf hin, dass auf Ebene der Gesamtskala relativ robuste Aussagen zur Qualität von Unterricht getroffen werden können, sofern man bereit ist, einen moderaten Informationsverlust und geringe Verzerrungen in den Parametern zugunsten einer einfachen Kommunizier- und Darstellbarkeit in Kauf zu nehmen.

### 2.4 Item- und Skalenqualität

Trotz des Befundes der tendenziellen Mehrdimensionalität ist es somit möglich, das Konstrukt „Unterrichtsqualität“ als eindimensionale Skala darzustellen. Diese 30-Item-Skala hat eine interne Konsistenz, gemessen als Cronbachs $\alpha$, von 0,928. Dabei weisen die eingesetzten Items, wie Tab. 2 zeigt, Mittelwerte von 1,51 (Die Schüler/innen arbeiten zeitweise selbstgesteuert.) bis 3,38 (Die Schüler/innen gehen freundlich und respektvoll miteinander um.) auf. Inwieweit die eingesetzten Items geeignet sind, das Konstrukt „Unterrichtsqualität“ als Item-Response-Modell zu beschreiben, lässt sich weiterhin mithilfe von *Mean-Square-Fit-Statistik* (MNSQ, vgl. Smith et al. 1998) und Trennschärfen der Items ($r_{it}$) überprüfen.

Der MNSQ ist ein Residualmaß, das Aufschluss über den Unterschied von empirisch beobachteter und empirisch auftretender Häufigkeit von Itemlösungen gibt und somit ein Maß für Verzerrungen in der Messung ist. Damit Items als passgenau im Sinne einer IRT-Modellierung von Bewertungen gelten können, sollten die MNSQ-Werte möglichst unter 1,40 bzw. bei Bewertungen, bei denen eine Übereinstimmung erwünscht ist, bei unter 1,20 liegen (vgl. Pietsch u. Tosana 2008; Wright u. Linacre 1994).

Bei Betrachtung der Trennschärfe ordinaler Daten ist es wiederum relevant, dass diese einerseits auf Einzelitemebene hoch genug sein sollte, um einzelne Unterrichtseinheiten innerhalb der Skala möglichst genau zu diskriminieren und dass, betrachtet über alle Items der Skala, sowohl Items mit hoher als auch mittlerer Trennschärfe vorkommen sollten, da so gewährleistet wird, dass durch die eingesetzten Items sowohl gut zwischen Unterrichtseinheiten mit hoher und niedriger Qualität als auch im Mittelbereich der Skala diskriminiert wird (vgl. Rost 2004).

Legt man diese Kriterien zugrunde, so lässt sich *grosso modo* feststellen, dass die eingesetzten Items der Skala zur Unterrichtsqualität diese Gütekriterien erfüllen. So lassen sich mit Blick auf die Trennschärfen der eingesetzten Items keine Auffälligkeiten feststellen. Diese liegen durchweg im Bereich von 0,37 bis 0,68 und differenzieren somit über die gesamte Skala gut aus. Insgesamt weisen sechs der 30 eingesetzten Items eine mittlere Trennschärfe ($0{,}30<r_{it}<0{,}50$) und 24 Items eine hohe Trennschärfe ($r_{it}\geq 0{,}50$) auf. Auch bei Betrachtung des MNSQ fallen keine Items auf. Einzig Item 19 (Im Unterricht werden überfachliche Zusammenhänge aufgezeigt.) und Item 23 (Die Schüler/innen

**Tab. 2:** Itemkennwerte für die Gesamtskala „Unterrichtsqualität“

| Dimension | Nr. | Itemformulierung | MW | SE | $r_{it}$ | MNSQ |
|---|---|---|---|---|---|---|
| Klassenmanagement und Klassenklima (D1) | 1 | Die Unterrichtszeit wird effektiv genutzt. | 2,94 | 0,02 | 0,54 | 1,01 |
| | 2 | Das Unterrichtstempo ist angemessen. | 3,04 | 0,01 | 0,61 | 0,87 |
| | 3 | Der Unterricht erfolgt auf Basis eines festen Regelsystems, für dessen Einhaltung die Lehrkraft sorgt. | 3,12 | 0,02 | 0,58 | 0,90 |
| | 4 | Die Lehrkraft behält den Überblick über unterrichtsbezogene und unterrichtsfremde Aktivitäten der Schüler/innen. | 3,05 | 0,02 | 0,54 | 0,96 |
| | 5 | Die Lehrkraft geht mit Störungen angemessen und effektiv um. | 2,82 | 0,02 | 0,57 | 0,94 |
| | 6 | Die Arbeitsaufträge und Erklärungen sind von der Lehrkraft angemessen, klar und präzise formuliert. | 3,09 | 0,02 | 0,60 | 0,89 |
| | 7 | Dem Unterricht liegt eine klare Struktur zugrunde; ggf. reagiert die Lehrkraft schüler- und situationsgemäß flexibel. | 3,11 | 0,02 | 0,66 | 0,80 |
| | 8 | Der Umgangston zwischen Lehrkraft und Schüler/innen ist wertschätzend und respektvoll. | 3,32 | 0,01 | 0,52 | 0,94 |
| | 9 | Die Schüler/innen gehen freundlich und rücksichtsvoll miteinander um. | 3,38 | 0,01 | 0,43 | 1,02 |
| Unterricht strukturieren, Methoden variieren (D2) | 10 | Die Lernziele der Unterrichtsstunde werden thematisiert oder sind den Schüler/innen offensichtlich bekannt. | 2,86 | 0,02 | 0,58 | 0,95 |
| | 11 | Die Schülerinnen und Schüler sind über den geplanten Unterrichtsablauf und die einzelnen Unterrichtsschritte informiert. | 2,81 | 0,02 | 0,57 | 1,03 |
| | 12 | Die Unterrichtsmethoden werden angemessen eingesetzt und ggf. variiert. | 2,90 | 0,02 | 0,68 | 0,81 |
| Motivieren, intelligent Üben, aktiv Lernen (D3) | 13 | Der Unterricht eröffnet Spielräume und ist nicht nur auf eine richtige Antwort fixiert. | 2,67 | 0,02 | 0,56 | 1,05 |
| | 14 | Die Lehrkraft gestaltet den Unterricht so, dass bei den Schüler/innen mehrere Sinne angesprochen werden. | 2,74 | 0,02 | 0,47 | 1,19 |
| | 15 | Die Schüler/innen werden angeregt/angeleitet den Unterricht aktiv mitzugestalten, oder sie gestalten den Unterricht aktiv mit. | 2,24 | 0,02 | 0,56 | 1,03 |
| | 16 | Der Erwerb von Arbeitstechniken und Lernstrategien wird durch die Lehrkraft gezielt unterstützt. | 2,58 | 0,02 | 0,64 | 0,89 |

