



HAL
open science

Measuring instructional quality in mathematics education

Armin Jentsch, Lena Schlesinger

► **To cite this version:**

Armin Jentsch, Lena Schlesinger. Measuring instructional quality in mathematics education. CERME 10, Feb 2017, Dublin, Ireland. hal-01949106

HAL Id: hal-01949106

<https://hal.science/hal-01949106v1>

Submitted on 9 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Measuring instructional quality in mathematics education

Armin Jentsch¹ and Lena Schlesinger¹

¹University of Hamburg, Germany; armin.jentsch@uni-hamburg.de

Instructional research has recently become more important within the framework of teaching effectiveness research. Various instruments have been developed within this research discipline in order to gain a better insight of what is really happening in the classroom. Most of these instruments mainly focus on generic aspects of instructional quality. In this paper we first describe a subject-specific dimension of instructional quality. Second, we show how these subject-specific aspects could be measured empirically with a standardized observational instrument. The results point out both good interrater agreement and satisfying reliability measures. The presented observational instrument has been developed within the study TEDS-instruct in which relations to teachers' competencies and students' achievement are analyzed.

Keywords: Mathematics instruction, instructional quality, classroom observation techniques, instructional improvement.

Introduction

In German instructional research, subject-specific aspects of instructional quality have been investigated rarely until now. Three generic dimensions have been introduced which are classroom management, personal learning support and cognitive activation (e.g. Lipowsky et al., 2009). Although these basic dimensions were developed for mathematics instruction in the first place, they are now deemed relevant for every subject at school (Baumert et al., 2010; Helmke, 2012). Other important aspects of mathematics instruction were apparently disregarded then (e.g. representations, examples, modelling, proof). Blum and others (2006) therefore ask for a high *mathematical* quality of the lesson beyond the three basic dimensions.

At the same time, several instruments for measuring subject-specific aspects of instructional quality have been developed within the American debate (e.g. Learning Mathematics for Teaching Project, 2011). However, some of these instruments do not contain generic aspects even though the prognostic validity of the three basic dimensions of instructional quality has been shown empirically more than once (see Baumert et al., 2010; Kunter et al., 2013; Lipowsky et al., 2009). In conclusion: to our knowledge there is no standardized instrument in existence that is based on a sound theoretical framework (e.g. the three basic dimensions) covering also subject-specific aspects regardless of the mathematical content discussed in class. Starting from these three basic dimensions, we would like to point out subject-specific characteristics of instructional quality that are deemed relevant in the literature and show how these characteristics can be measured with a standardized instrument. The purpose of this paper then is to describe the development of an observational protocol that is used to assess instructional quality in secondary mathematics classes and to present first empirical results.

Theoretical framework

In recent years educational research has shown a great scientific interest in teacher knowledge and instruction (see Hattie, 2012). At the same time, the relation of teacher competencies and students'

achievement has been analyzed (Baumert et al., 2010; Hill, Rowan & Ball, 2005; Scheerens, 2004). The so-called process-mediation-product-paradigm is regarded as a theoretical framework in which these research questions are grounded. This framework describes a relation between students' learning and instruction which has to be *offered* by teachers and *used* by students (Brophy, 2000; Brophy, 2006; Helmke, 2012; Oser, Dick & Patry, 1992).

Three basic dimensions of instructional quality

As mentioned before, three generic dimensions of teaching quality have been developed in recent instructional research which are classroom management, personal learning support and cognitive activation. These dimensions were shown to have a positive impact on both students' learning and their motivation in class (Baumert et al., 2010; Lipowsky et al., 2009).

Classroom management focuses on quality-oriented learning time provided for students and on how effectively the teacher deals with disciplinary conflicts (Brophy, 2000). Effective classroom management is also characterized by a lesson that is organized well and has clear routines (Lipowsky et al., 2009). The second dimension, *personal learning support*, includes the individual support provided by the teacher, the relationship between students and teacher as well as constructive feedback (e.g. Rakoczy, 2008). Finally, *cognitive activation* refers to how problem-solving tasks are used to activate learning processes (Baumert et al., 2010; Brophy, 2000). This dimension includes the activation of previous knowledge and whether challenging tasks and questions are presented that foster students in high-level thinking activities (Lipowsky et al., 2009; Praetorius et al., 2014).

