Educational Assessment, Evaluation and Accountability (2018) 30:343–346
https://doi.org/10.1007/s11092-018-9290-8

Reliability, validity and fairness—key issues in assessing the quality of teaching, instructional leadership and school practice

Guri Skedsmo [1,2] · Stephan Gerhard Huber [1]

Published online: 9 November 2018

Many attempts have aimed at measuring the quality of teachers and teaching; recently, the number of research studies investigating different models and approaches and their intended and unintended consequences has increased significantly. Models and methods of teacher evaluation, including value-added models (VAMs), have been highly criticised, especially in the US context. Another external method employed to investigate and report on educational quality is school inspection, which often includes observations of teaching quality. School inspection procedures have been increasingly characterised by audit control in many countries (Pollitt and Bouckaert 2004). In general, externally produced data are trusted as an objective source to form the basis for quality development in educational institutions (Lingard et al. 2013; Ozga 2012).

In this issue, we follow up on the topic of fair and valid teaching evaluation in the first two articles, where the authors present findings from studies examining various methods and measures used for evaluating teaching quality. The third article considers factors influencing teachers' perceptions of feedback from school inspections, while the fourth reports on a study of multiple-rater techniques, where patterns of principals' and teachers' ratings of instructional leadership are tested, and the authors challenge the application of the common understanding of principals' and teachers' rating behaviour that is often presented in school effectiveness research in the Western world.

✉ Guri Skedsmo
guri.skedsmo@ils.uio.no

✉ Stephan Gerhard Huber
stephan.huber@phzg.ch

[1] Institute for the Management and Economics of Education, University of Teacher Education Zug, Zug, Switzerland

[2] Department of Teacher Education and School Research, University of Oslo, Oslo, Norway

1 Articles in this issue of EAEA 4/2018

In the first article, van der Lans reports on a study in the Dutch secondary education system in which he examined the association between two measures of teacher effectiveness, namely classroom observation and student survey measures, and the extent to which this association depends on the study design. The two measures differ in their exact item content, but they aim to operationalise the same latent construct. In his analysis, the author tests and compares different models. The findings show that the correlations between the survey and observation measures depend on three factors, as follows: the number of classroom observations, the number of student ratings and whether the designs are nested or partially nested. Several implications can be drawn from this study. For example, it illustrates how the importance of conducting multiple observations of multiple lessons can increase the reliability and effect size. Moreover, the difference in reported correlations when using completely and partially nested designs indicates the need for more precise descriptions of study designs that use classroom observation or survey measures.

In the second article, Sloat, Amrein-Beardsley and Holloway examine the consistency of teacher-effectiveness ratings resulting from six VAM approaches. The approaches include the following: (1) the student growth percentile (SGP) model; (2) value-added linear regression model (VALRM); (3) value-added hierarchical linear model (VAHLM); (4) simple different (gain) score model; (5) rubric-based performance level (growth) model; and, finally, (6) simple criterion (per cent passing) model. The approaches were tested in a large suburban school district in Arizona, USA. The authors' analyses, which are based on the various approaches, show that some methods more closely correspond to the SGP model rating than they do to the other evaluations. However, there is substantial variety in the teacher-level rating assignments. Statistically and substantively, these assignments depend on the approach or model that is adopted to evaluate teachers' influence on their students' growth in achievement over time. The authors raise questions about whether local autonomy should be given to districts to choose the models. They also point out possible implications of the study related to the geographical area, and they suggest that the choice of model may determine the evaluation of teacher effectiveness.

In the third article, Quintelier, Vanhoof and De Maeyer report on a qualitative study investigating teachers' cognitive and affective responses to feedback from school inspection in the Flemish education context in Belgium. Furthermore, they consider how these responses lead to feedback acceptance or rejection during a school inspection process. The process includes three phases; specifically, there are preliminary inquiry and audit, resulting in an inspection reporting on the school's strengths and weaknesses, and finally, a quality judgement according to the categories of 'positive', 'restricted positive' and 'negative'. However, the authors report that teachers receive limited feedback in the debriefings after the teaching observations. They find that teachers who distrust the inspectors' expertise and trustworthiness are more likely to display an unreceptive reaction to negative feedback. Moreover, unjust feedback or feedback that fall short of the teachers' judgements tends to affect teachers' perceptions negatively, while feedback that exceeds their judgement leads to relief and happiness, even if it is negatively formulated.