**Tab. 2:** (Fortsetzung)

| Dimension | Nr. | Itemformulierung | MW | SE | $r_{it}$ | MNSQ |
|---|---|---|---|---|---|---|
| | 17 | Die Schüler/innen haben die Möglichkeit, Kommunikations- und Argumentationstechniken zu lernen bzw. anzuwenden. | 2,54 | 0,02 | 0,52 | 1,11 |
| | 18 | Die Lehrkraft bezieht den Erfahrungshorizont und/oder die Interessen der Schüler/innen in den Unterricht mit ein. | 2,85 | 0,02 | 0,55 | 1,00 |
| | 19 | Im Unterricht werden überfachliche Zusammenhänge aufgezeigt. | 1,99 | 0,02 | 0,41 | 1,27 |
| | 20 | Die Schüler/innen bearbeiten Aufgaben, die problemlösendes und/oder entdeckendes Lernen fördern. | 2,47 | 0,02 | 0,56 | 1,05 |
| | 21 | Die Schüler/innen bearbeiten Aufgaben, die einen klaren Alltags- und/oder Berufsbezug haben. | 2,82 | 0,02 | 0,48 | 1,14 |
| | 22 | Die Schüler/innen arbeiten (zeitweise) selbstorganisiert an vorgegebenen Aufgaben. | 2,34 | 0,02 | 0,56 | 1,13 |
| | 23 | Die Schüler/innen arbeiten (zeitweise) selbstgesteuert. | 1,51 | 0,02 | 0,37 | 1,29 |
| | 24 | Die Reflexion eigener Lernprozesse ist Bestandteil des Unterrichts. | 1,89 | 0,02 | 0,51 | 1,12 |
| Schülerorientierung und Unterstützung (D4) | 25 | Die Lehrkraft geht mit Schülerfehlern konstruktiv um. | 2,92 | 0,01 | 0,55 | 0,90 |
| | 26 | Die Lehrkraft gibt den Schüler/innen differenzierte Leistungsrückmeldungen. | 2,53 | 0,02 | 0,56 | 0,98 |
| | 27 | Die Lehrkraft verstärkt individuelle Lernfortschritte und/oder Verhaltensweisen durch Lob und Ermutigung. | 2,80 | 0,02 | 0,56 | 0,97 |
| Individuelle Förderung (D5) | 28 | Die Lehrkraft berücksichtigt die individuellen Lernvoraussetzungen der einzelnen Schüler/innen in der Unterrichtsgestaltung. | 2,19 | 0,02 | 0,63 | 0,93 |
| | 29 | Die Lehrkraft fördert die Schüler/innen entsprechend ihrer individuellen Lernvoraussetzungen. | 1,96 | 0,02 | 0,49 | 1,13 |
| (D6)* | 30 | Das Erreichen der Lernziele wird angemessen überprüft. | 2,59 | 0,02 | 0,59 | 0,95 |

*Lernerfolgssicherung

arbeiten zeitweise selbstgesteuert.) weisen mit MNSQ-Werten in Höhe von 1,27 und 1,29 bei gleichzeitig vergleichsweise geringen Trennschärfen von 0,41 und 0,37 eine leichte Tendenz zu ungenauen Differenzierungen auf.

Weiteren Aufschluss über die Qualität der Skala gibt Abb. 1. Dargestellt ist hier ein *Item-Mapping* (vgl. Stone et al. 1999; Zwick et al. 2001) der Analysedaten, das wie folgt zu lesen ist: Jede Ausprägung („trifft nicht zu" bis „trifft zu") der 30 eingesetzten Items (zu erkennen an der Itemnummer 1 bis 30 mit dem Zusatz .2 für „trifft eher nicht zu", .3 für „trifft eher zu" und .4 für „trifft zu" in der Spalte Itemschwierigkeit) hat eine spezifische Auftretenswahrscheinlichkeit, die zwischen den Items konstant gehalten wird und aufgrund der gleichzeitigen Darstellung von Itemschwierigkeit und Qualität des Unterrichts auf einer gemeinsamen Skala, mit einen Mittelwert von Null und einer Standardabweichung von Eins, ein Indikator für das Auftreten einer bestimmten Unterrichtsqualitätsausprägung ist.

Sichtbar wird hier, dass die 30 Items mit ihren insgesamt 120 Ausprägungen das Spektrum der Unterrichtsqualitätsskala komplett abdecken; es gibt somit für alle Qualitätsausprägungen Indikatoren im Unterrichtsbeobachtungsbogen der Schulinspektion Hamburg. Besonders wichtig ist dabei, dass sich die einzelnen Itemstufen über das gesamte Spektrum der Unterrichtsqualität erstrecken, sodass diese auch in den Randbereichen der Skala gut ausdifferenziert werden kann. Decken- oder Bodeneffekte sollten somit beim Einsatz der Skala nicht zu erwarten sein. Diesen Befund unterstützt eine berechnete Item-Separationsstatistik, die darüber Auskunft gibt, wie gut die 30 eingesetzten Items die Qualität der einzelnen Unterrichtssequenzen ausdifferenzieren, und die entsprechend eines klassischen Cronbachs α interpretiert werden kann (vgl. Clauser u. Linacre 1999): Mit einer WLE-Separationsreliabilität in Höhe von 0,913 differenzieren die eingesetzten Items die Unterrichtseinheiten auf dem Qualitätskontinuum hochreliabel aus.[3]

Abbildung 1 verdeutlicht weiterhin, dass die Schwellen der einzelnen Items weder als äquidistant noch als parallel zu betrachten sind. Beides variiert zwischen einzelnen Items: Während beispielsweise bei Item 6 („Die Arbeitsaufträge und Erklärungen sind von der Lehrkraft angemessen und präzise formuliert.") bereits bei einem Skalenwert von unter minus Eins, also mehr als einer Standardabweichung unterhalb des Skalenmittelwertes von Null, die Kategorie „trifft eher zu" (Kategorie 3) mit hoher Wahrscheinlichkeit gekreuzt wird, ist eine solch überdurchschnittliche Bewertung des Items 24 („Die Reflexion eigener Lernprozesse ist Bestandteil des Unterrichts.") erst bei einem Skalenwert von größer Eins, also mehr als einer Standardabweichung oberhalb des Skalenmittelwertes von Null, zu beobachten. Das heißt auch, dass – während eine relativ hohe Bewertung des sechsten Items kein Indikator dafür ist, ob die Qualität eines Unterrichts insgesamt hochwertig ist, ergo viele Items der Skala positive Ausprägungen aufweisen – eine relativ hohe Bewertung des Items 24 ein Indikator für einen insgesamt hochwertigen Unterricht im Sinne der Schulinspektion Hamburg ist.