Although the three basic dimensions focus on generic aspects of instructional quality, the question remains whether they could be operationalized in subject-specific way. This holds specifically for cognitive activation (Drollinger-Vetter, 2011; Schlesinger & Jentsch, 2016). Moreover, it is not clear which subject-specific aspects are missing in this framework and how its dimensionality changes when generic and subject-specific aspects are measured simultaneously (Drollinger-Vetter, 2011). Due to these concerns, most instruments that have been developed for the analysis of instruction are only suitable for a very small number of situations (e.g. regarding the mathematical content, see Steinweg, 2011; Schoenfeld, 2013).

Measuring instructional quality

We will now focus on the question how instructional quality can be measured both reliably and validly. Praetorius and colleagues (2012) see classroom observations as a straight-forward way to do so, especially compared to the analyses of material or minutes conducted during the lesson. Helmke (2012) even claims that one can only speak of instructional research in a narrower sense when classroom observations are performed. The reliability of classroom observations is always an issue since observer ratings are often heavily biased and the stability of the measurement is sometimes questioned (for an overview see Praetorius, 2014). Most authors suggest analyzing variance components beyond measures of interrater agreement to understand better which sources of error have added to the variance of the observed score (e.g. Praetorius et al., 2012; Praetorius et al., 2014).

Observational instruments may contain both low and high inference items (Praetorius, 2014). Codings with low inference are operationalized in a way that is strictly observable. High inference

items, in comparison, need the observer to interpret what he or she sees which makes the observation much more complex (Hugener, 2006). However, at the same time one gains a higher validity because instructional research has shown that low inference items explain only little variance when students' achievements are measured as outcome variables (e.g. Baumert et al., 2010). This is because the surface structure of instruction (e.g. which method is used by the teacher) and its quality may sometimes vary independently from each other (Kunter & Voss, 2013).

Developing an instrument for measuring instructional quality

The observational instrument that is presented in this paper was developed within the study of TEDS-Instruct which is a Follow-Up Study of TEDS-M (Teacher Education and Development Study in Mathematics). The main goal is to empirically investigate teachers' competencies in mathematics education at the secondary level and their influence on students achievement mediated by instructional quality. As a matter of fact, students' achievements will be collected to describe the prognostic validity of both teachers' competency tests and the observational protocol that is presented here.

Subject-specific aspects of instructional quality

For developing a subject-specific dimension of instructional quality we first analyzed which subject-specific aspects of instruction are assumed to have an impact on students' learning which has to be examined empirically. The main goal for developing a fourth dimension with subject-specific aspects was hence to extend the existing generic theoretical framework of instructional quality. Such an extension with subject-specific aspects is not established until now (e.g. Steinweg, 2011). During the development of the fourth dimension it became apparent that it is necessary to discuss the subject-specificity of some aspects that are included within the former three-dimensional generic framework. This discussion leads to the assumption that the three basic dimensions of instructional quality are not completely generic. However, in the fourth dimension there were included only such subject-specific aspects that were not already used to operationalize the other three dimensions.

For conceptualizing this dimension, a systematic literature survey within the databases of Web of Science, ERA and ERIC was conducted (see Schlesinger & Jentsch, 2016 for more detail). At the same time, the national debate on mathematics education and the German common core standards was reviewed. Based on the described approach, the following aspects were operationalized for the observational instrument:

- the teacher's mathematical correctness
- the use of representations
- mathematical competencies (modelling, problem-solving, the use of mathematical language, argumentation and proof, training mathematical tools and operations)
- a constructive approach to students' mathematical errors
- the quality of exercises and tasks
- sense-making
- teachers' mathematical explanations
- appropriate examples
- mathematical depth (e.g. generalizations)

Method

Based on the three basic dimensions of instructional quality and the subject-specific aspects that were condensed into a fourth dimension, we developed an observational instrument that can be used for in vivo ratings without needing videos of the lesson and that can be utilised independently from both the specific mathematical content and the academic year. Instructional quality was rated by assessing the items on a four-point scale (1=“not at all true“; 4=“completely true“). The instrument consists of four dimensions which are classroom management (five items), personal learning support (seven items), cognitive activation (five items) and mathematics educational quality of instruction (nine items). The data for TEDS-Instruct was collected in Hamburg from a sample of 38 teachers at the secondary level. The teachers participated on a voluntary basis. Therefore it can be assumed that they were greatly motivated to have their lessons observed. Each teacher was observed for two lessons (90 min each). Within one lesson, the instructional quality was assessed four times (every 22.5 min).

