

Unterrichtsentwicklung durch Unterrichtsfeedback – Erste Entwicklungsschritte eines Beobachtungsbogens für Schulleitungen

Holger Gärtner  · Felicitas Thiel · Max Nachbauer  · Christopher Kellermann 

Eingegangen: 2. Februar 2021 / Überarbeitet: 21. Juni 2021 / Angenommen: 24. August 2021 / Online publiziert: 20. September 2021
© Der/die Autor(en) 2021

Zusammenfassung Im vorliegenden Beitrag werden erste Entwicklungsschritte eines Beobachtungsbogen zur Erfassung von Unterrichtsqualität vorgestellt, der bei Unterrichtsevaluationen durch Schulleitungen eingesetzt werden soll, die mit dem Ziel eines entwicklungsbezogenen Feedbacks an Lehrpersonen erfolgen. Im ersten Teil wird die theoretische Grundlage des Beobachtungsbogens beschrieben. Unter Rückgriff auf lehr-lernpsychologische Theorien der Informationsverarbeitung, der Motivation und Theorien der sozialen Interaktion im Klassenzimmer wurden drei übergeordnete Anforderungsbereiche des Unterrichts (Unterstützung des Wissenserwerbs, Motivierung und Klassenmanagement) identifiziert und Unterrichtsmerkmale aus den entsprechenden Bezugstheorien hergeleitet. Zusätzlich wurde Individualisierung als vierte, querliegende Anforderung berücksichtigt. Im zweiten Teil erfolgt im Rahmen einer explorativen Studie eine erste empirische Überprüfung zentraler psychometrischer Eigenschaften des Beobachtungsbogens. Das Ergebnis

Die Autoren H. Gärtner und F. Thiel teilen sich die Erstautorenschaft.

Prof. Dr. Holger Gärtner (✉)

Fachbereich Erziehungswissenschaft und Psychologie, Arbeitsbereich Unterrichts- und Schulevaluation, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Deutschland
E-Mail: holger.gaertner@fu-berlin.de

Prof. Dr. Felicitas Thiel · Dipl.-Päd. Max Nachbauer · Christopher Kellermann, M.A.

Fachbereich Erziehungswissenschaft und Psychologie, Arbeitsbereich Schulpädagogik/Schulentwicklungsforschung, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Deutschland

Prof. Dr. Felicitas Thiel

E-Mail: felicitas.thiel@fu-berlin.de

Dipl.-Päd. Max Nachbauer

E-Mail: max.nachbauer@fu-berlin.de

Christopher Kellermann, M.A.

E-Mail: christopher.kellermann@fu-berlin.de

einer explorativen Faktorenanalyse zeigt eine weitestgehend theoriekonforme Faktorenstruktur. Die Unterrichtsmerkmale weisen moderate bis hohe Korrelationen, zufriedenstellende interne Konsistenzen und Stabilitäten auf. Die Ergebnisse sprechen dafür, dass das theoretische Modell empirisch abgebildet werden kann. Weitere Entwicklungsschritte des Beobachtungsbogens werden skizziert.

Schlüsselwörter Unterrichtsevaluation · Unterrichtsentwicklung · Beobachtungsbogen · Feedback · Schulleitung

Instructional development through feedback—first development steps of an observation sheet for school management

Abstract The present paper presents first development steps for an observation sheet developed for teacher evaluations by school leaders aiming to provide formative feedback to teachers. The first part of the paper describes the theoretical foundation of the observation sheet. With reference to cognitive learning theories, motivation theories and social interaction theories, three crucial requirements of teaching (supporting knowledge acquisition, motivating and classroom management) were identified. For each requirement, various teaching characteristics were derived from the corresponding theories. In addition, individualization was considered as a forth, transverse teaching requirement. In the second part, the psychometric properties of the observation sheet were examined in an exploratory study. Results of an exploratory factor analyses show a factor structure largely consistent with theory. Correlations between the teaching characteristics were moderate to high. Furthermore, internal consistency and stabilities were satisfying. In summary, the results indicate that the theoretical model can be empirically represented. Further development steps of the observation sheet are outlined.

Keywords Feedback · Observation sheet · School improvement · School leader · Teacher evaluation

1 Unterrichtsevaluation durch Schulleitungen und ihre Bedeutung für Unterrichtsentwicklung

Zahlreiche Studien belegen, dass Unterrichtsevaluationen für die Entwicklung des Unterrichts grundsätzlich einen hohen Stellenwert haben. In den letzten Jahren wurden im deutschsprachigen Raum dabei vor allem die Nutzung von Schülerleistungsdaten oder Schülerbefragungen implementiert (Gärtner 2010; Gärtner und Wurster 2019; Göllner et al. 2016). Anders als in angelsächsischen Ländern (Darling-Hammond et al. 1983) spielt die Evaluation des Unterrichts durch Schulleitungen in der deutschen Schulentwicklungsdiskussion bislang nur eine untergeordnete Rolle. Zwar schreiben alle Beamten-gesetze dienstliche Beurteilungen und damit eine regelmäßige oder anlassbezogene Evaluation des Unterrichts durch Schulleitungen vor (Meetz 2007), in der Praxis scheinen dienstliche Beurteilungen unter Lehrpersonen allerdings wenig akzeptiert und kaum mit einem entwicklungsbezogenen Unter-

richtsfeedback verbunden zu sein (Rauscher 2000). Obwohl empirische Befunde die Wirksamkeit von Führungsstilen wie *Instructional Leadership* oder *Leadership for Learning* bestätigen (Hallinger 2011; Pietsch et al. 2019), kann von einer systematischen Implementation eines Unterrichtsfeedbacks durch Schulleitungen allenfalls in Ansätzen gesprochen werden.

Grundsätzlich erscheinen Unterrichtsbeobachtungen als besonders aussichtsreicher Ansatz der Unterrichtsentwicklung. Für Unterrichtsbeobachtungen sprechen vor allem das Eröffnen einer neutralen Außenperspektive (Praetorius 2013), die hohe Augenschein-Validität und dadurch hohe Akzeptanz durch Lehrpersonen (Cohen und Goldhaber 2016) sowie die prädiktive Validität bezüglich des Lernerfolgs von Schüler/-innen (Fauth et al. 2014). Trotz gewisser methodischer Schwierigkeiten (Praetorius 2013), haben Unterrichtsbeobachtungen ein großes Potenzial für Unterrichtsentwicklung (Kraft et al. 2018; Steinberg und Sartain 2015; Taylor und Tyler 2012). Sollen Unterrichtsbeobachtungen durch Schulleitungen systematisch implementiert werden, müssen geeignete Instrumente vorliegen, die eine valide Beurteilung des Unterrichts erlauben und einen Rahmen für ein entwicklungsbezogenes Feedbackgespräch abstecken, in dem konkrete Hinweise auf die Optimierung von Unterricht thematisiert werden können (Kellermann et al. [under review](#); Heneman und Milanowski 2003). Da Schulleitungen nur in wenigen Fächern über fachliche Expertise verfügen, muss sich ein entsprechendes Instrument auf generische, fächerübergreifende Merkmale des Unterrichts beziehen. Aus diesen Gründen sollte eine Beschränkung auf zentrale lernwirksame Merkmale des Unterrichts und eine Zuordnung dieser Merkmale zu theoretisch trennscharfen übergeordneten Dimensionen erfolgen.