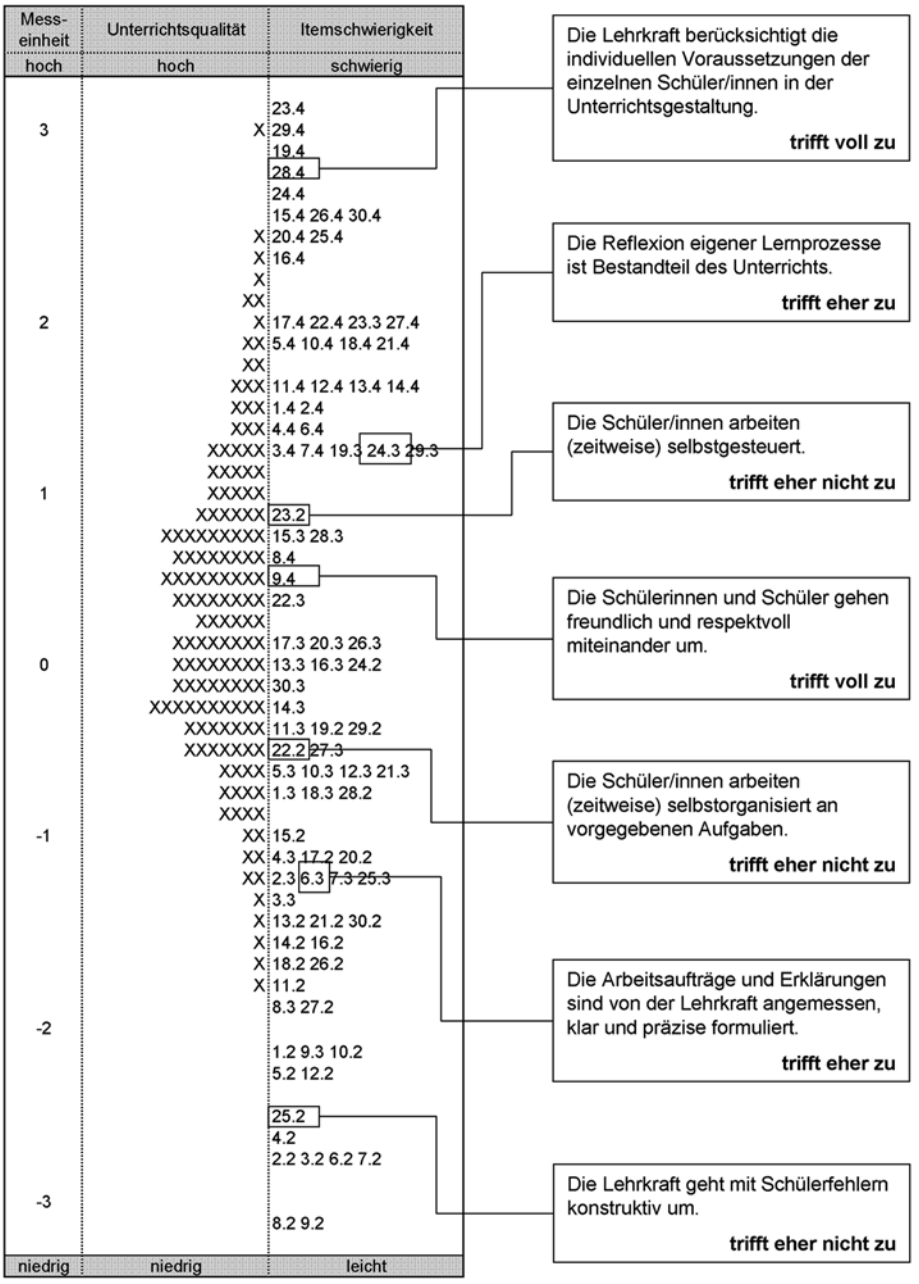


**Abb. 1:** Unterrichtsqualität sowie Verteilung der Itemschwierigkeitsparameter auf dem Qualitätskontinuum

## 2.5 Differenzielle Item-Funktionen

Anders als bei der Entwicklung nationaler Bildungsstandards geht es bei der Entwicklung von Standards für den Unterricht im Rahmen von Schulinspektionsverfahren darum zu gewährleisten, dass ein Instrument entwickelt wird, das möglichst universell einsetzbar ist. Ein Multimatrixdesign, in dem z. B. verschiedene Kriterien für unterschiedliche Schulformen genutzt werden, erscheint zwar theoretisch möglich, widerspricht aber dem Anspruch, alle Schulen am gleichen Maßstab zu messen. Eine weitere relevante Frage ist in diesem Kontext daher, ob ein solches Modell allgemeingültig ist oder ob bestimmte Schulen, Schulformen etc. bei der Bestimmung von Unterrichtsqualität systematisch benachteiligt respektive bevorzugt werden.

Ob Unterschiede zwischen Schulen und Schulformen vorliegen, lässt sich relativ einfach mithilfe einer hierarchischen Varianzzerlegung herausfinden. Im Rahmen der Messung kann so gezeigt werden, dass bedeutsame Unterschiede kaum auf institutionelle Effekte zurückzuführen sind (vgl. Abb. 2). Rund 12 % der Gesamtvariation liegen zwischen Schulen und Schulformen; 88 % der Unterschiede in der Unterrichtsqualität hingegen finden sich innerhalb von Hamburger Schulen, sind also auf einzelne Lehr-Lern-Settings zurückzuführen. Dabei liegen die geringen institutionellen Schulformunterschiede (fünf Prozent der Gesamtvariation) mit einem Anteil von zwei Dritteln des Effektes (67 %) vor allem zwischen reinen Grundschulen und Schulen, die reine Sekundarschulen sind, bzw. solchen, die neben einem Grundschul- auch einen Sekundarschulzweig führen. Folglich lassen sich mit einer Varianzaufklärung von unter zwei Prozent bzw. einem Anteil am Schulformeffekt von 33 % nur äußerst geringe Schulformeffekte im Sekundarschulbereich nachweisen.

Entsprechend ist in erster Linie zu überprüfen, inwieweit Unterschiede zwischen reinen Grundschulen und anderen Schulformen vorliegen. Dies lässt sich mithilfe von Analysen zum *Differential Item Functioning* (DIF, vgl. Holland u. Wainer 1993) überprüfen. Geprüft wird hier, ob einzelne Kriterien zwischen verschiedenen Gruppen invariant sind,

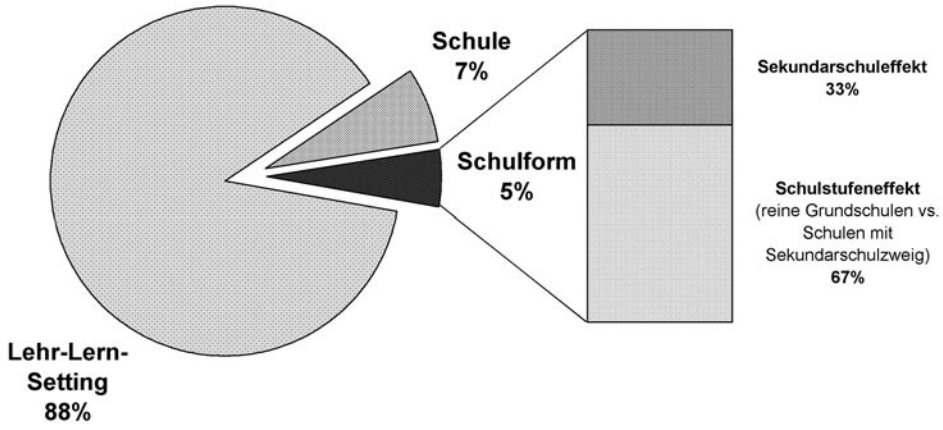


**Abb. 2:** Institutionelle und Lehr-Lern-Setting-bedingte Varianzanteile der Unterrichtsqualität

also für alle Schulformen gleichermaßen gut messen und Vergleiche somit legitim sowie fair sind. Hierfür werden die Kennwerte für die Subpopulationen gemeinsam berechnet und der Einfluss der Schulformen auf die Schwierigkeitsparameter der Kriterien bestimmt. Um hierbei zwischen Effekten, die sich auf tatsächliche Unterschiede zurückführen lassen, und solchen, die auf unfaire Kriterien zurückzuführen sind, zu unterscheiden, können Vorgaben zur Bestimmung substanzieller DIF-Effekte genutzt werden, die Draba (1977) bereits in den 1970er-Jahren vorgeschlagen hat. Demnach sind substanzielle DIF-Effekte nachweisbar, wenn die Unterschiede zwischen Subpopulationen einerseits statistisch signifikant sind und andererseits mehr als eine halbe Standardabweichung betragen.

Die Ergebnisse dieser Analyse zeigen, dass die Unterschiede zwischen reinen Grundschulen und Schulen mit einem Sekundarschulzweig bei rund 36 % einer Standardabweichung liegen. Gleichwohl ist diese Differenz grundsätzlich auf wahre Unterschiede in der Performanz und nicht auf generelle Unterschiede der Eignung des Messinstruments für unterschiedliche Schulformen zurückzuführen. Denn nur für die Kriterien 2, 7, 17, 18, 20, 24, 25 und 26 lässt sich ein statistisch signifikanter Unterschied zwischen Grundschulen und Schulen mit Sekundarschulzweig konstatieren, der 20 % einer Standardabweichung oder mehr beträgt. Auffallend ist, dass es bei allen genannten Kriterien für Grundschulen leichter ist, besser bei den Bewertungen abzuschneiden, als für Schulen mit Sekundarschulzweig. Gleichwohl liegen für sieben der acht genannten Kriterien die Abweichungen zwischen 20 und 28 % einer Standardabweichung; auffällig ist nur Kriterium 20, bei dem der Unterschied 43 % einer Standardabweichung beträgt. Zieht man jedoch alle zu bewertenden Kriterien in Betracht, dann zeigt sich, dass der Unterschied qua Schulstufe bei einem Prozent liegt, eine systematische Benachteiligung von Sekundarschulen in den Messungen somit nicht konstatiert werden kann.