Example items	Indicators
Mathematical depth	<ul style="list-style-type: none"> • The teacher provides generalizations • The teacher provides mathematical connections • The teacher deepens and structures mathematical knowledge
Representations	<ul style="list-style-type: none"> • The teacher provides various representations for mathematical objects • The teacher illustrates the linking between different representations

Table 1: Two example items for the subject-specific dimension

Altogether, there were six observers involved in the classroom study, all of which held at least a university degree from a mathematics teacher program. The observers were trained for the classroom observations in advance which took around 20 hours. The training had three main goals: 1) a joint understanding of the theoretical underpinnings of each rating dimension, 2) familiarizing with the observational protocol, 3) ensuring a satisfactory amount of interrater reliability. By doing so, all items and indicators were discussed thoroughly with the help of a rating manual. The goal of a joint theoretical understanding also involved the *object of measurement*, i.e. instructional quality. Based on the process-mediation-product-paradigm, instruction is regarded as a learning opportunity that is individually adapted to students' skills and dispositions. Even though the focus of the observations lies mainly on the teachers' behaviour, the latter is dependent of the students' behaviour and student-teacher-interactions. Students' reactions to the learning opportunities are crucial to understand and assess the *quality* of instruction and are hence part of the observation, too. Nonetheless, due to pragmatic reasons no student self-reports or cognitive tests have been collected.

Before stepping into real classrooms, the observers trained their skills on videotaped lessons until they reached a certain amount of interrater agreement. Finally, a pilot study was conducted with 13 teachers in three German federal states. After each observation, the ratings were discussed intensively between the two raters. For the data collection the lessons were observed directly without using videotapes, i.e. the raters assessed all items within the lesson in vivo. Two raters were chosen randomly and rated the lesson independently from each other. For this reason it was possible to avoid systematical agreement between certain raters. In addition to these ratings, there were also produced minutes for every lesson. These minutes included teaching methods, teacher-student-

interactions, students' behaviour and reactions, the mathematical content and provided materials and tasks from the lessons for a detailed description of the learning opportunities.

Results

As a first step we calculated descriptive statistics for the data that was collected by external observers. The following table contains the results of all items from the three basic dimensions and the newly developed subject-specific dimension. For ensuring interrater reliability *Spearman's* ρ was calculated. This is a common measure since in educational research one is more often interested in relative than in absolute decisions (Praetorius et al., 2012; see also Shavelson & Webb, 1991). In the present study we reached satisfying results of $.75 \leq \rho_{inter} < .97$ which can be interpreted as high or very high correlations between both observer ratings. In order to calculate the descriptive statistics the data was aggregated to a single datum per person ($N = 38$). By doing so, we first took the average rating of both observers and then calculated the mean of all eight measurement points per teacher.

Items	<i>M</i>	<i>SD</i>	<i>r_{it}</i>
Classroom management ($\alpha = .83$)			
Effective use of lesson time	3.58	.33	.59
Clear rules and routines	2.97	.19	.66
Preventing disruptions	3.39	.45	.83
Advance organization	2.89	.49	.55
Working atmosphere	3.23	.51	.77
Personal learning support ($\alpha = .714$)			
Students' individual support	2.05	.45	.37
Approach to heterogeneity/differentiation	1.26	.38	.64
Self-regulated learning	1.48	.35	.58
Teacher's feedback	3.07	.37	.49
Teacher approval	3.10	.38	.49
Students' feedback	1.05	.11	.22
Fostering cooperative learning	1.75	.46	.27
Cognitive activation ($\alpha = .821$)			
Challenging tasks and questions	2.54	.47	.79
Supporting metacognition	1.25	.29	.42
Activating prior knowledge and co-construction	2.66	.37	.76
Quality of teaching methods	2.81	.41	.66
Securing knowledge	2.43	.48	.50
Mathematics educational characteristics ($\alpha = .820$)			
Constructive approach to students' errors	2.79	.56	.69
Teacher's mathematical correctness	3.64	.37	.54
Representations	2.29	.65	.39
Exercises and tasks	2.37	.52	.63
Examples	2.99	.42	.54

Mathematical competencies	1.62	.15	.44
Sense-making	2.09	.49	.32
Teacher's explanations	2.93	.54	.62
Mathematical depth	2.34	.40	.69

Table 2: Descriptive statistics for all items

When looking at the measures in table 2, we see that correctness has reached the highest values in the subject-specific dimension. Even though this could be seen as a ceiling effect, the statistical discrimination is quite high. The same holds for the items in the first dimension. Nonetheless, these ceiling effects are not surprising as the sample consisted of professional teachers only (Baumert et al., 2010; Blömeke et al., 2010). On the other hand, the average individual support that was observed in the lessons is quite low which is also supported by the low measures of the items “Self-regulated learning” and “Differentiation”. Finally, the standard deviation of most items is high enough to conclude that a decent amount of variance was measured.

