Forschungsmonitor Schule
Rezensionen für die Praxis

Mirko Krüger

Rezension zu

Wenger, M., Lüdtke, O. & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene. Ergebnisse aus 81 Ländern. *Zeitschrift für Erziehungswissenschaft, 21*(5), 929–950.

Kommentierter Kurzbefund

Unterrichtsqualität wird oft anhand von Schülerurteilen ermittelt, beispielsweise in der pädagogischen Praxis bei externen Evaluationen (Schulinspektionen) oder in der Schul- und Unterrichtsforschung. Eine derartige Bestimmung der Unterrichtsqualität auf Schulebene setzt jedoch voraus, dass a) die Schülerurteile innerhalb der Schulen angemessen übereinstimmen, b) sich die Einschätzungen zwischen den Schulen systematisch unterscheiden und c) die Angaben hinreichend zuverlässig bzw. messgenau (reliabel) sind.

Wenger et al. untersuchen im Rückgriff auf Daten der PISA-Studien von 2000 bis 2012, inwiefern diese Voraussetzungen erfüllt sind im Hinblick auf Basisdimensionen der Unterrichtsqualität: Klassenführung, kognitive Aktivierung und konstruktive Unterstützung. In die Analysen flossen Daten von mehr als 1,3 Millionen Lernenden aus 55.300 Schulen und 81 Bildungssystemen ein. Die Auswertungen basieren auf Berechnungen von Übereinstimmungsmaßen: $r_{WG(J)}$, ICC(1) und ICC(2).

Im Ergebnis liegen für die meisten Merkmale von Unterrichtsqualität zufriedenstellende Übereinstimmungen zwischen den Schülerurteilen einer Schule vor und es gibt (geringe) systematische Variabilität zwischen den Schulen, jedoch sind die aggregierten Schülerurteile häufig nicht ausreichend messgenau, was u. a. daran liegt, dass an der PISA-Studie maximal 35 Schülerinnen und Schüler von einer Schule teilnehmen.

Die Studie fördert das Verständnis darüber, unter welchen Bedingungen aggregierte Schülerurteile zur Unterrichtsqualität genutzt werden können und auf welche Voraussetzungen von Seiten z. B. der Schulinspektion dabei zu achten ist. Die insgesamt eher ungünstigen Voraussetzungen zur Bestimmung von Unterrichtsqualität auf Schulebene anhand von Schülerurteilen weisen auf ihre eingeschränkte Nutzbarkeit hin. Dies umso mehr, als jenseits der betrachteten technischen Aspekte fraglich erscheint, inwiefern ein Durchschnittswert zur Unterrichtsqualität hilfreiche Informationen liefert, ob die Bestimmung von Unterrichtsqualität auf Grundlage allein von Schülereinschätzungen sinnvoll ist und in welchen Zusammenhängen eine Betrachtung auf Schulebene zweckdienlich sein kann. Letzteres ist nicht zuletzt deshalb fraglich, weil auf Klassenebene lernrelevantere Unterschiede in der Unterrichtsqualität zu verorten sein dürften.

Hintergrund

Einleitend konstatieren Wenger et al., dass Schülerurteile eine Informationsquelle sind, die in verschiedenen Kontexten zur Einschätzung der Unterrichtsqualität genutzt wird, z. B. in der empirischen

Schul- und Unterrichtsforschung oder in der pädagogischen Praxis, etwa bei externen Evaluationen wie der Schulinspektion. Dabei stelle sich die Frage nach der Aussagekraft dieser Informationsquelle. Für eine ausreichende Validität der Schülerurteile sollten nach Wenger et al. a) eine angemessene Übereinstimmung innerhalb der Schulen, b) eine systematische Variabilität der Schülerurteile zwischen Schulen sowie c) ein ausreichendes Maß an Zuverlässigkeit (Reliabilität) der aggregierten Urteile vorliegen. Im Kontext international vergleichender Analysen wurde dies bislang kaum systematisch untersucht, so dass sich das Autorenteam dieses Desiderats annimmt.