## 3 Erstellung und Beschreibung eines abgestuften Modells der Unterrichtsqualität

Generell zeigen die berichteten empirischen Befunde, dass es möglich ist, die Qualität von Unterricht, wie sie mithilfe des Unterrichtsbeobachtungsbogens der Schulinspektion Hamburg gemessen wird, empirisch valide als eindimensionales IRT-Modell darzustellen. Um die so bestimmte Unterrichtsqualität inhaltlich interpretierbar machen zu können, wurde final ein *Proficiency Scaling* (vgl. Beaton u. Allen 1992) durchgeführt, um auf diesem Wege die kriterienorientierte Interpretation von Werten auf der Skala Unterrichtsqualität zu ermöglichen. Hierzu wurde wie folgt verfahren: In einem ersten Schritt wurde die Gesamtskala „Unterrichtsqualität“ psychometrisch motiviert in diskrete Abstufungen eingeteilt. Wichtig hierbei ist zu wissen, dass sowohl Anzahl als auch Abstände zwischen einzelnen Abstufungen in gewissem Maße arbiträr sind und es keine richtige Methode gibt, mit deren Hilfe solche Abstufungen definiert werden können (vgl. Kolen u. Brennan 2004). Gerechtfertigt werden können diese Abstufungen immer erst im Nachhinein über die Merkmale, die sie umfassen, sowie über die Einschätzung von Experten – also letztlich anhand ihrer Praktikabilität und theoretischen Anbindungsmöglichkeit. Wie viele Abstufungen empirisch angemessen sind und eine adäquate Beschreibung erlauben, lässt sich jedoch berechnen, indem man die in den Messungen beobachtete Standardabweichung unter Berücksichtigung von Verteilungsmaßen und Meßfehlersignifikanzen in

Relation zum durchschnittlichen Standardfehler der Messung setzt (vgl. Pietsch u. Tosana 2008; Wright u. Masters 2002).

Nach Berechnungen mit diesem Verfahren ist es möglich, die Skala in rund 4,5 Stufen zu unterteilen. Somit besteht die Möglichkeit, vier oder fünf Abstufungen zu modellieren. Gleichwohl ist bei der Modellierung von diskreten Abstufungen metrischer Skalen zu beachten, dass neben der Genauigkeit, mit der evaluiert wurde, also der Reliabilität der Messung, auch die Anzahl der im Modell eingezogenen Stufen einen Einfluss darauf hat, wie genau die Zuordnung einzelner Einheiten zu den jeweiligen Abstufungen erfolgen kann (vgl. Huynh 1990; Ercikan 2006; Ercikan u. Julian 2002; Pietsch et al. 2009b). Bei einer Reliabilität von rund 0,91, wie sie im Rahmen der statistischen Modellierung beobachtet wurde, erscheint es sinnvoll, nicht mehr als vier Stufen zu beschreiben, da so eine akkurate Zuordnung von Unterrichtssequenzen zur Qualitätsstufe mit ca. 80-prozentiger Wahrscheinlichkeit erfolgen kann (vgl. Ercikan u. Julian 2002). Mit jeder Stufe, die zusätzlich eingezogen würde, würde bei gleichbleibender Reliabilität der Messungen auch die Genauigkeit der Zuordnung um ca. 10 Prozentpunkte und somit die Interpretierbarkeit der Ergebnisse zunehmend sinken (vgl. Ercikan 2006).

Entsprechend wurde die metrische Skala in Abstimmung mit Fachleuten für den Bereich der Unterrichtsentwicklung des Hamburger Landesinstituts für Lehrerbildung und Schulentwicklung in vier diskrete Abstufungen eingeteilt. Die Modellierung der Abstufungen folgte dabei im weitesten Sinne dem von Beaton u. Allen (1992) vorgeschlagenen und im Rahmen von TIMSS 1995 genutzten Ansatz zur Modellierung von Kompetenzstufen (vgl. Klieme et al. 2000). Entsprechend wurden nach einer ersten Inspektion der vorliegenden Itemschwierigkeitsparameter die einzelnen Abstufungen als äquidistant angenommen, wobei es relevant war, dass sich eine hinreichend große Anzahl von Items (mindestens 5) und Itemstufen (mindestens 10) in einer Abstufung befanden, um eine inhaltliche Beschreibung der Stufe vornehmen zu können. Die Schwellen (*Cut Scores*) wurden daher auf dem Mittelpunkt der Skala sowie bei neun Zehnteln einer Standardabweichung ober- und unterhalb des Skalenmittelwertes gelegt. Theoretisch sollten somit rund 63,2 % aller Unterrichtssequenzen auf den Stufen II und III und je 18,4 % auf den Stufen I und IV liegen.

Anschließend wurden die definierten Abstufungen auf Basis einer holistischen Betrachtung aller Schwierigkeitsparameter der Itemschwellen sowie der auf Itemebene kumulierten Stufenschwierigkeitsparameter für den jeweiligen Schwierigkeitsbereich *post-hoc,* vergleichbar der Interpretation von Faktoren im Rahmen einer Faktorenanalyse, inhaltlich beschrieben. Dabei wurde jedes Item resp. jede Itemausprägung genau einer Stufe zugeordnet und zur inhaltlichen Beschreibung der jeweiligen Abstufung diejenigen Charakteristika herangezogen, die einen gemeinsamen Schwierigkeitsbereich hinreichend beschreiben, sich jedoch von den darunter bzw. darüber liegenden Bereichen unterscheiden. Für die Abstufungen gilt, dass sie sich zueinander probabilistisch verhalten, d. h., dass bei Erreichen einer Abstufung darunter liegende Merkmale mit höherer und darüber liegende Merkmale mit niedrigerer Wahrscheinlichkeit zu beobachten sind. Diese Itemcluster ermöglichen es, die Stufen wie in Tab. 3 dargestellt zu charakterisieren.

Die inhaltliche Beschreibung von Itemgruppen erlaubt es nun, beobachtete Unterrichtssequenzen danach einzuteilen, welche Abstufung sie erreicht haben. So bilden z. B. alle Sequenzen, bei denen der IRT-skalierte Gesamtwert über alle 30 Items des Unterrichtbe-

**Tab. 3:** Inhaltliche Beschreibung der Abstufungen von Unterrichtsqualität

**Stufe 1: Lernklima und pädagogische Strukturen sichern**
Im Unterricht, dessen Qualität auf Niveaustufe 1 liegt, werden die notwendigen Voraussetzungen für erfolgreiches und anspruchsvolles Lernen gelegt. In der Regel gelingt auf diesem Niveau bereits die Sicherung eines lernförderlichen Unterrichtsklimas. Auch liegen den Unterrichtseinheiten normalerweise klare Strukturen zugrunde, die durch die Lehrkräfte im Bedarfsfall schüler- und situationsgemäß abgewandelt werden. Darüber hinaus werden vereinbarte Regeln zumeist eingehalten und Arbeitsaufträge durch die Lehrkräfte überwiegend klar, präzise und an der unterrichteten Schülerschaft orientiert formuliert.