Altogether we can conclude that an acceptable internal consistency could be reached for all four dimensions. When $r_{it} = .25$ is regarded as a threshold for acceptable measures of statistical discrimination, the item “Student’s feedback” did not reach acceptable values and was thus excluded from further analyses which is also due to a floor effect. All other items show at least mediocre correlations to the corresponding scale which supports the claim of three generic dimensions. This is, however, supposed to be confirmed by factor analyses. Recent both exploratory and also confirmatory approaches once again support the hypothesis of three generic dimensions but suggest dividing the subject-specific dimension into two sub-dimensions which will be discussed in more detail in the presentation (Blömeke et al., submitted).

To sum up, this present study has mainly an explorative character concerning the mentioned subject-specific aspects. However, from a more content-related standpoint one can conclude that fostering specific mathematical competencies like modelling or proof has often been disregarded during lessons. Precise analyses of the used material might then be fruitful to understand better what has happened in the classroom.

Discussion and outlook

The presented instrument for measuring instructional quality shall finally be discussed concerning advantages and disadvantages compared to other instruments in the field. Since this instrument has been developed in order to be used in classrooms without analyzing video there is a chance that it could possibly be used in a broader way than instruments from video studies. Second, measuring instructional quality more than once in a given lesson may describe the learning process in more detail and can lead to more reliable data because certain aspects of instructional quality may change a lot during the lesson. The ratings then tend to be biased heavily since the observer has to give a single rating for the whole lesson (Praetorius et al., 2012). Third, the instrument is suitable for most mathematics classes, academic years and mathematical contents. Finally, in this instrument generic and subject-specific aspects are combined which, in addition, can then be analyzed on their relation.

The question remains whether the present instruments’ prognostic validity can be shown by analyzing the relation of instructional quality and students’ achievements. It should be tested whether instructional quality can be seen as a mediator variable between teachers’ competencies and

students' learning, too. This might especially be interesting for mathematics educational scholars since the impact of generic aspects of instructional quality has already been shown in some studies (Baumert et al., 2010; Helmke, 2012; Lipowsky et al., 2009). The important mathematical or mathematics educational aspects of instructional quality and their impact on both learning and other outcome variables as motivation or metacognition have still to be found. Here, our study could help to gain a little more insight.

References

- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal*, 47(1), 133–180.
- Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.) (2010). *TEDS-M 2008 - Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.
- Blömeke, S., Casale, G., Jentsch, A., Schlesinger, L., Hennemann, T., Kaiser, G., & König, J. (2016). *Bringing the content back into instructional quality research: Generalizability of generic and domain-specific observations*. Manuscript submitted for publication.
- Blum, W., Drücke-Noe, C., Hartung, R., & Köller, O. (2006). *Bildungsstandards Mathematik: Konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichts Anregungen, Fortbildungsideen*. Berlin: Cornelsen Scriptor.
- Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 755–780). Mahwah, N.J: Erlbaum.
- Brophy, J. (2000). *Teaching*. Brüssel: International Academy of Education.
- Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann.
- Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.
- Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371–406.
- Hugener, I. (2006). Überblick über die Beobachtungsinstrumente. In E. Klieme, C. Pauli, & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis"* (Teil 3: Hugener, Isabelle; Pauli, Christine & Reusser, Kurt: Videoanalysen, S. 45-54). Frankfurt am Main: GPPF.
- Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251. doi:10.1007/s10984-006-9015-7
- Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: a multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers. Results from the COACTIV project* (pp. 97-124). New York, NY: Springer.

- Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. doi:10.1037/a0032583
- Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, (14), 25–47.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537. doi:10.1016/j.learninstruc.2008.11.001
- Oser, F., Dick, A., & Patry, J.-L. (1992) (Eds.). *Effective and responsible teaching: the new synthesis*. San Francisco, CA: Jossey Bass.
- Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann.
- Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*, 387–400.
- Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12.
- Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht*. Münster: Waxmann.
- Scheerens, J. (2004). *Review of school and instructional effectiveness. Background paper prepared for the Education for All Global Monitoring Report 2005*. Paris: UNESCO.
- 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.
- Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM-The International Journal on Mathematics Education*, *45*(4), 607-621.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.
- Steinweg, A. S. (2011). Einschätzung der Qualität von Lehr-Lernsituationen im mathematischen Anfangsunterricht - ein Vorschlag, *Journal für Mathematik-Didaktik*, *32*(1), 1-26.