Evaluation of a rating system for the assessment of metacognitive-discursive instructional quality
Edyta Nowińska, Anna-Katharina Praetorius

To cite this version:
Edyta Nowińska, Anna-Katharina Praetorius. Evaluation of a rating system for the assessment of metacognitive-discursive instructional quality. CERME 10, Feb 2017, Dublin, Ireland. hal-01949097

HAL Id: hal-01949097
https://hal.archives-ouvertes.fr/hal-01949097
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.
Evaluation of a rating system for the assessment of metacognitive-discursive instructional quality

Edyta Nowińska¹ and Anna-Katharina Praetorius²

¹Osnabrück University, Osnabrück, Germany; enowinska@uni-osnabrueck.de
²German Institute for International Educational Research, Germany; praetorius@dipf.de

Metacognition plays an essential role in learning mathematics. However, due to the lack of observational systems for evaluation of metacognition in mathematics instruction, rarely anything is known about how metacognition is practised and fostered when teaching and learning mathematics in class. This paper presents an observational system (a rating system) developed to reliably assess metacognitive activities in mathematics instruction. It also explains the methodology used to evaluate the reliability of ratings achieved with this tool and to investigate the stability of metacognitive-discursive practices between lessons of an individual teacher/class. Despite the high inference of conclusions needed to assess metacognitive-discursive instructional quality in seven dimensions, highly reliable ratings have been achieved for six dimensions. The paper discusses reasons for and consequences of the high reliability.

Keywords: Metacognition, discourse, rating system, generalizability study, decision study.

The role of metacognition in teaching and learning mathematics

Metacognition has been ascribed an essential role in regulating students’ cognitive processes in problem solving as well as in learning mathematics in general, in particular when constructing, organising, systematising, and connecting (pieces of) knowledge (cf. Schraw & Moshman, 1995; Wilson & Clark, 2004). However, hardly anything is known about how metacognition is practised and fostered in mathematics instruction. Assuming that enhancing learners’ metacognition is essential for promoting learning, research on metacognition in this area definitely merits future research (cf. Mevarech & Kramarski, 2014; Depaepe et al., 2010). For this kind of research a tool is needed that allows to reliably assess metacognitive practices when teaching and learning mathematics in a class. This paper reports on a research project that aimed at developing such a tool, named the rating system for analysing and assessing the metacognitive-discursive instructional quality (Nowińska, 2016). This tool can be used to first describe metacognition during class discussion, and to second evaluate how metacognitive activities are used to foster understanding in mathematics, in particular by elaborating students’ ways of thinking and reasoning, and by discussing them in a coherent and comprehensible way. One important research question underlying our work on developing and evaluating this tool was how stable metacognitive-discursive instructional quality is across lessons of an individual teacher in one class. In addition to advancing our knowledge about the occurrence of metacognitive-discursive instructional quality, investigating its stability allows identifying the number of lessons per teacher/class which would be needed to reliably measure metacognitive-discursive instructional quality.

Metacognition is often understood as knowledge about cognition and regulation of cognition (Flavell, 1976; Schraw & Moshmann, 1998). The groundwork for investigating metacognition in the domain of teaching and learning mathematics in a class has been done by Cohors-Fresenborg and...
Kaune (2007) as they developed a category system for an interpretative, transcript-based analysis of metacognitive and discursive activities (CMDA). This category system decomposes metacognition in planning, monitoring, and reflection. Examples are planning the structure of a proof or definition; monitoring the correctness of an argumentation; and reflecting on misconceptions or on difficulties experienced in interpreting a definition or in solving an equation. According to studies suggesting that the effects of metacognition on students’ understanding when learning in class seem to depend on the quality of the class discussion (cf. Mevarech & Kramarski, 2014; Depaepe et al., 2010), it was necessary to combine the analysis of metacognition with a deep analysis of precision, coherence, and accuracy of teacher’s and students’ contributions. For this purpose, CMDA also includes the categories discursivity and negative discursivity. Discursivity means activities enhancing precision, accuracy, and coherence in a class discussion, e.g. by making connections between different answers, or between external concept representations and students’ conceptions. Negative discursivity means activities with a negative influence on precision, accuracy, and coherence. The results of many years research conducted on metacognitive and discursive activities led the authors of CMDA to the conviction that discursive ways of practicing metacognition are crucial for supporting students’ understanding when learning mathematics in class. The term “discursive” does not simply mean “in a discourse” but is meant as a characteristic of discussions elaborating, explaining, and linking students’ ways of thinking in a coherent and comprehensible way.