Der in diesem Beitrag vorgestellte Beobachtungsbogen zur Erfassung von Unterrichtsqualität soll vorrangig im Rahmen von Unterrichtsevaluationen durch Schulleitungen eingesetzt werden, deren Ziel ein entwicklungsbezogenes Feedback an Lehrpersonen ist. Im Zentrum steht dabei das Handeln der Lehrperson; im Unterschied zu anderen Instrumenten der Unterrichtsevaluation, die die Interaktion zwischen Lehrkräften und Schüler/-innen in den Mittelpunkt stellen (Praetorius et al. 2018). Die Entwicklung des Beobachtungsbogens schließt an das Konzept der drei Basisdimensionen des Unterrichts an (Klieme et al. 2001). Es liegen bereits mehrere Beobachtungsbögen vor, die auf diesem Modell von Unterrichtsqualität basieren (Praetorius et al. 2018). Ausgangspunkt des Modells war eine empirische, faktoranalytische Ermittlung von Unterrichtsdimensionen, welche nachträglich theoretisch unterfüttert wurden. Durch das Zusammenführen mehrerer Unterrichtsmerkmale in wenige, inhaltlich umfassende Dimensionen genügt das Modell forschungsrelevanten Anforderungen wie Sparsamkeit, Vermeidung von Multikollinearität oder hoher Erklärkraft für den Lernerfolg von Schüler/-innen. Die drei Dimensionen wurden allerdings in verschiedenen Studien teilweise unterschiedlich operationalisiert (Praetorius et al. 2018). Zudem sind mehrere empirisch abgesicherte Merkmale wirksamen Unterrichts wie Klarheit, Strukturiertheit und Übung nicht im Modell enthalten (Praetorius et al. 2018).

Für die Entwicklung eines Beobachtungsbogens für Schulleitungen wurde das Modell der drei Basisdimensionen erweitert. Der Beobachtungsbogen sollte einerseits das Unterrichtsgeschehen fachübergreifend möglichst breit abdecken, ander-

seits sollten die Items aus erhebungswirtschaftlichen Gründen auf zentrale lernwirksame Merkmale des Unterrichts beschränkt werden. Die Operationalisierung mittels des Beobachtungsbogens weicht aus diesem Grund an einigen Stellen von der ursprünglichen Konzeption der drei Basisdimensionen ab: Die Dimension der kognitiven Aktivierung, die als Anregung „zum vertieften Nachdenken und zu einer elaborierten Auseinandersetzung mit dem Unterrichtsgegenstand“ (Lipowsky 2009, S. 93) definiert ist und Aspekte des Verstehens sowie des Transfers fokussiert, wird breiter operationalisiert und um die Aspekte der strukturierten Präsentation von Wissen und das Üben ergänzt. Anstelle der Dimension der konstruktiven Unterstützung wurde eine Dimension „Motivierung“ operationalisiert, die, nicht aus der Selbstbestimmungstheorie der Motivation von Deci und Ryan (1993), sondern aus dem Erwartungs-mal-Wert-Modell der Motivation abgeleitet ist, das den Nutzenaspekt einbezieht. Die Dimension Klassenführung wurde als Klassenmanagement vor dem Hintergrund der Arbeiten von Ophardt und Thiel (2013) sowie Piwowar (2013)

Tab. 1 Unterrichtsmerkmale nach Anforderungsbereichen und Beispiel-Items

Unterrichtsmerkmal	Beispiel-Item
<i>Unterstützung des Wissenserwerbs</i>	
Erklärung und Präsentation	Erklärungen der Lehrperson sind gut verständlich
Unterstützung des Verstehens	Die Fragen der Lehrperson regen schlussfolgerndes Denken an (z. B. „Wie erfolgt ...?“, „Warum geschieht ...?“)
Speicherung durch Üben	Übungsaufgaben werden schriftlich oder mündlich korrigiert
Anwendung und Transfer	Der Transfer des erlernten Wissens wird durch komplexe Anwendungsaufgaben gezielt angeregt
Individualisierter Wissenserwerb	Für Schülerinnen und Schüler unterschiedlicher Niveaustufen wird unterschiedliches Übungsmaterial eingesetzt
<i>Motivierung</i>	
Weckung von Interesse	Die Lehrperson demonstriert Interesse für den Lerngegenstand
Förderung von Freude	Explorierende oder produktive Tätigkeiten (z. B. Knobeln, Testen, Entwerfen ...) sind ein wichtiger Bestandteil des Unterrichts
Verdeutlichung des Nutzens	Die Lehrperson macht deutlich, welchen Nutzen das erworbene Wissen für das spätere Leben hat
Förderung von Selbstwirksamkeit	Die Lehrperson signalisiert den Schülerinnen und Schülern deutlich, dass sie ihnen etwas zutraut
Individualisierte Motivierung	Die Lehrperson geht auf die unterschiedlichen Interessen der Schülerinnen und Schüler ein
<i>Klassenmanagement</i>	
Lerndienliches Unterrichtsklima	Die Schülerinnen und Schüler gehen respektvoll miteinander um
Monitoring	Die Lehrperson bekommt mit, was die Schülerinnen und Schüler im Unterricht machen
Steuerung des Unterrichtsflusses	Die Lehrperson schweift nicht vom Thema ab
Gruppenaktivierung	Auch für einzelne Schülerinnen und Schüler entsteht während der Stunde kein Leerlauf
Umgang mit Störungen	Kleinere Störungen werden beiläufig behoben

operationalisiert. Entsprechend wurden auch der Aspekt der Steuerung des Unterrichtsflusses und das Unterrichtsklima einbezogen. Außerdem wurde ergänzend eine weitere Dimension „Individualisierung“ eingeführt.

Im Folgenden wird zunächst die theoretische Grundlage und damit die Herleitung der dem Beobachtungsbogen zugrunde gelegten Indikatoren beschrieben, bevor anschließend im Rahmen einer explorativen Studie eine erste empirische Überprüfung zentraler psychometrischer Eigenschaften des Beobachtungsbogens erfolgt.

2 Theoretische Grundlage des Beobachtungsbogens

Die theoretische Grundlage des Beobachtungsbogens bildet eine Beschreibung übergeordneter Anforderungsbereiche des Unterrichtens in der Perspektive von lehr-lernpsychologischen Theorien der Informationsverarbeitung, der Motivation und der sozialen Interaktion im Klassenzimmer: a) den Wissenserwerb der Schüler/-innen zu unterstützen, b) Schüler/-innen zu motivieren und c) das Verhalten der Schüler/-innen so zu regulieren, dass möglichst alle Schüler/-innen möglichst viel Lernzeit aktiv nutzen können. Zusätzlich ist eine angemessene Individualisierung des Unterrichts erforderlich, welche hier als querliegende Anforderung verstanden wird. Die linke Seite von Tab. 1 gibt einen Überblick über die Anforderungsbereiche und Unterrichtsmerkmale des Beobachtungsbogens.

2.1 Unterstützung des Wissenserwerbs

Die theoretische Grundlage für den Anforderungsbereich Unterstützung des Wissenserwerbs bildet das Lehr-Lern-Prozessmodell von Klauer und Leutner (2012). Das Modell unterscheidet vier Komponenten eines kognitiven Lernprozesses: 1) Informierung, 2) Informationsverarbeitung, 3) Speicherung und Abruf von Informationen sowie 4) Transfer.

Das Unterrichtsmerkmal *Erklärung und Präsentation* bezieht sich auf die Darbietung von Informationen. Die Herausforderung besteht darin, dass im sensorischen Register viele Reize aufgenommen, aber im Arbeitsgedächtnis nur wenige Reize gleichzeitig weiterverarbeitet werden können. Informationen sollten daher so aufbereitet werden, dass die kognitive Belastung des Arbeitsgedächtnisses möglichst gering ist (Sweller 1988). Eine gute Informationsdarbietung ist durch eine strukturierte Präsentation, verständliche Erklärungen, Veranschaulichungen durch Beispiele und das Aufgreifen des Vorwissens der Schüler/-innen gekennzeichnet.