Im theoretischen Hintergrund werden von Wenger et al. grundlegende Informationen zu statistischen Kennwerten zur Analyse der Unterrichtsqualität von Schulen und zu Theorien sowie Befunde zu Unterschieden der Unterrichtsqualität von Schulen aus Sicht von Schülerinnen und Schülern gegeben:

Statistische Kennwerte zur Analyse der Unterrichtsqualität von Schulen

Unterrichtsmerkmale werden in der Schul- und Unterrichtsforschung häufig durch Schülerurteile erhoben, da es sich um ein bewährtes, ökonomisches und valides Verfahren handelt. Um die Unterrichtsqualität auf Klassen- oder Schulebene zu messen, werden die individuellen Daten zu einem Schulmittelwert aggregiert, der „bei angemessener Übereinstimmung zwischen den Schülerinnen und Schülern ein von allen bzw. der Mehrzahl geteiltes Urteil der Unterrichtsqualität repräsentiert“ (S. 931).

Die Übereinstimmung zwischen den Schülerinnen und Schülern wiederum wird mit Hilfe von Indizes ermittelt. Eines der wichtigsten Übereinstimmungsmaße ist der r_{WG} (bzw. r_{WGJ} bei mehreren Items), für dessen Berechnung die Varianz der empirischen Verteilung durch die Varianz einer (theoretischen) Gleichverteilung dividiert und von 1 abgezogen wird. Bei perfekter Übereinstimmung der Schülerurteile ist ihre Varianz null und der $r_{WG} = 1$. Bei großer Diskrepanz geht er gegen null und kann sogar negative Werte annehmen. Der kritische Wert wird in der Literatur häufig auf $r_{WG} = 0.7$ festgelegt und sollte nicht unterschritten werden, um verlässliche Aussagen über die Unterrichtsqualität an einer Schule treffen zu können.

Da sich der r_{WG} auf die einzelnen Schulen bezieht, bleibt jedoch offen, ob sich die Schulen hinsichtlich ihrer Unterrichtsqualität systematisch unterscheiden und wie messgenau die Schulmittelwerte sind. Hierfür benötigt man weitere Indizes, die sog. Intraklassenkorrelationen ICC(1) und ICC(2).

„Die ICC(1) ist ein wichtiges Maß, um zu bestimmen, inwiefern sich die auf Schulebene gemittelten Urteile der Schülerinnen und Schüler zur Unterrichtsqualität systematisch zwischen Schulen unterscheiden. Die ICC(1) informiert jedoch nicht über die Reliabilität dieser Mittelwerte und berücksichtigt nicht die Anzahl an Schülerurteilen, die in die Berechnung der Mittelwerte einging. Hier setzt die ICC(2) an, die angibt, wie reliabel die Schulmittelwerte zur Unterrichtsqualität ausfallen, die auf Grundlage der Urteile der Schülerinnen und Schüler berechnet wurden“ (S. 933).

Die ICC(1) wird ähnlich wie Effektstärken interpretiert: Bei einer ICC(1) = 0.01 wird von einem „kleinen“, ab einer ICC(1) ≥ 0.10 von einem „mittleren“ und ab einer ICC(1) ≥ 0.25 von einem „großen“ Effekt gesprochen. Der kritische Wert bei der ICC(2) ist in der Literatur auf 0.70 festgelegt.