**Stufe 2: Klassen effizient führen und Methoden variieren**
Im Unterricht, dessen Qualität die Niveaustufe 2 erreicht, gelingt es in der Regel, nicht nur die Grundstrukturen zum Gelingen von Unterricht (Niveaustufe 1) abzusichern, sondern darüber hinaus auch eine effiziente Klassenführung – z. B. durch vorausplanendes Handeln der Lehrkräfte und Optimierung der aktiven Lernzeit – und eine Variation von Methoden im Unterrichtsgeschehen zu gewährleisten. Erste Grundlagen, die zum Gelingen von Individualisierung und Differenzierung beitragen können, sind ebenfalls beobachtbar: Die Verstärkung individueller Lernfortschritte durch Lob und Ermutigung, aber auch die Anpassung des Unterrichtstempos an die Bedürfnisse der Schülerschaft und die Gegebenheiten der jeweiligen Lernsituationen sind zumeist vorhanden.

**Stufe 3: Schüler motivieren, aktives Lernen und Wissenstransfer ermöglichen**
Im Unterricht, dessen Qualität die Niveaustufe 3 erreicht, gelingt es zumeist, Schülerinnen und Schüler auf vielfältige Art und Weise zum Lernen zu motivieren. Der Unterricht erfolgt teilweise schülerorientiert und das Lernverständnis ist nicht mechanistisch geprägt. Schülerinnen und Schüler können das Unterrichtsgeschehen in angemessenem Rahmen mitbestimmen und werden befähigt, sowohl aktiv als auch selbständig zu lernen. Die Diagnose von Lernständen erfolgt häufig mithilfe transparenter Verfahren. Lernfortschritte werden durch die Bereitstellung von Transfermöglichkeiten konsolidiert, was eine nachhaltige Auseinandersetzung mit Unterrichtsinhalten ermöglicht.

**Stufe 4: Differenzieren, Schüler wirkungs- und kompetenzorientiert fördern**
Im Unterricht, der die höchste Niveaustufe 4 erreicht, gelingt es nicht nur, die bisher beschriebenen Gelingensbedingungen guten Unterrichts zu gewährleisten, sondern es werden darüber hinaus auch hohe Anforderungen hinsichtlich der Schülerorientierung, Binnendifferenzierung und Individualisierung des Lernens erfüllt. Der Fokus des Unterrichts liegt auf der Ermöglichung eines nachhaltigen Kompetenzerwerbs und ist in der Regel sowohl durch den Einbezug überfachlicher Zusammenhänge als auch durch die Nutzung transparenter Diagnose- und Feedbackverfahren charakterisiert. Schälerinnen und Schüler erhalten teilweise auch die Möglichkeit, an selbst gewählten und für sie bedeutsamen Lerninhalten zu arbeiten. Die Reflexion des eigenen Lernens und der eigenen Lernprozesse ist ein wichtiger Bestandteil des Unterrichts.

obachtungsbogens neun Zehntel einer Standardabweichung unterhalb des Skalenmittels liegt, die Gruppe der Sequenzen, in denen es vornehmlich darum geht, grundlegende Gelingensbedingungen effektiven Unterrichtens zu sichern. Die Sequenzen, die über alle Items hinweg einen Gesamtwert von mehr als neun Zehntel, einer Standardabweichung oberhalb des Skalenmittels, erreichen, repräsentieren hingegen die Teilpopulation, in der eine kompetenzorientierte Förderung der Schülerinnen und Schüler mit hoher Wahrscheinlichkeit gelingen kann, da hier mit hoher Wahrscheinlichkeit eine kognitive Aktivierung stattfindet. Das Konstrukt „Unterrichtsqualität“ wird so gegenüber Schulpraktikern und Schulöffentlichkeit leichter und eindeutiger kommunizier- und darstellbar und es ist möglich, empirisch verlässliche, inhaltlich aussagekräftige Rückmeldungen zur

Qualität von Unterricht zu geben, ohne dabei komplexe empirische Kennzahlen nutzen oder elaborierte methodische Verfahren erklären zu müssen.

## 4 Zusammenfassung und Diskussion

Im vorliegenden Beitrag wurde untersucht, ob und wieweit es möglich ist, mithilfe von Daten aus Schulinspektionsverfahren ein gestuftes Modell der Unterrichtsqualität zu erstellen, das Qualitätsstandards der empirischen Schul- und Sozialforschung genügt. Mithilfe von Daten der Schulinspektion Hamburg, die in einer repräsentativen Zufallsstichprobe an Hamburger Schulen erhoben wurden, wurde ein Vorschlag zur Modellierung von Unterrichtsstandards aufgegriffen, den Meyer u. Klapper (2006) und Meyer 2008) unterbreitet haben. Unterrichtsstandards sollen sich demnach an den länderübergreifenden Bildungsstandards orientieren, auf Merkmalslisten ‚guten Unterrichts' rekurrieren, für die empirische Evidenz besteht, indem diese Merkmale zu Lernerfolgen führen, und als eindimensionales, abgestuftes Modell auf Basis vergleichender empirischer Forschung definiert werden. Dabei sollten die Standards auf eine empirisch geeichte, gültige Messskala übertragen werden können.

Mit Blick auf das erstellte Stufenmodell der Unterrichtsqualität zeigt sich, dass es möglich ist, ein solch abgestuftes Modell zu konstruieren. Dabei ist die größte Herausforderung die Dimensionalität des Modells. Wie die Analysen verdeutlichen, ist es zwar möglich, ein Generalfaktormodell, d. h. eine eindimensionale Struktur von Unterrichtsqualität, zu erstellen; gleichwohl ist eine mehrdimensionale Modellstruktur tendenziell angemessener und bildet die Unterrichtswirklichkeit besser ab. Die vorgelegten Analysen untermauern damit die rezenten Befunde von Klieme et al. (2001, 2006), dass effektiver Unterricht anhand differenzierter Facetten beschrieben werden muss, die jedoch nicht als unabhängig voneinander zu betrachten sind. Denn einerseits passt ein mehrdimensionales Modell besser auf die vorliegenden Daten als ein eindimensionales Modell, andererseits lassen sich teilweise sehr hohe Zusammenhänge zwischen einzelnen Subdimensionen des Modells nachweisen. Gleichwohl wird es im Rahmen von Einzelschulevaluationen nahezu unmöglich sein, reliable Aussagen zur Qualität von Unterricht auf Ebene von Subdimensionen zu treffen, da diese Dimensionen aus ökonomischen Gründen nur durch eine geringe Anzahl von Items indiziert werden. Insofern ist es sinnvoll, das eindimensionale Modell als verhältnismäßig robuste Approximation des mehrdimensionalen Modells für die Bestimmung und Rückmeldung zur Qualität von Unterricht auf Ebene einzelner Schulen zu nutzen.[4] Die Reduzierung auf eine Dimension führt dabei ggf. zu leichten Verzerrungen der Item- und Personenparameter. Die beobachtete Größenordnung scheint jedoch, wie Analysen der Itemresiduen zeigen, nicht bedenklich. Auch Fehlklassifikationen von Unterrichtssequenzen auf Abstufungen sollten durch die Unterkomplexität des IRT-Modells nur in geringem Maße zu erwarten sein (vgl. Walker u. Beretvas 2003).