Im Anschluss erfolgt die Informationsverarbeitung. Lehrpersonen sollten in dieser Phase die *Verstehensprozesse der Schüler/-innen unterstützen*. Das betrifft die Elaboration von Informationen und ihre anschließende Reduktion (Klauer und Leutner 2012). Elaborieren umfasst Prozesse wie die Klärung von Zusammenhängen, die Kontextualisierung von Informationen oder die Entfaltung unterschiedlicher Detailspekte. Angeregt werden können sie durch Vergleiche und kognitiv aktivierende Fragen, die auf Voraussetzungen, Bedingungen und Folgen abzielen. Reduktive Prozesse haben die Funktion, Informationen zu komprimieren. Lehrpersonen können diese anregen, indem sie behandelte Inhalte noch einmal zusammenfassen. Neben

elaborativen und reduktiven Strategien ist der Einsatz von metakognitiven Strategien hilfreich für Schüler/-innen (Friedrich und Mandl 1992). Lehrpersonen können den Einsatz metakognitiver Strategien dadurch unterstützen, dass sie die Schüler/-innen dazu auffordern, ihre Vorgehensweisen zu planen und ihre Lösungen selbst zu bewerten.

Die langfristige Verfügbarkeit von Wissen ist von einer nachhaltigen *Speicherung im Langzeitgedächtnis durch ausreichende Übung* abhängig. Durch Übung wird Wissen verinnerlicht und Kompetenzen werden automatisiert, sodass ein problemloser Abruf gelingt. Neben einer ausreichenden Häufigkeit von Übungsphasen ist relevant, wie die Übungen konkret ausgestaltet werden. Wichtig ist, dass besonders schwierige Aspekte intensiv geübt werden und dass die Resultate des Übens überprüft und gegebenenfalls korrigiert werden (Anderson 1982).

Die vierte kognitive Funktion des Lernens ist der *Transfer*. Hier geht es darum, erworbenes Wissen so zu flexibilisieren, dass es auf neue Probleme übertragen werden kann, bzw. darum, gelernte Prinzipien oder Strategien in anderen Zusammenhängen anzuwenden (Klauer und Leutner 2012). Transfer kann angeregt werden, indem die Schüler/-innen ihr theoretisches Wissen auf praktische Probleme anwenden oder indem die Schüler/-innen zu einem Thema mehrere Aufgaben bearbeiten, die den Kontext variieren.

2.2 Motivierung

Neben den kognitiven Funktionen sind die motivationalen Funktionen für das Lernen zentral. Motivation sorgt dafür, dass Lernaktivitäten begonnen und aufrechterhalten werden. Grundlage für die Identifikation von Indikatoren der Motivierungsqualität ist das in der Motivationspsychologie weit verbreitete Modell von Eccles und Wigfield (2002; siehe auch: Rheinberg 2008), das Erwartungs- und Wertkomponenten der Motivation unterscheidet. Vier Aspekte der Motivierung, die für den Unterricht eine zentrale Bedeutung haben, lassen sich hieraus ableiten: durch 1) das Wecken von Interesse, 2) die Förderung von Lernfreude und 3) die Verdeutlichung des Nutzens können Anreize für Lernaktivitäten geschaffen werden, während 4) die Stärkung der Selbstwirksamkeitserwartung eine günstige Einschätzung der eigenen Erfolgswahrscheinlichkeit beim Lernen fördert.

Das *Wecken von Interesse* ist ein wichtiger Ansatzpunkt für die Motivierung der Schüler/-innen. Interesse bezieht sich auf die jeweiligen Lerninhalte und stellt damit einen gegenstandsbezogenen Anreiz dar. Es setzt sich aus wertbezogenen Überzeugungen und Gefühlsüberzeugungen zusammen (Schiefele 2009). Lehrpersonen können Interesse wecken, indem sie Schüler/-innen mit neuen Einsichten überraschen oder indem sie im Unterricht mit Enthusiasmus auftreten und so selbst Interesse an den Lerninhalten demonstrieren.

Im Gegensatz zum Interesse als gegenstandsbezogenem Anreiz bezieht sich *Lernfreude* auf die Art der Lernaktivitäten und wird daher auch als tätigkeitsbezogener Anreiz bezeichnet (Rheinberg 2008). Lehrpersonen sollten demnach darauf achten, dass Schüler/-innen im Unterricht Freude und Spaß erleben. Dazu ermöglichen sie beispielsweise explorierende, gestalterische oder spielerische Aktivitäten.

Ein weiterer Anreiz bezieht sich darauf, welchen *Nutzen* die Lerninhalte für die Schüler/-innen haben (Hulleman und Harackiewicz 2009). Der Anreiz des Nutzens hat, anders als Interesse und Freude, keinen unmittelbaren Bezug zum Lernprozess selbst, sondern verweist auf günstige Wirkungen im Anschluss an das erfolgreiche Lernen. Für diese Motivierungsstrategie ist es entscheidend, dass die Lehrperson verdeutlicht, warum bestimmte Lerninhalte für die Schüler/-innen relevant sind und dass sie Beispiele gibt, in welchen Lebenssituationen bestimmte Kenntnisse oder Kompetenzen benötigt werden.

Anreize alleine motivieren noch nicht zum Lernen. Die Schüler/-innen müssen auch die Erwartung haben, das Lernziel mit ihren eigenen Kompetenzen erreichen zu können. Aus diesem Grund stellt die *Förderung von Selbstwirksamkeitserwartungen* (Bandura 1976) einen wichtigen Ansatzpunkt für Motivierung dar. Relevant ist hierbei, dass die Lehrperson den Schüler/-innen zeigt, dass sie ihnen etwas zutraut, dass sie spezifisch und aufgabenbezogen lobt und Aufgaben stellt, die zwar Herausforderungen darstellen, deren Lösung aber möglich ist. Bei Schwierigkeiten sollte die Lehrperson die Schüler/-innen ermutigen, aber die Lösung nicht vorwegnehmen, so dass die Schüler/-innen Erfolgserfahrungen machen können.

2.3 Klassenmanagement

Schulklassen sind dynamische Interaktionssysteme, in denen Lernen nur dann gelingt, wenn ein gewisses Maß an sozialer Ordnung sichergestellt ist (Doyle 2006). Diese sozialen Faktoren des Lernens werden im Konstrukt des Klassenmanagements konzeptualisiert. Im Vordergrund steht die Maximierung der aktiven Lernzeit für alle Schüler/-innen. Ausgehend von den Arbeiten von Ophardt und Thiel (2013) sowie Piwowar (2013) können fünf Funktionen im Hinblick auf das Klassenmanagement unterschieden werden, die Überschneidungen mit dem Linzer Bogen zur Klassenführung aufweisen (Lenke und Mayr 2015): 1) Unterrichtsklima, 2) Monitoring, 3) Steuerung des Unterrichtsflusses, 4) Gruppenaktivierung und 5) Umgang mit Störungen.

Für ein *lernförderliches Unterrichtsklima* muss zunächst eine Interaktionsordnung durch die Einführung von Regeln etabliert werden. Die Beachtung von Regeln durch die Schüler/-innen in der Form von gegenseitigem Respekt, Rücksichtnahme und die Bereitschaft zur Unterstützung findet ebenso Ausdruck im Unterrichtsklima wie entsprechende Verhaltensweisen der Lehrperson (Ophardt und Thiel 2013).

Ein effektives Klassenmanagement setzt voraus, dass die Lehrperson das Geschehen in der ganzen Klasse wahrnimmt (Kounin 2006). *Monitoring* ist ein Teilaspekt der professionellen Wahrnehmung. Damit ist gemeint, dass die Lehrperson die Aktivitäten aller Schüler/-innen im Blick behält und bemerkt, wenn sich einzelne Schüler/-innen mit unterrichtsfremden Dingen beschäftigen. Dadurch bemerkt die Lehrperson sofort, wenn sich einzelne Schüler/-innen mit unterrichtsfremden Dingen beschäftigen.