The category system CMDA allows a detailed interpretation and categorization of local, single metacognitive and discursive activities, but it does not provide any additional tool for the global, comprehensive assessment of their instruction-related quality, thus of the extent to which they facilitate understanding of the mathematical subject discussed in class. The new rating system discussed in this paper is a result of extending the category system CMDA to a video-based observational system aimed to analyse and measure ‘metacognitive-discursive instructional quality’ in a comprehensive way. For this aim, several dimensions of the metacognitive-discursive instructional quality as well as evaluation criteria to rate them have been developed (for details see Nowińska, 2016). To allow its application, the rating system needed to be valid and reliable despite the complexity and high inferences required for rating metacognitive-discursive instructional quality.

In the following, we first explain the design of the rating system. Second, we describe the methodology used to evaluate its reliability (G study), present the achieved results, and discuss their consequences for generalizable evaluation of metacognitive-discursive instructional quality. Finally, we discuss consequences of our study for further research aimed to deepen our understanding of metacognitive-discursive instructional quality and to improve teaching and learning practices in class.

---

1 The complete German version of CMDA is presented in Cohors-Fresenborg, Kaune, & Zülsdorf-Kersting (2014).

2 www.mathematik.uni-osnabrueck.de/fileadmin/didaktik/Projekte_KM/Kategoriensystem_EN.pdf
The design of the rating system

Our conceptualization of metacognitive-discursive instructional quality and its decomposition in seven dimensions is based on research literature concerning relations between metacognition and learning gains (e.g. Mevarech & Kramarski, 2014; Depaepe et al., 2010), and on the preliminary research work related to the category system CMDA (e.g., Cohors-Fresenborg et al., 2010; Gretzmann, 2011). Furthermore, we analysed more than 20 videotaped lessons to deepen our understanding of these dimensions. Each dimension is described by means of a guiding question (GQ) focusing raters’ attention on aspects to be analysed and evaluated, as well as of several answering categories. For each GQ, the answering categories describe particular aspects of classroom discussions that differ in quality. The categories are ordered so that they reflect increasing quality of the classroom discussion with regard to the relevant aspects, and constitute a rating scale. In the following, the seven guiding questions are described briefly (for the detailed version see Nowińska, 2016).