Theorien und Befunde zu Unterschieden der Unterrichtsqualität von Schulen aus Sicht von Schülerinnen und Schülern

Aus den von Wenger et al. referierten Bezügen zu Diskursen und Befunden der Schulforschung wird deutlich, dass sich Schulen mit Blick auf verschiedene Aspekte (z. B. die Unterrichtsqualität) unterscheiden. Es liegt also aus theoretischer sowie datengestützter Perspektive zwischen einzelnen

Schulen ein gewisses Maß an Variabilität vor, die für die Forschung (bspw. bei Fragen ihrer Effektivität oder der Wirkung der Schülerkomposition auf ihren Bildungserfolg) und die pädagogische Praxis (bspw. bei externer Evaluation durch die Schulinspektion) relevant ist.
Dabei merkt das Autorenteam an, dass in der empirischen Bildungsforschung bislang wenig Studien existieren, die sich mit der „Beurteilungsübereinstimmung, systematischen Unterschieden in der Unterrichtsqualität zwischen Schulen aus Sicht der Schülerinnen und Schüler oder der Reliabilität von aggregierten Schülerurteilen befassen“ (S. 934). Zwar lägen im Rekurs auf die Analyse der Unterrichtsqualität zahlreiche Ergebnisse für Unterschiede zwischen einzelnen Klassen, aber nicht zwischen einzelnen Schulen vor. Die wenigen Befunde zur Variabilität zwischen einzelnen Schulen zu anderen Indikatoren (z. B. Disziplinprobleme) lieferten zumindest Annahmen darüber, dass diese mit Blick auf die Unterrichtsqualität auch schulvergleichend vorliegen könnte.

Vor dem Hintergrund dieser Ausführungen werden folgende Fragestellungen formuliert:

1. Inwieweit stimmen Schülerurteile von Unterrichtsmerkmalen in den Schulen überein?
2. Wie stark unterscheidet sich die Unterrichtsqualität aus Sicht der Schülerinnen und Schüler zwischen Schulen?
3. Wie reliabel sind die aggregierten Schülerurteile zur Unterrichtsqualität?

Design

Stichprobe

In dieser Studie wurden Schülerdaten (15-Jährige) aus den internationalen PISA-Erhebungen der Jahre 2000, 2003, 2009 und 2012 berücksichtigt. Durch eine mehrstufige systematische Stichprobenziehung (mit Rotationsdesign) flossen in diese Untersuchung die Daten von insgesamt mehr als 1,3 Millionen Jugendlichen von über 55.300 Schulen aus 81 verschiedenen Bildungssystemen ein.

Instrumente

In den PISA-Erhebungen kamen in verschiedenen Fachkontexten Fragebögen zur Erfassung unterschiedlicher Unterrichtsmerkmale zum Einsatz, welche für die weiteren Analysen den drei Dimensionen der Unterrichtsqualität *Klassenführung*, *kognitive Aktivierung* und *konstruktive Unterstützung* zugeordnet wurden: Die Dimension Klassenführung umfasste eine Skala zur Disziplin (5 Items) und eine Skala zur Klassenführung (4 Items). Die Reliabilitätswerte für beide Skalen lagen im Mittel über alle Bildungssysteme bei Cronbachs $\alpha_{\text{Med OECD}} \geq 0.72$. Zur Erfassung der kognitiven Aktivierung wurde eine Skala mit 9 Items berücksichtigt (Cronbachs $\alpha_{\text{Med OECD}} = 0.83$). Der Dimension konstruktive Unterstützung wurden die Skalen Unterstützung (5 Items), Leistungsdruck (4 Items), lehrerzentrierte Instruktion (5 Items), Schülerorientierung (4 Items), Strukturierungsstrategien der Lehrkräfte (9 Items), Verhalten der Lehrkräfte bei Rückmeldungen (4 Items) und Beziehungsqualität (5 Items) zugeordnet. Im Mittel über alle Länder lagen die Reliabilitätswerte der Skalen bis auf eine Ausnahme (Leistungsdruck: Cronbachs $\alpha_{\text{Med OECD}} = 0.54$) im Bereich von $0.68 \leq$ Cronbachs $\alpha_{\text{Med OECD}} \leq 0.83$.

