Your full name: Armin Jentsch

Affiliated authors with institutions:

Affiliation: University of Greifswald

Current position: Postdoctoral researcher

Investigating generalizability interpretations for an instructional quality framework in secondary mathematics classrooms

Abstract (300 words)

The study presents research from the projects TEDS-Instruct and TEDS-Validate. Based on the established conceptual framework with Three Basic Dimensions (TBD), we have developed an observational instrument to measure instructional quality in secondary mathematics classrooms. Assuming equivalence of the measures across two German samples, we investigate into the validity of our assumptions in the sense of Kane (2013) by combining measurement invariance and generalizability analysis. To explore the psychometric properties of the observational instrument, we collected data from secondary mathematics classrooms (*N* = 76) in different parts of Germany and observed two lessons per classroom (90 minutes each). Four ratings per lesson were performed by extensively trained observers. Instructional quality was rated using an observational instrument with three generic dimensions, as well as two subject-specific dimensions (classroom management, student support, cognitive activation, subject-related quality, teaching-related quality). The results indicate similar levels of measurement invariance for all dimensions but cognitive activation. Additional analyses show that lack of measurement invariance could be due to measurement error, but more interestingly, cognitive activation items did not load on the same factor either. In the presentation, we show how a critical reception of the findings has led the developers to a simpler framework of instructional quality which can still be regarded as an enhancement of TBD. In general, however, the observational instrument translated well to a different educational context, and substantial amounts of variability in instructional quality were due to differences between classrooms, as expected. As an outlook to future research, we discuss how our findings could inform the development of other observational instruments involving both generic and subject-specific dimensions.

Extended summary (1000 words, excluding reference list) Include introduction, theoretical background, methods, aims, preliminary findings/findings, results, theoretical and education significance, relevance to the QUINT ambition and the reference list.

The study presents research from the projects Teacher Education and Development Study—Instruct (TEDS-Instruct) and TEDS-Validate. Based on a theoretical framework of instructional quality, we developed an observational instrument to measure generic and subject-specific dimensions in two samples of German secondary mathematics classrooms. Kane (2013) points out that generalizing from the findings of a study to a different educational context represents an interpretation of test scores that should be grounded in empirical evidence. Accordingly, exploring the generalizability of test scores is a major exercise in validity research. This study aims at investigating this "generalizability interpretation" (Kane, 2013).

Conceptual framework

Following the TIMSS video study, German educational research brought up a generic framework of instructional quality with three basic dimensions (TBD, Praetorius, Klieme, Herbert, & Pinger, 2018). They are *efficient classroom management*, potential for *cognitive activation* and *student support.* By classroom management, we refer to such procedures or strategies that enable efficient use of time to study. Learning opportunities in which students are confronted with intellectual challenge or higher-order thinking (e.g., problem solving) are considered to have potential for cognitive activation. Finally, student support refers to a positive classroom climate that satisfies students' needs for self-determination. It should be noted that the TBD have substantial overlap with the likewise three-dimensional Classroom Assessment Scoring System (CLASS). The generic dimensions of teaching quality described by these frameworks are widely regarded to have a positive influence on students' achievements in the classroom. However, in a literature review, Charalambous and Praetorius (2018) gather empirical evidence that subject-specific and generic dimensions explain more variability in students' learning outcomes than generic dimensions alone. They argue that it can be assumed that generic and subject-specific measures can provide additional value over and above each other (see also Berlin & Cohen, 2020). For this reason, the present study makes use of an enhanced version of the TBD. By conducting a literature review (Schlesinger & Jentsch, 2016), we have developed two additional subject-specific dimensions of instructional quality. While subject-related quality addresses the subject matter that is discussed in class (e.g., mathematical correctness or rigor), teaching-related quality focuses on the implementation of the mathematical tasks and material (e.g., by using various representations of the content).

Method

We collected data from 76 teachers in years 7 to 10 and observed two lessons per teacher. The lessons lasted 90 minutes each and were spread over approximately two weeks. Four ratings per lesson were performed by extensively trained observers. Instructional quality was rated using an observational instrument with three generic dimensions, as well as two subject-specific dimensions, with a total of 26 items which were rated on a four-point rating scale (classroom management: four items, e.g., use of time, α = .82/.80, student support: five items, e.g., collaboration, α = .66/.68,

cognitive activation: four items, e.g., co-construction, α = .76/.59, subject-related quality: four items, e.g., dealing with students' errors, α = .75/.56, teaching-related quality: four items, e.g., use of representations, α = .67/.65). Rater agreement was adequate across all items (ICC > .70).

We carried out analysis of measurement invariance using a multi-group approach to find out whether the instructional quality observer ratings conducted in TEDS-Instruct and TEDS-Validate lead to similar findings in two different German educational contexts. To gain further insights into the causes of potential differences, we performed additional generalizability studies on the item level (G studies, Cronbach et al., 1972). We used a design in which the total variability of scores is decomposed into classroom, lesson, segment, and rater effects.

Results

For classroom management, student support and the two subject-specific dimensions, at least partial scalar measurement invariance was found. For cognitive activation, however, there was a lack of measurement invariance already on the configural level. This means that the observer ratings did not result in a similar meaning of cognitive activation across the two samples. An additional exploratory factor analysis was carried out and resulted in a two-dimensional structure for cognitive activation in TEDS-Validate. As for the generalizability analysis, substantial amounts of variance could be explained by variability due to differences between teachers or classrooms. Overall, subject-specific items showed larger amounts of lesson variability, whereas generic items varied more within lessons, e.g., teachers employed more variation in how they supported students during lessons than across different lessons. Although the difference across studies were small in general, some cognitive activation and subject-related quality items suffered from low reliability in TEDS-Validate. This could also provide some explanation for the lack in measurement invariance.

Discussion

The results indicate similar levels of measurement invariance for all dimensions but cognitive activation. Additional analyses have shown that cognitive activation items did not load on the same factor either. In other words, the assumption of measurement equivalence across two samples of mathematics classrooms was violated, such that our measure of cognitive activation could not be used equivalently in different, but similar educational contexts. In the presentation, we discuss how a critical reception of the findings has led the developers to a simpler framework of instructional quality which can still be regarded as an enhancement of TBD. In our study, we also found considerable differences in generalizability between subject-specific and generic dimensions. We argue that these findings could be relevant when it comes to the development of future observational instruments, both in research and practice. A lack of generalizability will have an impact both on estimating relations with other variables (e.g., student achievement) and also on decision-making processes in instructional practice that involve observer ratings.

References

Berlin, R. & Cohen, J. C. (2020). The Convergence of Emotionally Supportive Learning Environments and College and Career Ready Mathematical Engagement in Upper Elementary Classrooms. *AERA Open, 6*(3), 1-20.

Charalambous, C., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM Mathematics Education, 50*(3), 355-366.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic Dimensions of Teaching Quality: The German Framework of Three Basic Dimensions. *ZDM Mathematics Education, 50*(3), 407-426.

Schlesinger, L. & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education, 48*(1), 29-40.