In the fourth article, Guo and Lu demonstrate how perceptions of power distance help explain the variation in the patterns of principal–teacher rating differences on

instructional leadership. This quantitative study was conducted in a central province, Henan, in China. Earlier, this province implemented a new curriculum reform introducing new instructional requirements. Based on a literature review on the application of multiple-rater techniques, the authors show that self–other rating agreement has been used to examine the effectiveness of leadership and its relevant outcomes. Moreover, rating gaps are often reported when leaders' self-ratings disagree with their subordinates' evaluations. This is also the case in studies of instructional leadership where the aim is to better understand the real-world instructional leadership performance. However, most of these studies have been conducted in the USA. Due to the more hierarchical administrative structures in Chinese schools and general higher scores on the power distance index in China, the authors set out to test the extent to which principal and teacher perceptions of instructional leadership would converge. While most Western studies report that principals' ratings of instructional leadership tend to be higher than those given by the teachers, the Chinese results generally exhibit smaller rating gaps. When the principals reported a low power distance, their ratings on instructional leadership were higher than those from the teachers. In contrast, when the principals reported a high power distance, their ratings on instructional leadership were lower than the ratings from the teachers were. The findings have several implications. One is related to the interpretation of principal–teacher ratings in terms of effective leadership.

2 Reflections on perspectives in the contributions

At least three interesting themes arise from the four articles included in this issue. The first relates to the reliability, validity and fairness of models and methods aiming to evaluate teaching quality. The second theme concerns evocative aspects of assessment and evaluation. Finally, the third theme highlights cultural awareness in educational research.

Regarding the first theme on the reliability, validity and fairness of models and methods aiming to evaluate teaching quality while considering design, van der Lans' findings from the Netherlands confirm those of other studies (e.g. Lei et al. 2018; Mintrop et al. 2018; Santelices et al. 2017), in that greater attention has to be paid to the design of such models to provide reliable data and a fair evaluation. The analysis across six value-added models by Sloat et al. in the US context demonstrates validity problems, and the authors emphasise that, although the VAM data may be statistically sophisticated, a range of facts, such as the type of tests used, policy contexts and geographical locations, will affect how VAMs play out in practice. Such problems were confirmed in a recent review of VA research by Everson (2017); this author points out that our knowledge about how well VAMs estimate teacher or school contributions to student achievement is still evolving. Even if research has offered insights, conclusions need to be validated using different populations and testing instruments. Sloat et al. argue that there is an imperative to offer counterarguments; specifically, more data and increased accountability policies and initiatives will offer solutions for improving educational quality and help work against inequity in education.

Teachers' cognitive and affective responses to feedback from school inspection, which are at the centre of the third article, can be associated with Dubnick's (2007)

distinction between 'accountability as evocative' and 'accountability as performance'. Responses to performative aspects are often cognitive in character, and they can result in an intentional act carried out to improve practice or accomplish a result. However, evocative elements depend on how accountability plays out in practice and how it is understood or interpreted, which sometimes contradicts formal intentions. In other words, attention needs to be paid to the evocative aspects of accountability practices.

The fourth article, that by Guo and Lu on how power distance plays out in the Chinese context, raises an important theme not only for research in education but also for the social sciences in general. It is essential to underline the need for other country perspectives to balance and nuance the knowledge that has been predominantly generated from studies in Western countries, especially the USA.

References

Dubnick, M. J. (2007). Situating accountability: Seeking salvation for the Core concept of modern governance. University of New Hampshire.

Everson, K. C. (2017). Value-added modeling and educational accountability: Are we answering the real questions. *Review of Educational Research, 87*(1), 35–70.

Lei, X., Li, H., & Leroux, A. J. (2018). Does a teacher's classroom observation rating vary across multiple classrooms? *Educational Assessment Evaluation and Accountability, 30*(1), 27–46. https://doi.org/10.1007/s11092-017-9269-x.

Lingard, B., Martino, W., & Rezai-Rashti, G. (2013). Testing regimes, accountabilities and educational policy: Commensurate global and national developments. *Journal of Education Policy, 28*(5), 539–556.

Mintrop, R., Pryor, L., & Ordenes, M. (2018). A complex adaptive system approach to evaluation: Application to a pay-for-performance program in the USA. *Educational Assessment Evaluation and Accountability, 30*(3), 285–312. https://doi.org/10.1007/s11092-018-9276-6.

Ozga, J. (2012). Governing knowledge: Data, inspection and education policy in Europe. *Globalisation, Societies and Education, 10*(4), 439–455. https://doi.org/10.1080/14767724.2012.735148.

Pollitt, C., & Bouckaert, G. (2004). *Public management reform. A comparative analysis (2 ed.)*. New York: Oxford University Press.

Santelices, M. V., Valencia, E., Gonzalez, J., & Taut, S. (2017). Two teacher quality measures and the role of context: Evidence from Chile. *Educational Assessment, Evaluation and Accountability, 29*(2), 111–146. https://doi.org/10.1007/s11092-016-9247-8.