Für die Maximierung aktiver Lernzeit ist weiterhin eine effektive *Steuerung des Unterrichtsflusses* wichtig, einerseits, weil bei unnötigen Pausen Lernzeit verloren geht, andererseits, weil das Handlungsprogramm durch Leerläufe oder Nachfragen unterbrochen wird und dann mühevoll wiederaufgebaut werden muss (Ophardt und

Thiel 2013). Für die Aufrechterhaltung des Unterrichtsflusses sind nahtlose Übergänge zwischen Unterrichtsphasen, gut vorbereitete Unterrichtsmaterialien und die Vermeidung von inhaltlichen Abschweifungen relevant.

Auf die Bedeutung der *Gruppenaktivierung* hat bereits Kounin (2006) mit seiner Videostudie aufmerksam gemacht. Vorrangig geht es darum, alle Schüler/-innen aktiv in den Unterricht einzubinden. Lehrpersonen müssen gezielt verhindern, dass sich einzelne Schüler/-innen aus dem Unterrichtsgeschehen zurückziehen oder in Leerlaufphasen verweilen. Hierfür müssen Lehrpersonen einen Gruppenfokus auf die gesamte Klasse und einen Individualfokus auf einzelne Schüler/-innen verbinden.

Störungen des Unterrichts entstehen, weil Schüler/-innen die Aufmerksamkeit nicht über eine ganze Stunde aufrechterhalten können, weil sie von ihren Mitschüler/-innen abgelenkt werden, diese beeindruckt werden wollen oder weil sie sich ungerecht behandelt fühlen (Thiel 2016). Ein angemessener Umgang mit Störungen berücksichtigt die Schwere des Störverhaltens. Bei kleineren Störungen sind beiläufige nonverbale Signale häufig ausreichend. Bei schwereren Störungen können dagegen verbale Ermahnungen und gegebenenfalls auch die Ankündigung von Sanktionen notwendig sein (Ophardt und Thiel 2013).

2.4 Individualisierung

Da sich Schüler/-innen hinsichtlich ihrer Lernvoraussetzungen unterscheiden, stehen Lehrpersonen schließlich vor der Herausforderung der *Individualisierung* des Unterrichts (Klieme und Warwas 2011). Individualisierung wird hier als querliegende Anforderung verstanden. Individualisierende Maßnahmen sind primär in den Bereichen Unterstützung des Wissenserwerbs und Motivierung erforderlich, da die Schüler/-innen sowohl unterschiedliche kognitive Lernvoraussetzungen (z. B. Vorwissen) als auch unterschiedliche motivationale Bedingungen (z. B. Interessen) aufweisen. Dementsprechend werden die zwei Aspekte individualisierter Wissenserwerb und individualisierte Motivierung unterschieden.

Die Überschneidungen des hier vorgestellten Ansatzes zum Modell der drei Basisdimensionen sind offensichtlich. Unterschiede ergeben sich vor allem durch eine stärkere Berücksichtigung von Prozessen der strukturierten Darbietung von Information und der Übung bei der Informationsverarbeitung sowie, bezogen auf den Anforderungsbereich der Motivierung, durch eine Unterscheidung von gegenstandsbezogener, prozessbezogener und utilitaristischer Motivation. Beim Klassenmanagement erfolgte ergänzend zur Operationalisierung dieser Dimension in den Instrumenten, die die drei Basisdimensionen erfassen, die Berücksichtigung von Aspekten der Steuerung des Unterrichtsflusses, der Gruppenaktivierung und des Unterrichtsklimas (vgl. Praetorius et al. 2018).

Für die Entwicklung eines Beobachtungsbogens für die Unterrichtsevaluierung durch Schulleitungen hat das Vorgehen mehrere Vorteile. Erstens ermöglicht die theoretische Herleitung eine eindeutige Zuordnung der einzelnen Unterrichtsmerkmale zu distinkten Dimensionen. Zweitens legt das Unterrichtsmodell seinen Schwerpunkt auf die Handlungen von Lehrpersonen und auf beeinflussbare Aspekte der Lernaktivitäten von Schüler/-innen (z. B. Lernfreude, Störverhalten), um konkrete Ansatzpunkte für eine Weiterentwicklung des Unterrichts aufzuzeigen.

Drittens weist das Modell einen relativ hohen Auflösungsgrad auf, da innerhalb jedes Anforderungsbereichs mehrere inhaltlich eng gefasste Unterrichtsmerkmale benannt werden. Auf der Grundlage der insgesamt 15 Unterrichtsmerkmale kann ein spezifisches und differenziertes Feedback gegeben werden.

2.5 Fragestellungen

Nachdem die theoretische Grundlage und die Herleitung der dem Beobachtungsbogen zugrunde gelegten Indikatoren beschrieben wurden, soll im Folgenden eine erste Überprüfung zentraler psychometrischer Eigenschaften des Beobachtungsbogens erfolgen. Im Rahmen einer explorativen Studie werden Aspekte der Reliabilität und Validität des entwickelten Bogens überprüft. Zunächst wird die faktorielle Validität im Vergleich zu den theoretischen Annahmen getestet. Als Aspekte der Reliabilität werden anschließend die internen Konsistenzen der gebildeten Skalen, deren Interkorrelationen sowie deren Stabilität im Sinne einer Test-Retest-Reliabilität überprüft.

3 Empirische Analysen zum Beobachtungsbogen

Das beschriebene Modell von Unterrichtsqualität wurde in der Form eines Beobachtungsbogens für Schulleitungen operationalisiert. Nachfolgend wird zunächst die Entwicklung des Beobachtungsbogens beschrieben. Anschließend werden die Ergebnisse einer explorativen Studie mit Beobachtungsdaten von Lehramtsstudierenden berichtet.

3.1 Entwicklung des Beobachtungsbogens

Bei der Entwicklung des Beobachtungsbogens wurde zunächst ein Itempool für die insgesamt 15 Konstrukte generiert. Als Vorlage dienten, wo dies unter inhaltlichen Gesichtspunkten sinnvoll erschien, Items aus bestehenden Beobachtungsbögen (z. B. Clausen 2002; Helmke und Schrader 1997; Piwowar 2013). Das Antwortformat ist eine vierstufige Likert-Skala. Basierend auf dem Itempool wurde eine vorläufige Version des Beobachtungsbogens erstellt und in einer Pilotstudie im Sommersemester 2016 primär an Schulen in Berlin überprüft. Basierend auf diesen Analysen wurde der Beobachtungsbogen optimiert. Die überarbeitete Version des Beobachtungsbogens umfasst insgesamt 15 Skalen und 78 Items, mit 4 bis 8 Items pro Skala¹ (Beispielitems siehe Tab. 1).

3.2 Beobachtertraining

Die hier analysierten Beobachtungsdaten wurden von geschulten Lehramtsstudierenden erhoben. Das Beobachtertraining erfolgte im Rahmen einer universitären

¹ Der vollständige Beobachtungsbogen kann auf folgender Internetseite heruntergeladen werden: <https://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/schulentwicklungsforschung/forschung/abge laufenprojekte/DFG/index.html>.

Lehrveranstaltung und dauerte insgesamt 10h. Neben einer theoretischen Einführung in die theoretischen Grundlagen des Beobachtungsbogens und der Sensibilisierung für Beobachtungsfehler wurde der Einsatz des Beobachtungsbogens in mehreren Durchgängen an Videomaterial geübt. Zunächst wurden Videoausschnitte aus unterschiedlichen Unterrichtsstunden von Studierenden unabhängig voneinander beurteilt. Anschließend wurden Übereinstimmungen und Abweichungen der individuellen Einschätzungen anhand von Profildiagrammen in Gruppen von bis zu vier Studierenden diskutiert. Im Sinne einer Kalibrierung wurden die Items in der Diskussion konkretisiert. Abschließend erfolgte eine Diskussion der Gruppenergebnisse im gesamten Seminar. Die gemeinsam erarbeiteten Anhaltspunkte zur Konkretisierung der Items wurden zusammengefasst. Der Umfang des Beobachtertrainings ist zwar erkennbar niedriger als bei Beobachtertrainings im Rahmen wissenschaftlicher Studien, allerdings waren die Studierenden bereits durch eine Vorlesung mit den entsprechenden Theorieansätzen der Informationsverarbeitung, der Motivierung und des Klassenmanagements vertraut.