Auswertung

Zur Beantwortung der Fragestellungen wurden mit Hilfe verschiedener R-Pakete (multilevel, lme4, metafor) aufwendige Auswertungen vorgenommen: Zunächst wurde zur Prüfung der Übereinstimmung von Schülerurteilen hinsichtlich der Unterrichtsmerkmale in den Schulen für jede Skala der r_{WGJ}-Index gebildet und für alle Schulen eines jeden Bildungssystems gemittelt. Anschließend wurden die Mediane der Ergebnisse der Bildungssysteme berechnet. Des Weiteren wurde der prozentuale Anteil an Schulen

ermittelt, die über dem Grenzwert von $r_{WG} = 0.70$ lagen.

Um Aussagen über die Variabilität und Reliabilität der Schülerinnen- und Schülerurteile zur Unterrichtsqualität treffen zu können, wurden die ICC(1) und ICC(2) inklusive der 95%-Konfidenzintervalle ermittelt. Dabei wurde auf die Verwendung von Stichprobengewichten verzichtet. Weiterhin wurden für mehrfach erhobene Unterrichtsmerkmale die Varianz zwischen den Ländern und zwischen den Zeitpunkten erhoben, sodass Aussagen darüber möglich sind, inwiefern Unterschiede in den Übereinstimmungsmaßen r_{WGJ}, ICC(1) und ICC(2) zwischen den Ländern zeitlich stabil sind oder variieren. Fehlende Werte wurden fallweise ausgeschlossen, da ihre Anzahl sehr gering war.

Ergebnisse

Für die meisten Länder zeigen sich moderate bis starke Übereinstimmungen der Schülerurteile in den Schulen. Davon ausgenommen sind die beiden Merkmale Schülerorientierung und Rückmeldung, die zur konstruktiven Unterstützung zählen. Die Werte für diese beiden Skalen liegen in der überwiegenden Zahl der Bildungssysteme unter dem Grenzwert, sodass die Aussagekraft der aggregierten Schülerurteile auf Schulebene eher begrenzt erscheint und die Durchschnittswerte dieser Unterrichtsmerkmale die Einschätzungen vieler Schülerinnen und Schüler an diesen Schulen nur unzutreffend abbilden.

Es ergeben sich systematische Unterschiede in der Unterrichtsqualität zwischen den Schulen, die gemäß gängiger Konventionen jedoch überwiegend „klein“ oder „mittel“ ausfallen. Im Hinblick auf die Messgenauigkeit der Einschätzungen zur Unterrichtsqualität belegen die ICC(2)-Berechnungen, dass die Reliabilitätskennwerte in der Mehrheit der Bildungssysteme nicht im akzeptablen bzw. zufriedenstellenden Bereich liegen.

Diskussion und Einschätzung

Hintergrund

Die Studie von Wenger, Lüdtke und Brunner greift vor dem Hintergrund der wissenschaftlichen und praxisbezogenen Diskussion über die Bestimmung der Unterrichtsqualität auf der Grundlage von Schülerurteilen ein für die Administration und Schule relevantes Forschungsdesiderat auf.

Zunächst führt die Autorengruppe in die statistischen Kennwerte zur Analyse der Aussagekraft von Schülerurteilen zur Unterrichtsqualität von Schulen ein, die in der Untersuchung Anwendung finden. Als Maß der Übereinstimmung von Schülerurteilen an einer Schule wird häufig der r_{WG} genutzt. Die ICC(1) und ICC(2) kommen als Übereinstimmungsmaße zum Einsatz, um die Variabilität und Reliabilität hinsichtlich interessierender Indikatoren (hier der Unterrichtsqualität) bezogen auf die Gesamtstichprobe zu analysieren.

Im Anschluss an diese Ausführungen beleuchten Wenger et al. kurz Theorien und Befunde zu Unterschieden der Unterrichtsqualität von Schulen aus Sicht von Schülerinnen und Schülern, aus denen sich ableiten lässt, dass die Erforschung der Beurteilungsübereinstimmung von systematischen Unterschieden in der Unterrichtsqualität zwischen Schulen auf der Basis von Schülerurteilen oder der Messgenauigkeit von Schülerurteilen ein Desiderat zu sein scheint.