GQ 1 puts the focus on using metacognitive activities for an elaborate discussion of mathematical content and on supporting learners’ autonomy in practicing such activities. To answer this GQ, the rater has to distinguish, among others, between metacognitive activities limited to monitoring results of calculations, on the one hand, and extended to reflection on mathematical ways of reasoning, methods, definitions, and conceptions related to them, on the other hand. Due to the essential role of argumentation in learning and understanding mathematics, GQ 2 focuses on justifications combined with metacognitive activities, and on supporting learners’ autonomy in providing and analysing justifications. To answer this GQ, the rater has to distinguish between fragmentary justifications, on the one hand, and efforts made in class to orchestrate single justifications in order to produce precise comprehensive argumentations, on the other hand. GQ 3 aims at assessing to which extent the interplay of metacognitive and discursive activities foster students’ understanding of subject-specific issues discussed in the particular lesson. The answering categories for this GQ distinguish among others between situations without any productive use of metacognitive and discursive activities, and situations in which (at least in the case of one single learner) metacognitive and discursive activities foster and express learners’ understanding of the subject-specific issues discussed in the class. GQ 4 analyses the use of discursive activities in producing precise and coherent discussion. Such discussion is an essential precondition for an effective use of metacognition in class in order to foster learners’ understanding. GQ 5, on the contrary, evaluates to what extent negative discursivity (e.g., not taking notice of inadequate mathematical vocabulary or of fragmentary answers as well as of answers not related to the discussed question) leads to ignoring students’ cognitive and metacognitive processes, and hinders the reciprocal understanding in class as well as the understanding of subject-specific issues. GQ 6 evaluates to what extent metacognitive and discursive activities are used to build coherent and stringently guided discourse units (i.e., debates). The answering categories for this GQ distinguish between classroom situations without any debates, and situations with at least one remarkable debate led by the teacher or by students. GQ 7 aims at assessing to which extent metacognitive and discursive activities are related to challenging and complex subject-specific issues (e.g., related to meta-mathematics), used to elaborate such issues, and to foster learners’ understanding of them.
To ensure reliable ratings despite the high level of inference needed to answer the guiding questions, the rating procedure was designed as a two-step procedure. In the first step of the rating process, the rater categorises each of the teacher and student contributions. Hereby, the rater uses the category system adopted from Cohors-Fresenborg and Kaune, and works with special software which at the end of the categorisation generates a graphic representation (i.e. category line; for more details see Nowińska, 2016). The category line includes all codes for metacognitive and (negative) discursive activities set by the rater, and distinguishes between codes for teacher and student activities. It serves as a basis for interpreting relations between teacher’s and students’ metacognitive and discursive behaviour, and for assessing students’ autonomy in practicing these activities. Thus, the purpose of the first step is to get insight into the kind and quality of each single metacognitive and (negative) discursive activity, and to prevent the rater from rushed and inadequate ratings. In the second step, the rater uses the category line and the video transcript to evaluate the lessons by means of the seven rating scales elaborated on above.

In order to be able to carry out these tasks, three raters (students at the end of their master study course in mathematics education) participated in an intensive rater training (6 months, 180 h in sum). They were qualified to understand the purpose of the rating system, the foci of the seven rating scales, and the use the rating system. During the rater training (and also after it) the raters were obligated to justify their decisions regarding their interpretation of each single metacognitive and (negative) discursive activity as well as their final evaluation of the instructional quality. This allowed the trainer to discuss the answers given by the raters in detail, to discuss reasons for differences between the raters, and to provide each rater with detailed feedback. The videos and transcripts used during the training were separate from the ones used in the current study.

Methodology

In the current study, sequences from 24 videotaped mathematics lessons (6 teachers/classes with 4 lessons per teacher/class) were analysed. For each teacher, four lessons were videotaped within two weeks, and should represent “normal” lessons in these classes. From each lesson, a 10-minute video sequence showing a discussion in the class was chosen. This was done by two independent experts who first analysed each lesson, and suggested the sequence in which the main topic of the lesson was discussed, and in which the students actively participated in the discussion. Finally, the experts agreed on one sequence. In many cases, however, only one 10-minute discussion could be indicated, whereas in the remaining time the students worked individually or in pairs. Each video sequence was evaluated by three independent raters, who had taken part in the rater training.

Generalizability theory was used (Shavelson & Webb, 1991; for an application to the instructional context, see Praetorius et al., 2012) for assessing the generalizability (which can be interpreted similarly to reliability in classical test theory) of the rating instrument. The reported relative G coefficient can be interpreted analogously to a reliability coefficient in the classical test theory. Thus, a coefficient ≥ 0.7 is needed for a satisfactory reliability. In addition to providing these G coefficients, generalizability studies (G studies) allow decomposing the variance in rating scores into different components (e.g., teachers, lessons, and raters), their interactions, and measurement error. Therefore, G study results provide more detailed and precise information regarding reliability than reliability coefficients used in classical test theory. Furthermore, decision studies (D studies)
can be conducted to estimate the reliability under multiple hypothetical measurement conditions, thus also allowing to analyse numbers of lessons per teacher/class higher than the number actually evaluated by the raters in our study. In the present study, it was investigated how many lessons per teacher/class would be necessary for a reliable assessment of the aspects of the metacognitive-discursive quality determined by the seven dimensions.