3.3 Stichprobe

Die Beobachtungsdaten wurden von den Lehramtsstudierenden im Rahmen eines Schulpraktikums im Sommersemester 2017 erhoben. Es handelt sich um Unterrichtsstunden, in denen die Lehramtsstudierenden hospitiert haben, d.h. die Lehramtsstudierenden beobachteten den Unterricht von regulären Lehrpersonen. Die Stichprobe besteht aus insgesamt $N=620$ beobachteten Unterrichtsstunden. In jeder Unterrichtsstunde war ein Beobachter/eine Beobachterin anwesend. Das Beobachtungsverfahren hatte den folgenden Ablauf: Zunächst beobachtete der Beobachter/die Beobachterin den Unterricht und machte sich stichpunktartige Notizen zu relevanten Verhaltensweisen der Lehrperson und der Schüler/-innen. Direkt nach der Unterrichtsstunde füllte der Beobachter/die Beobachterin den Beobachtungsbogen aus. Die Beobachter/-innen (65,7 % weiblich, 34,3 % männlich) wurden instruiert, jeweils vier Unterrichtsstunden derselben Lehrperson (67,9 % weiblich, 32,1 % männlich) einzuschätzen. Der überwiegende Teil der Beobachtungen (91,0 %) erfolgte an Schulen in Berlin. Die Beobachtungen verteilen sich folgendermaßen auf verschiedene Schulformen: 43,1 % Grundschulen, 29,8 % Gymnasien und 27,1 % Schulen mit mehreren Bildungsgängen. Hinsichtlich der beobachteten Unterrichtsfächer ergibt sich folgende Verteilung: 45,0 % MINT-Fächer, 28,7 % sprachliche Fächer, 9,5 % gesellschaftswissenschaftliche Fächer, 7,6 % kreative Fächer und 5,6 % weltanschauliche Fächer.

3.4 Analysen

Für die Überprüfung der Faktorenstruktur wurde eine explorative Faktorenanalyse genutzt. Da einige Items nicht normalverteilt waren, wurde für die explorative Faktorenanalyse ein robuster Maximum-Likelihood-Schätzer verwendet. Weiterhin wurden die Standardfehler bezüglich der hierarchischen Datenstruktur (Beobachtungen geschachtelt in Lehrperson) korrigiert. Für die Rotation der Faktoren wurde eine oblique Geomin-Rotation benutzt, da von bedeutsamen Faktorkorrelationen auszu-

gehen war. Die Überprüfung der Zusammenhänge zwischen den Unterrichtsmerkmalen erfolgte mithilfe von Produkt-Moment-Korrelationen zwischen den Skalen. Als Skalenwert wurde der Mittelwert aller Items der jeweiligen Skala verwendet.

Die interne Konsistenz wurde mithilfe des Reliabilitätskoeffizienten Ω (McDonald 1999) untersucht. Dieses Maß wurde verwendet, weil die vorliegenden Daten die Voraussetzungen des üblicherweise verwendeten Koeffizienten α (ein essenziell tau-äquivalentes Reliabilitätsmodell) nicht erfüllten, während die Berechnung von Ω zulässig war (Dunn et al. 2014). Als Indikator der Test-Retest-Reliabilität verwenden wir die Produkt-Moment-Korrelation zwischen den Skalenwerten des jeweils gleichen Konstrukts bei der ersten und zweiten Unterrichtsbeobachtung. Die Analysen wurden mithilfe von Mplus 8 (Muthen und Muthen 1998–2017) durchgeführt. Fehlende Werte wurden durch die Full-Information-Maximum-Likelihood-Methode behandelt, so dass jeweils alle Fälle eingeschlossen werden konnten. Der Anteil fehlender Werte lag je nach Item zwischen 0,3 und 9,5 %.

3.5 Ergebnisse

3.5.1 Deskriptive Statistiken

Zunächst wurden die Mittelwerte und Standardabweichungen der Items deskriptiv analysiert. Bei einer kleinen Gruppe von Items zeigten sich relativ hohe Mittelwerte, d. h. die Beobachter/-innen bewerteten die Unterrichtsqualität bei diesen Aspekten als sehr gut (siehe Online-Zusatzmaterial 1). Einhergehend mit den hohen Mittelwerten weisen diese Items rechtssteile Verteilungen auf. Davon abgesehen zeigten sich deskriptiv keine weiteren Auffälligkeiten.

3.5.2 Faktorenstruktur

Der erste Schritt der explorativen Faktorenanalyse bestand in der Ermittlung der Anzahl der Faktoren. Die Faktorenanzahl ist relevant für die Modellannahme, dass eine hohe Anzahl von inhaltlich eng gefassten Unterrichtsmerkmalen empirisch abgebildet werden kann. Aus einer theoretischen Perspektive war mit einem 15-Faktoren-Modell zu rechnen. Empirische Ansätze zur Ermittlung der Faktorenanzahl (O'Connor 2000) kamen zu unterschiedlichen Ergebnissen. Das Kaiser-Guttman-Kriterium ergab 14 Faktoren, der MAP-Test ergab 13 Faktoren und die Parallelanalyse ergab 9 Faktoren. Vor diesem Hintergrund wurden explorative Faktorenanalysen mit 9, 13, 14 und 15 Faktoren berechnet. Die Ergebnisse dieser vier Faktorenmodelle wurden wiederum unter theoretischen und empirischen Gesichtspunkten bewertet. In theoretischer Hinsicht erschien das 15-Faktoren-Modell am plausibelsten, da sich alle neu dazu kommenden Faktoren (im Vergleich zu Modellen mit weniger Faktoren) inhaltlich sinnvoll interpretieren ließen². In empirischer Hinsicht wurde die Passung der Modelle anhand des Akaike Information Criterion (AIC) überprüft. Dieser Indi-

² Bei einer explorativen Faktorenanalyse mit 16 Faktoren zeigte sich dagegen, dass sich der neu dazu kommende 16. Faktor inhaltlich nicht sinnvoll interpretieren ließ. Dieser Faktor bündelt Restvarianzen von inhaltlich unterschiedlichen Items.

kator wird für den Vergleich von nicht geschachtelten Modellen empfohlen (Kline 2011). Das 15-Faktoren-Modell wies den niedrigsten Wert auf ($AIC=94685$), somit erschien dieses Modell auch unter empirischen Gesichtspunkten am geeignetsten. Die Modellpassung ($CFI=0,93$, $RMSEA=0,04$, $\chi^2/df=1,83$) kann als zufriedenstellend angesehen werden (Kline 2011).

Als nächstes wurde überprüft, inwieweit die empirischen Faktorladungen den theoretischen Zuordnungen von Items zu Konstrukten entsprechen. Hierbei ist zu berücksichtigen, dass zwischen Unterrichtsmerkmalen teilweise bedeutsame Zusammenhänge bestehen können. Für die explorative Faktorenanalyse ergibt sich hieraus, dass der Anspruch, nach dem Items keine bedeutsamen Querladungen aufweisen sollten, zu streng erscheint. Stattdessen wird der Anspruch gestellt, dass die Items jeweils die höchste Faktorladung auf dem theoretisch zugehörigen Faktor haben, was bedeutet, dass die Items primär das vorgesehene Konstrukt erfassen. Weiterhin sollten theoriekonforme Faktorladungen inhaltlich bedeutsam ($\lambda > 0,30$, Wolff und Bacher 2010) und statistisch signifikant ($p < 0,05$) sein.