Gestützt wird dieses Vorgehen dadurch, dass verschiedene Untersuchungen zeigen, dass es theoretisch legitim (vgl. Diamantopoulus et al. 2008) und empirisch vertretbar (vgl. Reckase et al. 1988) ist, mehrdimensionale Konstrukte eindimensional darzustellen. Die eindimensionale Modellierung des mehrdimensionalen Konstrukts hinge dann von der Kombination der Subfaktoren ab (vgl. Robitzsch 2009). Für den vorliegenden

Fall bedeutet dies: Die normativ definierten Subdimensionen bilden in einer linearen Kombination das Gesamtkonstrukt „Unterrichtsqualität“ im Sinne des Hamburger Orientierungsrahmens Schulqualität formativ ab, wobei die jeweils in Gruppen zusammengefassten Merkmale, die einzelnen Subdimensionen reflektiv messen. In diesem Fall wäre die eindimensionale Skalierung dann eine Approximation des formativen Strukturmodells, die es erlaubt, die Vorteile des Modells, wie z. B. die einfache Interpretierbarkeit der Itemladungen, zu nutzen (vgl. Robitzsch 2009). Möchte man diesem Ansatz weiter folgen, ist in weiteren Analysen u. a. zu prüfen, inwieweit die einzelnen Subskalen die Gütekriterien eines eindimensionalen IRT-Modells erfüllen und welchen Einfluss die disproportionalen Itemmengen der jeweiligen Subdimensionen auf die Ausprägungen des Gesamtkonstruktes haben.[5]

Die Analyse der Item- und Skalenqualität hingegen förderte keine auffälligen Ergebnisse zutage. Sowohl die Kennwerte der klassischen Testtheorie als auch die Kennwerte der IRT-Analyse weisen darauf hin, dass es möglich ist, mit den 30 eingesetzten Kriterien eine eindimensionale Skala zur Unterrichtsqualität zu erstellen, die das gesamte Spektrum tatsächlich beobachtbarer Qualitätsausprägungen abdeckt. Die interne Konsistenz der Skala „Unterrichtsqualität“ liegt, trotz eventueller Verzerrungen, hoch. Auch sind beim Einsatz der Skala weder Decken- noch Bodeneffekte zu erwarten, da die 120 Itemausprägungen das gesamte Qualitätsspektrum von −3 bis +3 Standardabweichungen abdecken. Dadurch, dass sich sowohl Items mit mittlerer als auch solche mit hoher Trennschärfe finden, ist darüber hinaus zu erwarten, dass auch im Mittelbereich der Skala verhältnismäßig genau ausdifferenziert werden kann.

Eine weitere zu klärende Frage war, ob ein solches Modell fair zwischen Schulformen differenziert oder ob gegebenenfalls eine systematische Benachteiligung bzw. Bevorzugung von einzelnen Schulformen durch das eingesetzte Instrument zu beobachten ist. Um dies zu prüfen, wurden die Schulformen „reine Grundschulen“ und „Schulen mit Sekundarschulzweig“ voneinander unterschieden und ermittelt, ob differentielle Item-Funktionen für einzelne Merkmale des Beobachtungsbogens nachweisbar sind. Insgesamt konnte hier für sieben Merkmale ein signifikanter Unterschied zugunsten der „reinen Grundschulen“ diagnostiziert werden. Jedoch lagen die ermittelten Kennwerte außerhalb der Größenordnung, mit der auf eine systematische Benachteiligung geschlossen werden kann: Es ist somit anzunehmen, dass keine systematische Benachteiligung von „Schulen mit Sekundarschulzweig“ bei der Bewertung von Unterrichtsqualität erfolgt und beobachtete Unterschiede auf tatsächliche Qualitätsunterschiede an „reinen Grundschulen“ und „Schulen mit Sekundarschulzweig“ zurückzuführen sind.

Last but not least wurde eine Abstufung der metrischen Skala vorgenommen, die eine kriteriale Interpretation von Unterrichtsmerkmalen zulassen soll. Hierzu wurden auf der Skala „Unterrichtsqualität“ mithilfe eines *Proficiency Scaling* vier Schwellen (*Cut Scores*) eingezogen, die die einzelnen Abstufungen voneinander separieren. Die einzelnen Stufen wurden abschließend narrativ beschrieben, sodass eine anschauliche, inhaltliche Interpretation der einzelnen Abstufungen für die praktische Anwendung möglich ist. Das vorgestellte Modell bietet nun die Möglichkeit, zukünftige Befunde zur Unterrichtsqualität der Schulinspektion Hamburg in einem kriterialen Maßstab zu verorten, dessen Grundlage eine standardisierte, regional geeichte Messskala bildet. Schulverantwortliche und Bildungsadministration erhalten so leicht nachvollziehbare Rückmeldungen zur

Qualität von Unterricht, mit der sich absolute Fragestellungen, wie z. B. „An welcher Schule ist dringende Unterstützung bei der Entwicklung von Unterrichtsqualität vonnöten?“, beantworten lassen (vgl. Pietsch et al. in Vorb.). Darüber hinaus zeigt das Modell, wie einzelne Merkmale von Unterrichtsqualität gemeinhin aufeinander aufbauen. Ebenfalls kann ein *Feed Up, Feed Back* und *Feed Forward* in den Rückmeldungen der Schulinspektion Hamburg aufgrund der Stufung im Modell dargestellt werden. Aber auch im Rahmen eines Systemmonitorings ermöglicht es ein solches Modell, einen relevanten Teilaspekt von Bildungsqualität transparent, für Leser leicht nachvollziehbar darzustellen und auf Entwicklungspotenziale in der Unterrichtsgestaltung hinzuweisen (vgl. Diedrich 2009; Institut für Bildungsmonitoring 2009; Pietsch 2009a). Zusammengenommen bietet sich hier zukünftig die Chance, im Lichte der empirischen Befunde Maßnahmen der Schul- und Unterrichtsentwicklungsprozesse ebenso wie Fort- und Weiterbildungsangebote gezielt weiterzuentwickeln (vgl. Pietsch et al. 2009a).

Gleichwohl bleiben viele weitere Forschungsfragen offen. So stellt sich generell die Frage, ob und, falls ja, in welchem Maße die hier berichteten Befunde über Hamburg hinaus generalisierbar sind. Kommen andere Inspektionen mit ihren Instrumenten zu ähnlichen Ergebnissen und lässt sich ein vergleichbares Modell erstellen, sofern bei den Unterrichtsbeobachtungen in den Ländern grundsätzliche empirische Verfahrensstandards eingehalten werden? Oder handelt es sich beim vorgestellten Stufenmodell nur um ein Best-Practice-Modell mit regional begrenzter Gültigkeit, das Hinweise darauf gibt, wie einzelne Teilaspekte effektiven Unterrichts an Hamburger Schulen – aber auch nur dort – aufeinander aufbauen? Hierhinter steht auch die Frage, ob die erstellten Abstufungen für ein Modell von Unterrichtsstandards letztlich der Tatsache geschuldet sind, dass es an Hamburger Schulen in einigen Qualitätsbereichen von Unterricht stärkere Entwicklungspotenziale gibt als in anderen oder ob die vorgeschlagenen Abstufungen inhaltlich repräsentativ sind und somit eine Interpretation mit Blick auf das Konstrukt Unterrichtsqualität im Allgemeinen und nicht nur auf dessen Operationalisierung durch die Schulinspektion Hamburg im Speziellen zulassen. Um etwas hierüber in Erfahrung zu bringen, müsste ein empirischer Vergleich zwischen den Instrumenten der Schulinspektionen in den Ländern stattfinden. Ebenso wichtig ist es, den implizit angenommenen Zusammenhang von Prozess- und Produktmerkmalen empirisch zu untersuchen: Denn ohne eine empirisch nachweisbare Verbindung der wahrgenommenen Qualität von Unterricht auf der einen und tatsächlich erzielten Lernerfolgen auf der anderen Seite bleibt deren Zusammenhang eine ausschließlich auf Wahrscheinlichkeitsannahmen beruhende Unterstellung.