**Results**

The results of the G studies indicated satisfactory reliability of ratings concerning six out of the seven dimensions of the metacognitive-discursive instructional quality (guiding questions 1 to 6), for which the relative G coefficient varied between 0.78 and 0.98 (see Table 1). The ratings concerning dimension 7 were not reliable, with a relative G coefficient of 0.38.

<table>
<thead>
<tr>
<th>Lesson-unspecific (stable) component</th>
<th>GQ 1</th>
<th>GQ 2</th>
<th>GQ 3</th>
<th>GQ 4</th>
<th>GQ 5</th>
<th>GQ 6</th>
<th>GQ 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>89</td>
<td>45</td>
<td>60</td>
<td>52</td>
<td>71</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>Lesson-specific component</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l:t</td>
<td>0</td>
<td>46</td>
<td>21</td>
<td>39</td>
<td>22</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>Rater bias components</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>r*t</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>51</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r*(l:t); e</td>
<td>8</td>
<td>9</td>
<td>19</td>
<td>8</td>
<td>7</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>Relative G-coefficient</td>
<td>0.98</td>
<td>0.78</td>
<td>0.90</td>
<td>0.83</td>
<td>0.92</td>
<td>0.83</td>
<td>0.38</td>
</tr>
</tbody>
</table>

**Table 1: Relative G-coefficients and variance decomposition (in %) for the seven dimensions**

Based on the rating data, the variance in the ratings was decomposed in variance components attributable to the teacher/class (t), lessons nested in teachers (l:t), raters (r), the interaction between teachers and raters (r×t), and the unexplained variance, i.e. residual (r*(l:t); e). Table 1 shows the percentage of variance explained by the different variance components.

For dimensions 1 to 6, the amount of variance attributable to rater bias was very small (between 1% and 3% of the entire variance); this indicates that the raters do rarely differ in their ratings. However, further rater training would be needed to eliminate the very high amount of the variance (55%) attributable to rater bias for dimension 7 in order to get reliable ratings.

The ratio of t to l:t, which describes the stability of the given dimension across lessons of an individual teacher/class, indicates partly very high stability (see e.g., GQ 1) and partly very low stability (see e.g., GQ 7).

To determine how many lessons per teacher/class are necessary to measure metacognitive-discursive instructional quality in a stable and reliable way, D analyses were conducted with the hypothetical number of lessons per teacher/class varying between 1 and 10. The number of raters was fixed to the actual number in the study (i.e., three). Figure 1 illustrates the results of the D study for each of the seven dimensions.
To obtain relative G coefficients greater than 0.7, one lesson is needed for the dimensions related to GQ 1 as well as GQ 5. Two lessons are needed for the dimensions related to GQ 3, GQ 4, and GQ 6, and three lessons for the dimension related to GQ 2. Thus, three lessons are sufficient to achieve a G coefficient of 0.7 for GQ 1 to GQ 6.), whereas 5 lessons per teacher/class would be needed for the reliability greater than 0.8. Due to the high amount of the variance attributable to rater bias for GQ 7, no satisfactory reliability concerning this dimension could be reached, even with 10 lessons per teacher/class without further rater training.

Discussion

The aim of our research project was to develop a reliable rating system for assessing metacognitive-discursive instructional quality. For this purpose, seven dimensions of metacognitive-discursive quality had been developed. Despite the high amount of inferences needed to rate these dimensions, highly reliable ratings were achieved for six of them. This rather unusual result (for an overview on the amount of rater effects found in prior studies, see Praetorius et al., 2012) can likely be explained, among others, with the intensive rater training, and with the two-steps procedure of the rating process. Both aspects prevented the raters from a superficial analysis, and instead forced well-reasoned scoring.