Die Ergebnisse lassen sich folgendermaßen zusammenfassen (siehe Online-Zusatzmaterial 2): Bei 9 der 15 Skalen (Erklärung und Präsentation, Speicherung durch Üben, Anwendung und Transfer, Weckung von Interesse, Verdeutlichung des Nutzens, Förderung von Freude, Monitoring, Steuerung des Unterrichtsflusses und Gruppenaktivierung) entspricht das empirische Ladungsmuster der theoretischen Zuordnung der Items zu den Konstrukten. Bei diesen Skalen haben jeweils alle Items die höchste Faktorladung auf dem theoretisch zugehörigen Faktor. Alle Faktorladungen sind inhaltlich bedeutsam ($\lambda=0,35$ bis $\lambda=0,84$) und statistisch signifikant.

Bei den Skalen Unterstützung des Verstehens und Umgang mit Störungen deckt sich das Ladungsmuster weitestgehend mit der theoretischen Einteilung. Bei diesen beiden Skalen hat jeweils ein Item die höchste Faktorladung auf einem anderen als dem zugehörigen Faktor. Beide Querladungen sind inhaltlich plausibel. Die restlichen Items laden jeweils alle inhaltlich bedeutsam ($\lambda=0,37$ bis $\lambda=0,65$) und statistisch signifikant auf den zugehörigen Faktoren.

Bei der Skala Lerndienliches Unterrichtsklima ergeben sich abweichend von der theoretischen Erwartung zwei Faktoren. Beide Faktoren lassen sich inhaltlich sinnvoll interpretieren. Ein Faktor bezieht sich primär auf das Verhalten der Schüler/-innen, der andere primär auf das Verhalten der Lehrperson. Bei beiden Faktoren sind jeweils alle Faktorladungen inhaltlich bedeutsam ($\lambda=0,63$ bis $\lambda=0,88$) und statistisch signifikant.

Bei den drei verbleibenden Skalen Individualisierter Wissenserwerb, Individualisierte Motivierung und Förderung von Selbstwirksamkeit weicht das Ladungsmuster erkennbar von der theoretischen Einteilung ab. Die Items dieser drei Skalen ergeben zusammen zwei Faktoren. Auf dem ersten Faktor laden inhaltlich bedeutsam ($\lambda=0,41$ bis $\lambda=0,87$) und statistisch signifikant zwei Items des Individualisierten Wissenserwerbs und ein Item der Individualisierten Motivierung. Wir interpretieren, dass es sich um einen bereichsübergreifenden Individualisierungs-Faktor handelt. Auf dem zweiten Faktor laden alle Items der Förderung von Selbstwirksamkeit, drei Items der Individualisierten Motivierung sowie zwei Items des Individualisierten Wissenserwerbs inhaltlich bedeutsam ($\lambda=0,33$ bis $\lambda=0,68$) und statistisch signifikant. Da von der Skala Förderung von Selbstwirksamkeit alle Items relativ hohe

Tab. 2 Produkt-Moment-Korrelationen r , Reliabilitätskoeffizient Ω und Test-Retest-Reliabilität r_{tt}

	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
U1	(0,86/0,72)	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U2	0,60	(0,86/0,79)	-	-	-	-	-	-	-	-	-	-	-	-	-
U3	0,43	0,44	(0,82/0,63)	-	-	-	-	-	-	-	-	-	-	-	-
U4	0,47	0,55	0,47	(0,78/0,59)	-	-	-	-	-	-	-	-	-	-	-
U5	0,61	0,61	0,35	0,47	(0,87/0,75)	-	-	-	-	-	-	-	-	-	-
U6	0,52	0,50	0,40	0,47	0,66	(0,84/0,78)	-	-	-	-	-	-	-	-	-
U7	0,35	0,58	0,29	0,38	0,49	0,40	(0,89/0,68)	-	-	-	-	-	-	-	-
U8	0,48	0,55	0,43	0,31	0,62	0,40	0,54	(0,86/0,80)	-	-	-	-	-	-	-
U9	0,39	0,31	0,23	0,37	0,30	0,25	0,37	0,20	(0,90/0,76)	-	-	-	-	-	-
U10	0,53	0,48	0,38	0,35	0,56	0,30	0,52	0,64	0,45	(0,89/0,78)	-	-	-	-	-
U11	0,48	0,46	0,40	0,36	0,45	0,27	0,46	0,47	0,44	0,54	(0,93/0,80)	-	-	-	-
U12	0,53	0,45	0,37	0,44	0,48	0,25	0,49	0,49	0,59	0,52	0,67	(0,87/0,73)	-	-	-
U13	0,53	0,46	0,46	0,46	0,54	0,34	0,64	0,64	0,47	0,54	0,61	0,60	(0,89/0,80)	-	-
U14	0,44	0,39	0,31	0,39	0,45	0,25	0,48	0,48	0,67	0,55	0,67	0,75	0,61	(0,93/0,79)	-
U15	0,22	0,33	0,36	0,19	0,25	0,38	0,27	0,39	0,01	0,19	0,23	0,10	0,34	0,10	(0,84/0,76)

Die oberste diagonale Zellreihe in Klammern gibt die Koeffizienten Ω und r_{tt} wieder, alle anderen Zellen Produkt-Moment-Korrelationen r

U1 Erklärung und Präsentation, U2 Unterstützung des Verstehens, U3 Speicherung durch Üben, U4 Anwendung und Transfer, U5 Weckung von Interesse, U6 Förderung von Freude, U7 Verdeutlichung des Nutzens, U8 Förderung von Selbstwirksamkeit, U9 Lerndienliches Unterrichtsklima Schüler/-innen, U10 Lerndienliches Unterrichtsklima Lehrperson, U11 Monitoring, U12 Steuerung des Unterrichtsflusses, U13 Gruppenaktivierung, U14 Umgang mit Störungen, U15 Individualisierung

Faktorladungen aufweisen ($\lambda = 0,50$ bis $\lambda = 0,66$), von den zwei Individualisierungs-Skalen dagegen jeweils nur bestimmte Items, schlussfolgern wir, dass dem Faktor primär das Konstrukt Förderung von Selbstwirksamkeit zugrunde liegt.

Aufgrund der Ergebnisse der Faktorenanalyse wurden für die nachfolgenden Analysen zwei Modifikationen vorgenommen. Die Bildung von Skalenwerten (für Korrelationsanalysen) als auch der Reliabilitätskoeffizient Ω setzen jeweils eindimensionale Skalen voraus (Döring und Bortz 2016; Dunn et al. 2014). Eindimensionalität ist bei der Skala zum Lerndienlichen Unterrichtsklima und bei den zwei Individualisierungs-Skalen nicht gegeben. Daher wurden beim Lernförderlichen Unterrichtsklima zwei Skalen gebildet, entsprechend den zwei ermittelten Faktoren (Lernförderliches Unterrichtsklima Schüler/-innen und Lernförderliches Unterrichtsklima Lehrperson). Bezüglich Individualisierung wurde aus den drei Items, die auf einem Faktor laden, eine bereichsübergreifende Individualisierungs-Skala gebildet. Durch diese Modifikationen wird im Weiteren explorativ überprüft, ob sich die alternativen Modellierungen empirisch bewähren.

3.5.3 Zusammenhänge zwischen Unterrichtsmerkmalen

Die Analyse der Korrelationen zwischen den Skalen erlaubt Schlussfolgerungen über die empirische Trennbarkeit der Unterrichtsmerkmale. Hierbei muss wiederum berücksichtigt werden, dass bedeutsame Zusammenhänge zwischen Unterrichtsmerkmalen theoretisch plausibel und aufgrund früherer Studien zu erwarten sind. Problematisch sind nur sehr hohe Korrelationen ($r > 0,90$), da es in diesem Fall fraglich erscheint, ob die Skalen verschiedene Konstrukte messen (Kline 2011). Tab. 2 zeigt die vollständige Korrelationsmatrix.