## Anmerkungen

1 Dass diese Annahme auch in der Praxis zutrifft, zeigen deskriptive Datenanalysen. Die Kategorie „nicht beobachtbar“ wird auf Ebene der einzelnen Items in der Regel in weniger als 5 % aller Fälle genutzt. Ausnahmen bilden die Items „Die Lehrkraft geht mit Störungen angemessen und konstruktiv um.“, „Das Erreichen der Lernziele wird angemessen überprüft.“ und „Die Lehrkraft geht mit Schülerfehlern konstruktiv um.“. Bei diesen Items wird die Kategorie „nicht beobachtbar“ in 34, 21 und 16 % aller Fälle durch die Inspektorinnen und Inspektoren genutzt. Aufgrund der hohen Anzahl fehlender Werte wurden diese Items nicht zur inhaltlichen Beschreibung der Abstufungen herangezogen.

2 Items, bei denen die Inspektorinnen und Inspektoren die Kategorie „nicht beobachtbar" gewählt hatten, wurden als fehlende Werte unter der Annahme *Missing at Random* (MAR) behandelt. Im Mittel wurde diese Kategorie pro Sequenz 2,4-mal genutzt (SE: 0,05, SD: 1,65). Wie eine Sensitivitätsanalyse von Pietsch u. Leist (2009) mithilfe von *Latent-Class-Pattern-Mixture-Modellen* (LCPMM) zeigt, führt die Nutzung dieser Kategorie und deren Nicht-Handhabung als *Missing not at Random* (MNAR) in der Datenauswertung zu einer leichten Unterschätzung von Itemmittelwerten. Dies hat jedoch keinen nachweisbaren Effekt auf die Bestimmung zentraler Tendenzen, wie z. B. des Populationsmittelwertes, im IRT-Modell.

3 Da positive Korrelationen zwischen den Itemresiduen beobachtet wurden, ist davon auszugehen, dass die Reliabilität der Gesamtskala „Unterrichtsqualität" überschätzt wird.

4 Die Modellierung eines mehrdimensionalen Modells aus analytischen Gründen im Rahmen von Analysen auf Populations- resp. Stichprobenebene bleibt hiervon natürlich unbenommen. Im Rahmen von Rückmeldungen an Einzelschulen bietet es sich darüber hinaus ggf. an, neben dem Modell Kennwerte auf Ebene einzelner Items für innerschulische Analysezwecke zurückzumelden, um so eine differenzierte und detaillierte Auseinandersetzung mit den Evaluationsbefunden zu ermöglichen.

5 Betrachtet man die MNSQ-Werte der Items in der mehrdimensionalen Skalierung, dann liegen die Kennwerte für alle 30 Items im Bereich von 0,80 bis 1,20. Die beiden in der eindimensionalen Skalierung auffälligen Items 19 und 23 haben dann MNSQ-Werte i. H. v. 1,13 und 0,89. Dieser Befund unterstützt das Vorgehen, diese beiden Items trotz der im Rahmen der eindimensionalen Skalierung berichteten Kennwerte in der Gesamtskala zu belassen, da das eindimensionale Modell als Approximation des mehrdimensionalen Modells gilt (vgl. zu diesem Thema auch Goldstein 2004).

## Literatur

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scores with successive integers. *Applied Psychological Measurement, 2*(4), 581–594.

Bangert-Drowns, R. L., Kulik, C. L., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–237.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191–204.

Behörde für Bildung und Sport. (2006). *Orientierungsrahmen: Qualitätsentwicklung an Hamburger Schulen*. Hamburg: Behörde für Bildung und Sport.

Bonsen, M., Büchter, A., & Peek, R. (2006). Datengestützte Schul- und Unterrichtsentwicklung. Bewertung der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer. In W. Bos, H.-G. Holtappels, R. Pfeiffer & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (Bd. 14, S. 125–148). Weinheim: Juventa.

Böttcher, W., & Kotthoff, H.-G. (2007a). Schulinspektion zwischen Rechenschaftslegung und schulischer Qualitätsentwicklung: internationale Erfahrungen. In W. Böttcher & H.-G. Kotthoff (Hrsg.), *Schulinspektionen: Evaluation, Rechenschaftslegung und Qualitätsentwicklung* (S. 9–20). Münster: Waxmann.

Böttcher, W., & Kotthoff, H.-G. (2007b). Gelingensbedingungen einer qualitätsoptimierenden Schulinspektion. In W. Böttcher & H.-G. Kotthoff (Hrsg.), *Schulinspektionen: Evaluation, Rechenschaftslegung und Qualitätsentwicklung* (S. 223–230). Münster: Waxmann.

Bos, W., Holtappels, H.-G., & Rösner, E. (2006). Schulinspektionen in den deutschen Bundesländern – eine Baustellenbeschreibung. In W. Bos, H.-G. Holtappels, R. Pfeiffer & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (Bd. 14, S. 81–124). Weinheim: Juventa.

Bremerich-Vos, A., & Böhme, K. (2009). Lesekompetenzdiagnostik – die Entwicklung eines standardbasierten Kompetenzmodells für den Bereich Lesen. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 219–249). Weinheim: Beltz.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brophy, J. (2000). *Teaching*. Genf: IBE.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Clauser, B., & Linacre, J. M. (1999). Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions, 13*(2), 696.

Diamantopoulus, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203–1218.

Diedrich, M. (2009). 1. Jahresbericht der Schulinspektion: Trends für die beruflichen Schulen. *Informationen für Hamburger Berufliche Schulen, 19*(2), 10–11.

Ditton, H. (2000). Qualitätskontrolle und -sicherung in Schule und Unterricht. Ein Überblick zum Stand der empirischen Forschung. In A. Helmke, W. Hornstein & E. Terhart (Hrsg.), *Qualitätssicherung im Bildungsbereich* (Zeitschrift für Pädagogik: Beiheft Nr. 41, S. 73–92). Weinheim: Beltz.

Döbert, H., Rürup, M., & Dedering, K. (2008). Externe Evaluation von Schulen in Deutschland – die Konzepte der Bundesländer, ihre Gemeinsamkeiten und Unterschiede. In H. Döbert & K. Dedering (Hrsg.), *Externe Evaluation von Schulen. Historische, rechtliche und vergleichende Aspekte* (S. 63–152). Münster: Waxmann.

Dobbelstein, P. (2008). Qualitätsmaßstäbe in der Diskussion – die Suche nach dem guten Unterricht. In S. Müller, K. Dedering & W. Bos (Hrsg.), *Jahrbuch Schulische Qualitätsanalyse in NRW* (S. 84–92). Neuwied: LinkLuchterhand.

Draba, R. E. (1977). *The identification and interpretation of item bia*s. Chicago: University of Chicago.

Ehren, M. C. M., & Visscher, A. J. (2006). Towards a theory on the impact of school inspections. *British Journal of Educational Studies, 54*(1), 51–72.

Ehren, M. C. M., & Visscher, A. J. (2008). The relationships between school inspections, school characteristics and school improvement. *British Journal of Educational Studies, 56*(2), 205–227.

Ercikan, K. (2006). Examining guidelines for developing accurate proficiency level scores. *Canadian Journal of Education, 29*(3), 823–838.

Ercikan, K., & Julian, M. (2002). Classification accurancy of assigning student performance to proficiency levels. *Applied Measurement in Education, 15*(3), 269–294.

Fend, H. (1998*). Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung*. Weinheim: Juventa.

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education, 10*(2), 123–144.

Fraser, B. J., Walberg, H. J., Welch, W. W., & Hattie, J. A. (1987). Syntheses of educational productivity research. *International Journal of Educational Research, 11*(2), 145–252.

Goldstein, H. (2004). International comparison of student attainment: Some issues arising from the PISA study. *Assessment in Education, 11*(3), 319–330.

Habing, B., Finch, H., & Roberts, J. S. (2005). A Q3 statistic for unfolding item response theory model: Assessment of unidimensionality with two factors and simple structures. *Applied Psychological Measurement, 29*(6), 457–471.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Helmke, A. (2003). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze: Kallmeyer.