Die Argumentationsweise und Hinführung zur eigenen Studie erscheinen aus Sicht des Rezensenten bedingt gelungen. Es wird nicht herausgearbeitet, welchen Mehrwert eine international vergleichende Studie in diesem Kontext besitzt. Zudem fehlen Hinweise darauf, welche empirisch abgesicherten Erkenntnisse zur interessierenden Fragestellung es aus anderen Ländern geben könnte. Wurde die Literatur dahingehend systematisch gesichtet? Oder entschied sich die Autorengruppe bewusst dagegen? Mit welcher Begründung? Diese Fragen bleiben unbeantwortet.

Design
Das Studiendesign und die Durchführung werden ausführlich und nachvollziehbar benannt. Beeindruckend ist die Gesamtstichprobe, auf die zurückgegriffen werden kann. Die Angaben zu den verwendeten Forschungsinstrumenten werden unter Verweis auf die Primärquellen gegeben. Die statistische Auswertung überzeugt.

Ergebnisse
Insgesamt weisen die vorliegenden Ergebnisse darauf hin, dass die Aussagekraft von aggregierten Schülerurteilen (z. B. in Form von Schulmittelwerten) vielfach begrenzt ist und dass ein schulbezogener Durchschnittswert nur unzutreffend abbildet, wie bestimmte Unterrichtsmerkmale von den Schülerinnen und Schülern an diesen Schulen wahrgenommen werden. Wenger et al. stellen selbst infrage, inwiefern sie durch ein Konsensusmodell zutreffend beschrieben werden oder ob hier nicht von Dispersionsmodellen (s. z. B. Chan 1998) ausgegangen und entsprechende Indizes (z. B. die Standardabweichung der Schülerurteile in einer Schule; s. LeBreton und Senter 2008) berechnet werden sollten. Insbesondere für die Schulinspektionspraxis legt dies nahe, dass generell auch die Heterogenität der Unterrichtsqualität innerhalb einer Schule stärker in den Blick genommen (vgl. Wurster und Gärtner 2013) und Kennwerte berechnet werden sollten, die diese Unterschiede innerhalb von Schulen quantifizieren (z. B. Dispersionsmaße). Generell unterstreichen die vorliegenden Ergebnisse zur Übereinstimmung, dass im Kontext der Evaluation stets eine Maßzahl der Beurteilerübereinstimmung mit angegeben werden sollte, um eine Einschätzung zu ermöglichen, inwiefern aggregierte Urteile die Wahrnehmung aller Schülerinnen und Schüler abbilden (Lüdtke et al. 2006; Gärtner 2010).

Wenger et al. stellen abschließend gut nachvollziehbar die Limitationen ihrer Studie vor:
Aufgrund der Studienanlage konnten die Schülerinnen und Schüler lediglich auf Schul-, jedoch nicht auf Klassenebene zugeordnet werden. Deswegen sind die Befunde dieser Studie aus ihrer Sicht nicht auf Mehrebenenanalysen übertragbar, in denen die Klassenebene modelliert wird. Zudem wurde in dieser Studie der Fokus auf einen kleinen Bereich der Unterrichtsqualität und auf Schülerurteile gerichtet. Zukünftige Forschung sollte weitere Unterrichtsmerkmale und Perspektiven (z. B. externe Beobachter) berücksichtigen. Angesichts der Nutzung ungewichteter Daten, die flexible Analysen (z. B. Bootstrapverfahren für Konfidenzintervalle) mit Hilfe von R-Paketen ermöglichte, stellt sich auch die Frage nach der Verteilung der Werte mit gewichteten Daten, wenngleich das Autorenteam durch die Prüfung einzelner Variablen von sehr ähnlichen Resultaten ausgeht. Abschließend bemerken Wenger et al., dass durch den eher globalen Analysefokus über 81 Bildungssysteme dezidierte Länder- oder Ländergruppenanalysen vernachlässigt werden mussten.