Die Korrelationen variieren zwischen $r = 0,01$ und $r = 0,75$. Zwischen mehreren Skalen bestehen substanzielle Korrelationen, allerdings fällt keine Korrelation so hoch aus, dass anzunehmen wäre, dass zwei Skalen das gleiche Konstrukt messen ($r > 0,90$). Die moderate Korrelation ($r = 0,45$) zwischen den beiden Unterrichtsklima-Skalen bestätigt, dass es sich um zwei empirisch trennbare Aspekte handelt. Weiterhin zeigt sich, dass sich die bereichsübergreifende Individualisierungs-Skala gut von der Skala zur Förderung von Selbstwirksamkeit abgrenzen lässt ($r = 0,39$).

3.5.4 Interne Konsistenz

Die Ergebnisse zur internen Konsistenz in der Form des Reliabilitätskoeffizienten Ω finden sich in der oberen diagonalen Zellreihe von Tab. 2. Die interne Konsistenz variiert zwischen $\Omega = 0,78$ und $\Omega = 0,93$ und liegt somit durchgängig über dem zu erreichenden Schwellenwert für Reliabilitätskoeffizienten von 0,70 (Kline 2011). Somit fällt die interne Konsistenz bei allen Skalen zufriedenstellend aus, auch bei den explorativ gebildeten Unterrichtsklima-Skalen ($\Omega = 0,90$ und $\Omega = 0,89$) und der bereichsübergreifenden Individualisierungs-Skala ($\Omega = 0,84$).

3.5.5 Test-Retest-Reliabilität

Die Produkt-Moment-Korrelationen zwischen den Skalenwerten zum ersten und zweiten Messzeitpunkt befinden sich ebenfalls in Tab. 2. Die Korrelationen variieren zwischen $r_{tt}=0,59$ und $r_{tt}=0,80$. Bei der Interpretation der Test-Retest-Reliabilität ist die zeitliche Stabilität des untersuchten Merkmals zu berücksichtigen. Die Korrelation zwischen den Messungen an zwei Zeitpunkten (r_{tt}) entspricht der Reliabilität, wenn die wahren Werte und das Ausmaß an Messfehlern im betrachteten Zeitraum konstant bleiben. Es ist anzunehmen, dass Lehrpersonen verschiedene Unterrichtsstunden unterschiedlich gestalten, auch wenn die einzelnen Unterrichtsstunden wie in diesem Fall im Abstand von nur wenigen Wochen liegen. Demnach kommt es zu unsystematischen Veränderungen in den wahren Werten der Unterrichtsmerkmale. Daher unterschätzt die Korrelation r_{tt} die tatsächliche Reliabilität. Vor diesem Hintergrund erscheint für die meisten Skalenmittelwerte ein ausreichendes Maß an Stabilität gegeben.

4 Diskussion

Im vorliegenden Beitrag wurde ein Beobachtungsbogen zur Erfassung von Unterrichtsqualität vorgestellt, der bei Unterrichtsevaluationen durch Schulleitungen im Rahmen eines Unterrichtsfeedbacks für Lehrpersonen eingesetzt werden soll. Dazu wurde das bewährte Modell der drei Basisdimensionen der Unterrichtsqualität erweitert. Die Vorteile des Vorgehens liegen in einer theoretischen Herleitung der Unterrichtsmerkmale, dem Fokus auf das Handeln der Lehrperson im Unterricht und einer hohen Auflösung durch 15 inhaltlich eng gefasste Unterrichtsmerkmale. Die Ergebnisse einer explorativen Studie mit Beobachtungsdaten von Lehramtsstudierenden zeigen eine weitgehende Deckung von theoretischer Modellstruktur und empirischer Faktorenstruktur, moderate bis hohe Korrelationen zwischen den Unterrichtsmerkmalen, zufriedenstellende interne Konsistenzen und Stabilitäten. Insgesamt sprechen die Ergebnisse dafür, dass das theoretische Modell empirisch abgebildet werden kann.

Abweichungen zwischen theoretischem Modell und empirischer Faktorenstruktur ergeben sich zunächst bei der Skala zum Unterrichtsklima, welche zwei Faktoren umfasst. Es erscheint theoretisch plausibel, dass respektvolle und freundliche Verhaltensweisen der Schüler/-innen nicht zwangsläufig auch mit entsprechenden Verhaltensweisen der Lehrperson einhergehen und umgekehrt. Die moderate Korrelation zwischen den Skalen zeigt aber, dass sich die beiden Aspekte zumindest begünstigen. Weiterhin bilden die zwei Individualisierungsskalen einen bereichsübergreifenden Individualisierungs-Faktor und weisen Querladungen auf dem Faktor zur *Förderung von Selbstwirksamkeit* auf. Der bereichsübergreifende Faktor legt nahe, dass Beobachter/-innen nicht unterscheiden können, ob individualisierende Maßnahmen auf kognitive oder motivationale Lernvoraussetzungen der Schüler/-innen ausgerichtet sind.

Aus den empirischen Ergebnissen können zwei Modifikationen des theoretischen Modells abgeleitet werden. Erstens sollte die Unterscheidung zwischen schülerbe-

zogenem und lehrpersonenbezogenem Unterrichtsklima aufgegriffen werden und somit ein Unterrichtsmerkmal ergänzt werden. Zweitens sind die beiden Aspekte von Individualisierung in ein gemeinsames Unterrichtsmerkmal zusammenzuführen. Für die Weiterentwicklung des theoretischen Modells in dieser Form sprechen auch erste Analysen mit einem Schülerfragebogen, die zu ähnlichen Ergebnissen bezüglich der Faktorenstruktur kommen (Gärtner et al. [under review](#)). Aufbauend auf dieser Weiterentwicklung des theoretischen Modells kann der Beobachtungsbogen entsprechend überarbeitet werden.

Limitationen des vorliegenden Beitrags sollen im Folgenden thematisiert werden: Die empirische Überprüfung des Beobachtungsbogens erfolgte zwar mit einer vergleichsweise großen Stichprobe, allerdings handelte es sich nicht um Schulleitungen, sondern um trainierte Lehramtsstudierende. Die Rekrutierung einer entsprechenden großen Schulleitungsstichprobe war im Rahmen des Projekts nicht möglich. Auch der Umfang des Trainings war mit zehn Stunden vergleichsweise niedrig. Es spricht allerdings für die Konstrukte, dass trotz dieser Limitation die theoretisch angenommene Faktorenstruktur weitgehend belegt werden konnte. Weitere Kontextfaktoren konnten im Rahmen der Unterrichtsevaluation durch die Studierenden nicht erfasst werden. Auch die Inter-Rater-Reliabilität ließ sich mit den vorliegenden Daten nicht überprüfen. Zwar wurde im Beobachtertraining auf eine Kalibrierung der Beurteilung großer Wert gelegt, allein kann diese eine Überprüfung der Inter-Rater-Reliabilität nicht ersetzen.

Gleichwohl liefert die vorliegende explorative Analyse erste Evidenzen für die Eignung des Beobachtungsbogens für eine Unterrichtsevaluation durch Schulleitungen. Die Ergebnisse weisen darauf hin, dass es im Interesse der theoretischen Trennschärfe sinnvoll ist, die drei Basisdimensionen auf der Grundlage der gewählten Bezugstheorien zu operationalisieren und so die kognitive, die motivationale und die soziale Perspektive auf Unterricht eindeutig zu trennen. Dies erleichtert nicht nur die Konzeption und Durchführung von Beobachtungstrainings für Schulleitungen, sondern stellt auch einen klaren Rahmen für deren Feedback an die beurteilten Lehrkräfte bereit. Die Erfahrungen mit der Nutzung des Bogens im Rahmen eines Schulleitungstrainings bestätigen diese Einschätzung. Auch die eigentliche Zielgruppe des Bogens konnte ihn nach dem Training praktikabel einsetzen. Die Wirkung eines auf dem Beobachtungsbogen basierendem Unterrichtsfeedbacks von Schulleitungen auf die von Schüler/-innen wahrgenommene Unterrichtsqualität wurde im Rahmen einer Interventionsstudie untersucht. Schüler/-innen von Lehrpersonen, die ein Unterrichtsfeedback von ihrer Schulleitung erhielten, schätzten die Unterrichtsqualität nach der Feedbackintervention bei fünf der beschriebenen Unterrichtsmerkmale signifikant besser ein, als Schüler/-innen von Lehrpersonen, die kein Feedback erhielten (Kellermann et al. [under review](#)). Die Interventionsstudie liefert somit Belege dafür, dass der Beobachtungsbogen für wirksame Unterrichtsentwicklung verwendet werden kann und unterstreicht dessen Praktikabilität.

