Cramer, Colin

Anna-Katharina Praetorius: Messung von Unterrichtsqualität durch Ratings. Münster: Waxmann 2014. [Rezension]

Erziehungswissenschaftliche Revue (EWR) 14 (2015) 1

Erziehungswissenschaftliche Revue (EWR)

Rezensionszeitschrift für alle Teilbereiche der Erziehungswissenschaft
ISSN 1613-0677

Die Rezensionen werden in die Zeitschrift mittels eines Peer-Review-Verfahrens aufgenommen.

Weitere Informationen: http://www.klinkhardt.de/ewr/

Kontakt: EWR@klinkhardt.de

EWR 14 (2015), Nr. 1 (Januar/Februar)

Anna-Katharina Praetorius
Messung von Unterrichtsqualität durch Ratings
Münster: Waxmann 2014
(324 S.; ISBN 978-3-8309-2980-2; 29,90 EUR)

Anna-Katharina Praetorius hinterfragt in ihrer Dissertation die Qualität hoch-inferenter Ratings durch Beobachter, die oftmals zur Messung von Unterrichtsqualität herangezogen werden. Ausgangspunkt der Arbeit ist die Feststellung: „Trotz des häufigen Einsatzes von Unterrichtseinschätzungen externer Beobachter wurde bislang kaum erforscht, wie zuverlässig und wie valide diese Einschätzungen tatsächlich für Aussagen über Unterrichtsqualität sind“ (14). Diesem Desiderat folgend geht die Autorin durch Reanalyse empirischer Studien fünf Forschungsfragen zur Güte bzw. Eignung solcher Ratings nach: Sie betreffen die Verzerrung der Ratings durch Rater-Bias, die Reliabilität und Validität der Ratings von geschulten versus ungeschulten Beobachtern, das Ausmaß der situationellen Beeinflussung der Ratings, mögliche Probleme im Urteilsprozess der Rater sowie die zeitliche Stabilität der beurteilten Merkmale.

Der theoretische Teil der Arbeit führt zunächst in Grundlagen der Forschung zur Unterrichtsqualität ein (Kapitel zwei) und thematisiert in Kapitel drei Gründe für die Messung von Unterrichtsqualität sowie Anforderungen an Messinstrumente, Wege der Datenerhebung und zentrale Datenquellen (Lehrkräfte, Schüler und Beobachter). Definitionen, Arten, Ausmaß und Ursachen von Rater-Biases werden daraufhin in Kapitel vier grundgelegt. In Kapitel fünf wird das Anliegen der Arbeit konkretisiert, indem die Forschungsfragen expliziert und in ein eigenes Modell eingeordnet werden.

Kapitel sechs bereitet den empirischen Teil vor, indem die Generalisierbarkeitstheorie (G-Theorie) vorgestellt wird. Dieser sophistizierte Ansatz kritisiert die Axiomatik der klassischen Testtheorie und die üblichen, direkt daraus ableitbaren Konsequenzen für die Beurteilung der Zuverlässigkeit von Messungen (z. B. die Notwendigkeit unbegrenzter Beobachtungen zu potenziell allen Bedingungen). Diese in der Sozialforschung nicht zu realisierenden Bedingungen werden in der G-Theorie durch das Konzept des universe-score ersetzt. Hier wird die Differenz von Stichprobe und Grundgesamtheit explizit mit aufgenommen – Reliabilität ist dann nicht unabhängig von der Frage nach Generalisierbarkeit zu denken. Es kommen zahlreiche weitere quantitative und auch qualitative Verfahren zum Einsatz.

Nun folgen fünf empirische Studien, die jeweils in sich abgeschlossene Reanalysen darstellen. Sie beruhen auf Daten aus der VERA-Studie und der Pythagoras-Studie. Der ersten Studie (Kapitel sieben) zufolge entfallen erhebliche Varianzanteile in den Ratings der Klassenführung und Schülerorientierung auf Beobachtereffekte. Die Eignung der bestehenden Instrumente und des Mediums Video zur Messung von Unterrichtsqualität werden skeptisch beurteilt. Offenbar können Videos die Variabilität von Unterrichtsqualität nur bedingt einholen. Die zweite Studie (Kapitel acht) verweist auf die vergleichbare Reliabilität der Ratings von Klassenführung und Schülerorientierung zwischen geschulten und ungeschulten Ratern. Zwar fällt der Rater-Bias unter geschulten Beobachtern theoriekonform zumindest teilweise geringer aus, doch lassen die Befunde insgesamt Zweifel an der Effizienz der bislang