Reflexionsfragen für die Praxis

Nachfolgende Reflexionsfragen sind ein Angebot, die Befunde der rezensierten Studie auf das eigene Handeln als Lehrkraft oder Schulleitungsmitglied zu beziehen und zu überlegen, inwiefern sich Anregungen für die eigene Handlungspraxis ergeben. Die Befunde der rezensierten Studien sind nicht

immer generalisierbar, was z. B. in einer begrenzten Stichprobe begründet ist. Aber auch in diesen Fällen können die Ergebnisse interessante Hinweise liefern, um über die eigene pädagogische und schulentwicklerische Praxis zu reflektieren.

Reflexionsfragen für Lehrkräfte:

- Wie ist meine Einstellung gegenüber der Einholung von Schülerrückmeldungen zu meinem Unterricht? Welche Barrieren müssen beseitigt werden, um meine Haltung positiver zu gestalten?
- Inwieweit nutze ich Möglichkeiten, mir Rückmeldungen von Schülerinnen und Schülern zu meinem Unterricht einzuholen? Und inwiefern leite ich hieraus Optimierungsbedarf ab?
- Inwiefern greife ich auf die Beurteilung meines Unterrichts durch Kolleginnen und Kollegen oder externe Beurteiler zurück und kenne hierfür einsetzbare Instrumente?
- Welche konkreten Maßnahmen sind an unserer Schule implementiert oder sollten implementiert werden, um durch Unterrichtsbeurteilungen Unterrichtsentwicklung zu betreiben?

Reflexionsfragen für Schulleitungen:

- Welche Maßnahmen zur Beurteilung der Unterrichtsqualität werden an meiner Schule verwendet? Inwieweit werden die Ergebnisse zur Unterrichtsentwicklung genutzt?
- Inwiefern ermögliche ich an meiner Schule einen Austausch über Unterrichtsqualität in den einzelnen Fächern?
- Inwiefern ist die Qualitätssicherung und -steigerung des Unterrichts an unserer Schule konzeptionell verankert? Welche Verfahren werden dahingehend verwendet und wie werden diese kommuniziert?
- Welche verbindlichen Konsequenzen könnten sich aus der regelmäßigen Beurteilung der Unterrichtsqualität ergeben? Wie kann ich mit meinem Kollegium dahingehend in einen Austausch gelangen?

Literatur

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

Gärtner, H. (2010). Wie Schülerinnen und Schüler ihre Lernumwelt wahrnehmen. *Zeitschrift für Pädagogische Psychologie, 24*(2), 111–122.

LeBreton, J. M. & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 1*(4), 815–852.

Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Analyse von Lernumwelten. Ansätze zur Bestimmung der Reliabilität und Übereinstimmung von Schülerwahrnehmungen. *Zeitschrift für Pädagogische Psychologie, 20*(1/2), 85–96.

Wurster, S. & Gärtner, H. (2013). Erfassung von Bildungsprozessen im Rahmen von Schulinspektion und deren potenzieller Nutzen für die empirische Bildungsforschung. *Unterrichtswissenschaft, 41*(3), 217–236.

Rezensent/-in

Dr. Mirko Krüger, PD, Lehrer an der Georg-Müller-Gesamtschule in Wetter (Ruhr) und Lehrbeauftragter an der Fakultät für Bildungswissenschaften, Universität Duisburg-Essen. Arbeitsschwerpunkte: Schul- und Schulsportentwicklung, Sprachbildung im Sportunterricht, Professionalisierung von Lehrkräften

Zitiervorschlag

Krüger, M. (2020). Rezension zu Wenger, M., Lüdtke, O. & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene. Ergebnisse aus 81 Ländern. Zeitschrift für Erziehungswissenschaft, 21(5), 929–950. *Forschungsmonitor Schule, 100*. Abgerufen von https://www.forschungsmonitor-schule.de/print.php?id=109

Urheberrecht