Trust in results

Guri Skedsmo [1,2] · Stephan Gerhard Huber [3]

Over recent decades, many countries and states across the world have attempted to measure the effectiveness of their educational systems. The question is whether testing and benchmarking sometimes create their own dynamic by introducing new reforms which, in turn, require more testing and assessment to determine their effects (Baker and LeTendre 2005). Some scholars argue that this increase in governing systems based on measures of educational effectiveness has shifted the focus from a 'trust in the profession' to a 'trust in results' (Uljens et al. 2013). Tied to this shift of focus are various degrees of accountability pressure perhaps linked to incentives and sanctions to boost performance. Educators are expected to seek ways to help improve education where the results are used to give direction and legitimize selected strategies and actions. This issue of EAEA includes four studies of tools and measures of educational quality and effectiveness, and one study of how school principals respond to accountability demands.

1 Articles in this issue of EAEA, 4/2019

In the first article, Reynolds and Candee argue that defining *high quality* in early childhood classrooms is challenging because it relies on assumptions that there is a consensus what children should know and be able to do at kindergarten entrance, while kindergartens should be able to meaningfully identify, measure and support these behaviours and skills. The authors report on a study which assesses the psychometric characteristics of the Classroom Learning Activities Checklist (CLAC), which is a tool

✉ Guri Skedsmo
  guri.skedsmo@phsz.ch

✉ Stephan Gerhard Huber
  stephan.huber@phzg.ch

[1] Institute for Research on Professions and Professional Learning, Schwyz University of Teacher Education, Goldau, Switzerland

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

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

that measures the unique classroom behaviours that support students' self-regulation and task-oriented behaviours. CLAC was tried in 72 classrooms which are part of the Midwest Child-Parent Center. Based on the analysis from 1358 children, the authors tested the tool's dimensionality and validity. Findings show that CLAC measures two dimensions of classroom context: instructional responsiveness and student engagement. Both are independently associated with Pre–K learning gains. Moreover, the authors find that classroom strategies that promote students' early self-control and self-directed behaviour facilitate and sustain children's learning as children move to kindergarten and elementary grades.

In the second article, Hunter examines the relationship between instructional quality and a key component of No Child Left Behind (NCLB) policies in terms of holding schools accountable for student performance on standardized tests. More concretely, he analyses the extent to which NCLB-defined school failure to reach 'adequate yearly progress' in mathematics was associated with subsequent changes in middle-grade mathematics instructional quality. Using data from a large multi-year research project covering several school districts, the author finds that instructional quality improved more in schools that failed math in the prior year and faced severe sanctions than in schools that passed. However, the level of instructional quality in schools that failed was still lacking. Conversely, this was not the case in districts emphasizing the instructional leadership of school principals and teachers' access to a rigorous mathematics curricula as key mechanisms for instructional improvement. The author argues that school underperformance may allow school leaders to create urgency regarding the need for instructional improvement and thereby mobilize teachers for change. However, focusing on underperformance and pressuring schools to improve are not enough.

In the third article, Guarino, Stacy and Wooldridge compare two approaches to measuring school effectiveness. The first approach they call *beating the odds* (BTO), which compares predicted and actual school effectiveness scores based on a cross-sectional analysis to determine whether a school 'beats' its expected performance. In contrast, the second value-added approach (VAM) contains a type of control of prior performance on the student level. Generally, authors find that the VAM approach provides a more defensible measure of school effectiveness, but they note that both approaches have advantages and disadvantages, which they highlight as important to consider when policymakers choose a methodology for measuring school effectiveness. The authors argue that several general principles uncovered in this side-by-side comparison also are relevant to other approaches.

In the fourth article, Amrein-Beardsley and Geiger investigate potential sources of invalidity when using teacher value-added and principal observational estimates. According to the authors, statisticians construct and use value-added models (VAMs) and growth models to measure the predicted and actual 'value' a teacher 'adds' to student achievement from 1 year to the next. This is typically done by measuring student growth on large-scale standardized tests (e.g. as mandated throughout the U.S. by the No Child Left Behind [NCLB] Act, 2001) and aggregating this growth at the teacher level, while statistically controlling for confounding variables such as students' prior test scores and other student-level and school-level variables. However, as the authors point out, control variables vary by model. The authors illustrate in their analysis how those with the power to evaluate teachers (e.g. principals) within such contemporary evaluation systems might (1) artificially inflate or (2) artificially deflate observational

estimates when used alongside their value-added counterparts or (3) artificially conflate both estimates to purposefully exaggerate perceptions of validity.

In the final article, Qian and Walker explore the role school principals play in managing the intersection of external and internal accountability systems within Chinese schools. Chinese school leaders' work environment is, on one hand, highly political, which implies being conscious of their role as state employees and their dependence on various government agencies. On the other hand, they also must demonstrate expert knowledge and thereby gain legitimacy from teachers to produce better performance in schools, which is necessary to receive recognition from superiors in terms of being rewarded by better assignments. The authors demonstrate how school principals work on strengthening internal accountability by focusing on building mutually supportive and trusting relationships – also called *paternalistic leadership* – as a way of responding to external accountability demands.

2 Some reflections

Key topics across the five contributions in this issue are measures of educational effectiveness, pressure and accountability demands. Several studies have shown that differences in academic performance begin at an early stage and persist over time. Increasing attention is, therefore, being paid to measuring the quality of early childhood learning and development, from independent researchers to governments and international organizations. As Reynolds and Candee demonstrate in the first article in this issue, it is important to investigate instruments developed and implemented to measure quality. However, as Goldstein and Flake (2016) argue, if quality assessments are created to improve outcomes for young children, evidence should be required to show how data are used to improve programmes, which means that attention also must be paid to the role that quality assessments play in serving young children.

Regarding the article by Qian and Walker, it would also be interesting for further research to explore how teachers respond to internal accountability demands in an environment where paternalistic leadership is applied to ensure commitment and where mutual dependencies seem to be a key element in internal accountability systems. Moreover, Hunter's article demonstrates that the label of *underperforming* results in sanctions on the school, which can represent a 'wake-up call' and mobilize key actors, such as teacher and school leaders for change. However, to develop instructional quality sustainably, input-oriented investments are needed.

Both the articles by Guarino et al. and Amrein-Beardsley and Geiger in this issue concern measures of educational effectiveness. Guarino et al. compare BTO and VAM approaches to assess the overall effectiveness of educational systems and seem to favour value-added measures for this purpose. However, Amrein-Beardley and Geiger address critical issues concerning how VAM is used to evaluate teachers' work and hold them to account. A key question raised by Amrein-Beardsley and Geiger is what we do when we are presented with contradictory information, for example one measure suggesting high teacher quality and the other indicating low teacher quality. According to the authors, the observational estimate would typically be the higher one, and the value-added estimate, the lower one. The latter is often assumed to be the stronger and more exact and valid and would, therefore, often be trusted to trump the other. Hence,

more research and critical investigations into these issues are needed, as well as concerning what the estimates signify and how they are interpreted and used in practice.

References

Baker, D. P., & LeTendre, G. (2005). *Global similarities and national differences*. Stanford: Stanford University Press.

Goldstein, J., & Flake, J. K. (2016). Towards a framework for the validation of early childhood assessment systems. *Educational Assessment, Evaluation and Accountability, 28*(3), 273–293.

Uljens, M., Møller, J., Ärlestig, H., & Frederiksen, L. F. (2013). The professionalisation of Nordic school leadership. In L. Moos (Ed.), *Transnational influences on values and practices in Nordic educational leadership: Is there a Nordic model?* (pp. 133–157). Dordrecht: Springer Netherlands.