praktizierten Rater-Schulungen in der Unterrichtsqualitätsforschung aufkommen. Mit der dritten Studie (Kapitel neun) lassen sich quantitativ eine hohe zeitliche Stabilität und damit ein geringer Einfluss situativer Merkmale auf die Ratings konstatieren. Gleichwohl zeigt sich anhand einer qualitativen Inhaltsanalyse eine Diskrepanz zwischen den zeitlich stabilen Beurteilungen einerseits und deren instabilen Begründungen andererseits, die unter anderem darauf verweisen könnte, dass die Rater nicht alle für die Urteilsfindung relevanten Aspekte verbalisieren können und die implizite Logik der Urteilsfindung daher schwer zugänglich ist. Den Urteilsprozessen im Kontext hoch-inferenter Ratings geht die vierte Studie (Kapitel zehn) nach. Unter anderem mittels der Methode des lauten Denkens zeigt sich, dass entlang des gesamten Beurteilungsprozesses Probleme auftreten und es weiterer (experimenteller) Studien bedarf, um deren Relevanz für die Forschung quantifizieren zu können. Die fünfte Studie (Kapitel elf) zeigt die Abhängigkeit der zeitlichen Stabilität der Ratings von den Dimensionen der Unterrichtsqualität. Während Klassenführung und Schülerorientierung über mehrere Unterrichtsstunden hinweg sehr stabil beurteilt werden können, gelingt dies bei der kognitiven Aktivierung nur sehr bedingt.

Die hier stark verkürzte Darstellung einiger zentraler Befunde mündet im Text in einer Gesamtdiskussion (Kapitel zwölf). Die Eignung externer Beobachter zur Erfassung von Unterrichtsqualität wird aufgrund von Messproblemen, welche die Reliabilität und Validität der Ratings generell in Frage stellen, kritisch beurteilt. Die Güte der Ratings hängt nicht nur von unterschiedlichen Beobachterperspektiven, sondern auch von methodischen Erwägungen ab. Es seien daher in der Forschung die Qualität der Ratings anzugeben und qualifizierte Rater einzusetzen – gute Trainingskonzepte und Rater-Manuale vorausgesetzt.

Implikationen der Arbeit sind weitreichend und unter Umständen folgeschwer für die (videografische) Forschung zur Unterrichtsqualität. So kann die generelle Anfrage an die gegenwärtige Forschungspraxis gestellt werden, ob die teuren und offenbar nur begrenzt zuverlässigen hoch-inferenten Ratings eine akzeptable Effizienz aufweisen und ob sie letztlich zu Ergebnissen führen, die ihren Vorzug gegenüber alternativen Verfahren wie der Lehrer- und / oder Schülerbefragung rechtfertigen. Jedenfalls entmythifiziert die Arbeit die zuweilen als „Königsweg“ oder „state of the art“ titulierte videografische Forschung.

Der Band mündet in einer vielschichtigen Kritik an der für die Beurteilung der Qualität von Ratings insgesamt nicht hinreichenden Bestimmung von Reliabilität. Die Autorin betritt mit ihren Analysen ein weitgehend neues Terrain. Es ist insofern angemessen, den Zustand gegenwärtiger Rating-Praxis offenzulegen und die mit hoch-inferenten Ratings verbundenen Limitationen aufzuzeigen, ohne konkrete Hinweise oder gar eine methodische Anleitung für deren Behebung geben zu wollen.

Wenn an abschließender Stelle prominent auf die Kontext- und Situationsabhängigkeit unterrichtlicher Prozesse hingewiesen und diese als problematisch für hoch-inferente Ratings exponiert wird, wären allerdings Hinweise auf nicht-quantifizierende Forschungsansätze, etwa jene der strukturtheoretisch-rekonstruktiven oder ethnografischen Unterrichtsforschung wünschenswert gewesen. Diese verdeutlichen eindrücklich die Grenzen der Festlegung unterrichtlichen Handelns auf zu enge Muster und zu eindeutige Interpretationen. Dies gilt mehr für die metatheoretische Einordnung und Abgrenzung der Arbeit – für die konkreten

Analysen haben paradigmenfremde Überlegungen naturgemäß keinen nennenswerten Mehrwert. Würdigend hervorgehoben werden muss hier auch die Berücksichtigung verschiedener qualitativ-inspirierter Methoden, welche die quantitativen Analysen erweitern und teils zu divergierenden Ergebnissen führen.

Der Band ist trotz der umfangreichen statistischen Darlegungen gut lesbar. Eine gewinnbringende Lektüre setzt allerdings ein fortgeschrittenes Verständnis empirischer Sozialforschung voraus, wenngleich die Ergebnisse, Diskussionen und Zusammenfassungen auch von einem breiteren Fachpublikum erschlossen werden können. Schließlich sind die knappen, konsistenten und kundigen Einführungen in die jeweiligen theoretischen Vorarbeiten zu erwähnen, deren Lektüre unabhängig von den konkreten Befunden ein Gewinn für alle mit hoch-inferenten Ratings konfrontierten Forschenden sein dürfte. Der Band wendet sich insgesamt ausschließlich an einen wissenschaftlichen Leserkreis, ein für eine Dissertation angemessener Anspruch.

Colin Cramer (Tübingen)