Helmke, A. (2006). Was wissen wir über guten Unterricht? Über die Rückbesinnung auf den Unterricht als Kerngeschäft der Schule. *Pädagogik, 2,* 42–45.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts.* Seelze: Klett-Kallmeyer.

Helmke, A., & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Enzyklopädie der Psychologie: Psychologie des Unterrichts und der Schule* (Bd. 3, S. 71–176). Göttingen: Hogrefe.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational and Statistical Statistics, 15*(4), 353–368.

Huynh, H., Michaels, H., & Ferrara, S. (1995). A comparison of three procedures to identify item clusters with local dependence. Paper, präsentiert auf dem National Council on Measurement in Education, San Francisco.

Institut für Bildungsmonitoring. (2009). *Bildungsbericht Hamburg 2009*. Hamburg: Institut für Bildungsmonitoring.

Kiper, H. (2008). Diskurse zur Unterrichtsentwicklung: Eine kritische Betrachtung. In N. Berkemeyer, W. Bos, V. Manitius & K. Müthing (Hrsg.), *Unterrichtsentwicklung in Netzwerken. Konzeptionen, Befunde, Perspektiven* (S. 95–120). Münster: Waxmann.

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive; Kulturspezifische Perspektiven, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In PISA-Konsortium Deutschland. (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 333–359). Opladen: Leske + Budrich.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung. In Bundesministerium für Bildung und Forschung. (Hrsg.), *TIMSS – Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Videodokumente* (S. 43–58). Bonn: BMBF.

Klieme, E., Baumert, J., Köller, O., & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Pflichtschulzeit* (S. 85–134). Opladen: Leske + Budrich.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts ‚Pythagoras'. In M. Prenzel & L. Aloi-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 127–146). Münster: Waxmann.

Klieme, E., Avenarius, H., Blum, W., Döbrich, W., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2007). *Zur Entwicklung nationaler Bildungsstandards – eine Expertise*. Berlin: BMBF.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

KMK. (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4) – Beschluss vom 15.10.2004*. München: LinkLuchterhand.

Köller, O. (2008). Bildungsstandards in Deutschland: Implikation für die Qualitätssicherung und Unterrichtsqualität. In M. A. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (Zeitschrift für Erziehungswissenschaft: Sonderheft 9, S. 47–59). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279–308.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Maritzen, N. (2007). Schulinspektion – ein neues Element der Systemsteuerung. *Journal für Schulentwicklung, 11*(3), 6–14.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Meyer, H. (2004). *Was ist guter Unterricht?* Berlin: Cornelsen.

Meyer, H. (2006). Schulinspektion führt nicht automatisch zu Qualitätssicherung: Interview mit der westfälisch-lippischen Direktorenvereinigung. http://www.westfaelische-direktorenvereinigung.de/PDF/Jahrestagung%202006/Interview%20Schulinspektion.pdf. Zugegriffen: 05. Sep. 2009.

Meyer, H. im Gespräch mit M. A. Meyer (2008). Disput über aktuelle Probleme und Aufgaben der Didaktik. In M. A. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (Zeitschrift für Erziehungswissenschaft: Sonderheft 9, S. 77–86). Wiesbaden: VS Verlag für Sozialwissenschaften.

Meyer, H., & Klapper, A. (2006). Unterrichtsstandards für ein kompetenzorientiertes Lernen und Lehren. In R. Hinz & B. Schumacher (Hrsg.), *Auf den Anfang kommt es an: Kompetenzen entwickeln – Kompetenzen stärken* (S. 89–108). Wiesbaden: VS Verlag für Sozialwissenschaften.

Oelkers, J., & Reusser, K. (2008). *Qualität entwickeln, Standards sichern, mit Differenzen umgehen*. Berlin: BMBF.

Pietsch, M. (2009a). Die Qualität des Unterrichts an Hamburger Schulen aus Beobachterperspektive. In Institut für Bildungsmonitoring. (Hrsg.), *Jahresbericht der Schulinspektion Hamburg 2008* (S. 44–62). Hamburg: Behörde für Schule und Berufsbildung.

Pietsch, M. (2009b). Unterrichtsbeobachtungen & Co.: Die externe Evaluation hinterlässt einen Datenberg. Was steckt dahinter und wie können Sie damit arbeiten? In M. Bonsen, W. Hohmeier, & M. Reese (Hrsg.), *Handbuch Unterrichtsqualität sichern – Sekundarstufe* (Loseblattsammlung). Berlin: Raabe.

Pietsch, M., & Leist, S. (2009). The impact of „not observable" response options on the results of classroom observations: An application of Latent Class Pattern Mixture Models to outcomes that are potentially missing not at random. Paper präsentiert auf der 13. Biennale der European Association for Research on Learning and Instruction (EARLI), Amsterdam.

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

Pietsch, M., Bonsen, M., & Bos, W. (2007). Ein Index sozialer Belastung als Grundlage für die Rückmeldung ‚fairer Vergleiche' von Grundschulen in Hamburg. In W. Bos & M. Pietsch (Hrsg.), *KESS 4 – Kompetenzen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 an Hamburger Grundschulen* (S. 225–246). Münster: Waxmann.

Pietsch, M., Schnack, J., & Schulze, P. (2009a). Unterricht zielgerichtet entwickeln: Die Schulinspektion Hamburg entwickelt ein Stufenmodell für die Qualität von Unterricht. *Pädagogik, 2,* 38–43.

Pietsch, M., Böhme, K., Robitzsch, A., & Stubbe, T. C. (2009b). Das Stufenmodell zur Lesekompetenz der länderübergreifenden Bildungsstandards im Vergleich zu IGLU 2006. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 393–428). Weinheim: Beltz.

Pietsch, M., Schnack, J., Schulze, P., & Krause, M. (in Vorb.). Elaborierte Rückmeldungen zur Qualität von Unterricht: Über empirisch abgesicherte Bezugsnormen für die Weiterentwicklung von Schule und Unterricht. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektionen in Deutschland – eine erste empirische Zwischenbilanz.* Münster: Waxmann.

Ravitch, D. (1995). *National standards in American education: A citizen's guid*e. Washington: Brookings Institution Press.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193–203.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 42–107). Weinheim: Beltz.

Rolff, H.-G. (2007). *Studien zu einer Theorie der Schulentwicklung*. Weinheim: Beltz.

Rost, J. (2004). *Lehrbuch Testtheorie/Testkonstruktion*. Bern: Huber.

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London: OFSTED.

Scherens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon Press.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Seidel, T. (2008). Schuleffektivitätskriterien in der internationalen empirischen Forschung. *Zeitschrift für Erziehungswissenschaft, 11,* 348–367.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Psychometrika, 28*(3), 237–247.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*(1), 66–78.

Stone, M. H. , Wright, B. D., & Stenner, A. J. (1999). Mapping variables. *Journal of Outcome Measurement, 3*(4), 308–322.

Stralla, M. (2009). *Die Unterrichtsbeobachtungen im Rahmen der deutschen Schulinspektion. Analyse des Kerninstruments zur Beurteilung der Schulqualität* (Unveröffentlichte Diplomarbeit). Berlin: Freie Universität Berlin.

Visscher, A. J., & Coe, R. (2002). *School improvement through performance feedback*. Lisse: Swets & Zellinger.

Visscher, A. J., & Coe, R. (2003). School performance feedback systems. Conceptualisation, analysis and reflection. *School Effectiveness and School Improvement, 14*(3), 321–349.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classification: multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*(3), 255–275.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions, 16*(3), 888.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest. Generalised item response modelling software.* Melbourne: ACER Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145.

Yen, W. M. (1993). Scaling performance assessments. Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15–25.