Anschlussstudien sollten an den aufgezeigten Limitationen ansetzen. Die in der vorliegenden explorativen Studie durchgeführten Analysen sollten mit einer Schulleitungsstichprobe wiederholt werden. Zusätzlich sollte die Inter-Rater-Reliabilität nach Training, sowohl im Unterricht als auch mittels Videomaterial, überprüft werden. Zudem muss durch Invarianztestungen geprüft werden, ob die gefundene Fak-

toerenstruktur gleichermaßen für alle Schulformen und Jahrgangsstufen Gültigkeit besitzt. Dies würde die Voraussetzungen für einen breiten Einsatz des Beobachtungsbogens zu Feedbackzwecken, nicht nur durch Schulleitungen, schaffen.

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (<https://doi.org/10.1007/s35834-021-00321-7>) enthalten.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369–406.
- Bandura, A. (1976). *Lernen am Modell*. Stuttgart: Klett.
- Clausen, M. (2002). *Unterrichtsqualität. Eine Frage der Perspektive?* Münster: Waxmann.
- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45, 378–387.
- Darling-Hammond, L., Wise, A.E., & Pease, S.R. (1983). Teacher evaluation in the organizational context: a review of the literature. *Review of Educational Research*, 53(3), 285–328.
- Deci, E.L. & Ryan, R.M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik* 39(2), S.223-238. <https://doi.org/10.25656/01:11173>
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin: Springer.
- Doyle, W. (2006). Ecological approaches to classroom management. In C.M. Evertson & C.S. Weinstein (Hrsg.), *Handbook of classroom management* (S. 97–126). Mahwah: Lawrence Erlbaum.
- Dunn, T.J., Baguley, T., & Brunson, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105, 399–412.
- Eccles, J.S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual review of psychology*, 53, 109–132.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift für Pädagogische Psychologie*, 28, 127–137.
- Friedrich, H.F., & Mandl, H. (1992). Lern- und Denkstrategien – ein Problemaufriß. In H.F. Friedrich & H. Mandl (Hrsg.), *Lern- und Denkstrategien. Analyse und Intervention* (S. 3–54). Göttingen: Hogrefe.
- Gärtner, H. (2010). Das ISQ-Selbstevaluationsportal. Konzeption eines Online-Angebots, um die Selbstevaluation in Schule und Unterricht zu unterstützen. *Die Deutsche Schule*, 102(2), 163–175.
- Gärtner, H., & Wurster, S. (2019). Instrumente und Verfahren der Evaluation in Schule und Unterricht. In C. Buhren, Klein, G. & S. Müller (Hrsg.), *Handbuch Evaluation in Schule und Unterricht* (S. 113–127). Weinheim: Beltz.
- Gärtner, H., Thiel, F., Kellermann, C. & Nachbauer, M. (under review). Ein theoriebasierter Schülerfragebogen für Unterrichtsevaluation.
- Göllner, R., Wagner, W., Klieme, E., Lüdtke, O., Nagengast, B., & Trautwein, U. (2016). Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven.

- In BMBF (Hrsg.), *Forschungsvorhaben in Anknüpfung an Large-Scale-Assessments*. Bildungsforschung, (Bd. 44, S. 63–82). Berlin: BMBF.
- Hallinger, P. (2011). Leadership for learning: lessons from 40 years of empirical re-search. *Journal of Educational Administration*, 49(2), 125–142.
- Helmke, A., & Schrader, F.-W. (1997). Unterrichtsbeurteilungen durch externe Beobachter. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 510–518). Weinheim: Beltz.
- Heneman, H. G., & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, 17(2), 173–195.
- Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, 326, 1410–1412.
- Kellermann, C., Nachbauer, M., Gärtner, H. & Thiel, F. (under review). The effects of school leader's feedback on teaching quality – Results of an intervention study.
- Klauer, K. J., & Leutner, D. (2012). *Lehren und Lernen: Einführung in die Instruktionspsychologie*. Weinheim: Beltz.
- Klieme, E., & Warwas, J. (2011). Konzepte der Individuellen Förderung. *Zeitschrift für Pädagogik*, 57, 805–818.
- Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur“ und Unterrichtsgestaltung. In BMBF (Hrsg.), *TIMSS – Impulse für Schule und Unterricht, Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–58). Bonn: BMBF.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York: Guilford.
- Kounin, J. S. (2006). *Techniken der Klassenführung*. Münster: Waxmann.
- Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: a meta-analysis of the causal evidence. *Review of Educational Research*, 88, 547–588.
- Lenke, G., & Mayr, J. (2015). Das Linzer Konzept der Klassenführung (LKK). Grundlagen, Prinzipien und Umsetzung in der Lehrerbildung. In K. Zierer & al (Hrsg.), *Jahrbuch für Allgemeine Didaktik* (S. 71–84). Baltmannsweiler: Schneider.
- Lipowsky, F. (2009). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 73–102). Berlin: Springer.
- McDonald, R. P. (1999). *Test theory. A unified treatment*. Mahwah: Lawrence Erlbaum.
- Meetz, F. (2007). *Personalentwicklung als Element der Schulentwicklung. Bestandsaufnahme und Perspektiven*. Bad Heilbrunn: Klinkhardt.
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide*. Los Angeles: Muthén & Muthén.
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32, 396–402.
- Ophardt, D., & Thiel, F. (2013). *Klassenmanagement. Ein Handbuch für Studium und Praxis*. Stuttgart: Kohlhammer.
- Pietsch, M., Tulowitzki, P., & Koch, T. (2019). On the differential and shared effects of leadership for learning on teachers' organizational commitment and job satisfaction: a multilevel perspective. *Education Administration Quarterly*, 55(5), 705–741.
- Piwovar, V. (2013). Multidimensionale Erfassung von Kompetenzen im Klassenmanagement: Konstruktion und Validierung eines Beobachter- und eines Schülerfragebogens für die Sekundarstufe 1. *Zeitschrift für Pädagogische Psychologie*, 27, 215–228.
- Praetorius, A.-K. (2013). Einschätzung von Unterrichtsqualität durch externe Beobachterinnen und Beobachter. *Beiträge zur Lehrerinnen- und Lehrerbildung*, 31, 174–185.
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of three basic dimensions. *ZDM*, 50, 407–426.
- Rauscher, H. (2000). Lehrerbeurteilung und Beratung durch den Schulleiter. *Schulleiter-Handbuch*, 94, 6–98.
- Rheinberg, F. (2008). *Motivation*. Stuttgart: Kohlhammer.
- Schiefele, U. (2009). Situational and individual interest. In K. R. Wenzel & A. Wigfield (Hrsg.), *Educational psychology handbook series. Handbook of motivation at school*. (S. 197–222). New York: Routledge.
- Steinberg, M. P., & Sartain, L. (2015). Does better observation make better teachers? New evidence from a teacher evaluation pilot in Chicago. *Education Next*, 15, 71–76.
- Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive science*, 12, 257–285.
- Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102, 3628–3651.

Thiel, F. (2016). *Interaktion im Unterricht*. Opladen: Barbara Budrich.

Wolff, H.-G., & Bacher, J. (2010). Hauptkomponentenanalyse und explorative Faktorenanalyse. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 333–365). Wiesbaden: Springer.