No satisfactory reliability could be achieved for the seventh dimension. The reliability analyses showed that this is due to high rater bias. Obviously, the meaning of “complex subject-specific” issues, which is at the core of the seventh dimension, has not been interpreted in the same way by all raters. Thus, additional rater training or other raters with a more substantial background in mathematics education would be needed to get reliable ratings for this dimension. Such efforts seem highly desirable as the seventh dimension plays an important role in a long-term evaluation of the metacognitive-discursive quality in a class. In general, complex subject-specific issues are discussed in mathematical instruction rarely. Discussing such issues indicates teacher’s efforts to deepen and systematise students’ cognition related to (meta-)mathematical questions, methods or ways of reasoning, and therefore it is a significant characteristic of instructional quality.
The seven dimensions of the metacognitive-discursive instructional quality vary in their stability between lessons in a particular class. The lowest variability could be determined for the dimension concerning the extent to which metacognition is practised in a class, in interactions between the teacher and the students (GQ 1), and the highest for this concerning metacognitive activities combined with justifications (GQ 2). The quite stable first dimension is based on GQ 1 which also takes some observable aspects of patterns in interactions between metacognitive and discursive teacher and students activities, whereas GQ 2 focuses more on metacognitive activities in relation to the content discussed in class and to students’ reasoning concerning this content. The results indicate that the quite stable observational patterns in practicing metacognition do not necessarily imply the stability of metacognitive efforts to elaborate the mathematical issues discussed in class and to foster students’ understanding. A deep analysis of videos is needed to explain this observation. Our preliminary analysis shows that in some classes, providing justification seems to be well established as a social norm. This means that the learners and the teacher are used to justify their answers, i.e. to practice monitoring or reflection. However, by doing so, not always the necessary attempts are made in the class to reflect on these justifications, to control and correct them, and to orchestrate single and fragmentary justifications in order to produce a coherent global explanation related to mathematical issues discussed in the class. This can be observed in particular when new concepts, definitions or strategies are introduced. Despite the high number of single justifications combined with metacognition, the lack of a well-orchestrated mathematical justification related to the new issues may hinder understanding. Consequently, this leads to a low score for the second dimension of the metacognitive-discursive instructional quality (GQ 2). The score can be higher when the tasks discussed in the class do not require a global well-orchestrated mathematical justification, and the lack of it cannot be evaluated negatively, with very low scoring. Thus, the variability of the second dimension seems to be related to the complexity of the mathematical content. This observation seems plausible but it must be investigated more deeply. Considering only the number of justifications combined with metacognitive activities would most likely enhance the reliability of the ratings in the second dimension but it would distort the validity of the instructional quality.

Our D studies are of crucial importance for further research on metacognition in mathematics instruction. Due to the variability of metacognitive-discursive practices between lessons in a particular class, at least three lessons per teacher/class and three qualified raters would be needed for reliable (generalizable) evaluation of the metacognitive-discursive quality with regard to six dimensions of this construct. Thus, given these relative small numbers, the rating system can be considered as a practicable research tool although intensive rater training is needed.

Given this result, a pivotal next step for research on metacognitive-discursive instructional quality is to investigate the effects of each of the six dimensions on students’ mathematics achievement. In doing so, the empirical relevance of metacognitive-discursive instructional quality can be investigated, and implications for supporting metacognition to foster mathematical understanding can be suggested. Thus, continuing this research is highly desirable. It would shift the focus from measurement and evaluation to development and improvement. This would require the work with teachers, and not only research on teachers’ instructional practices. Thereby, the rating system presented in this paper can be used as an analytical tool in teacher trainings for guiding teachers’ reflection on their own practices and on learners’ metacognitive and discursive behaviour.
References


