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. Due to missing tools for observation and evaluation of metacognitive activities in mathematics instruction, rarely anything is known about how metacognition is fostered in mathematics instruction. This paper presents the design of such an innovative tool. It also explains the methodology used to evaluate its reliability as well as to investigate the stability of metacognitive-discursive instructional features in seven dimensions. Due to the high inference of conclusions needed to assess instructional quality, highly reliable ratings were achieved for six of seven dimensions. The paper discusses the quite unusual result and its consequences for evaluating the metacognitive-discursive instructional quality.

Keywords: Metacognition, Discourse, Rating system, Generalizability study, Decision study.

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 of mathematics, in particular when constructing, organising, systematising, and connecting (pieces of) knowledge (cf. Schraw & Moshman, 1995; Wilson & Clark, 2004). Though, rarely anything is known about how metacognition is 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). One central first step for this kind of research is to develop a tool 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 instructional metacognitive-discursive quality (RSMDQ) (Nowińska, in press). This tool not only describes the amount of metacognition during class discourse. It also evaluates how metacognitive activities are used to foster understanding in mathematics, and to build a coherent, elaborated discourse involving students’ ways of thinking.

Metacognition is often understood as knowledge about cognition and regulation of cognition (Flavell, 1976; Schraw & Moshmann, 1998). The groundwork for interpreting metacognition in the domain of teaching and learning mathematics in a class had 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)¹. 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

¹ The complete German version of CMDA is presented in Cohors-Fresenborg, Kaune, & Zülsdorf-Kersting (2014).
experienced in interpreting a definition or in solving an equation. For the purpose of a detailed description of effects that metacognitive activities may have on the coherence of mathematical discourse, the category system also includes the categories called *discursivity* and *negative discursivity*. *Discursivity* means activities carried out to enhance the precision, accuracy and coherence in a discourse, e.g. by making connections between questions and answers, or between external concepts’ representations and students’ conceptions and their ways of thinking as they emerge in a discourse. On the contrary, *negative discursivity* means activities with a negative influence on the precision, accuracy and coherence of the discourse, e.g. the use of inadequate vocabulary or superficially clear sentences with an unclear sense or without connections to what was asked or said before.

The category system CMDA allows a detailed interpretation of local, single metacognitive and discursive activities but it does not provide any tool for the global assessment of their instructional quality in the discourse, it means of the extent to which the interplay of these activities facilitates learning and understanding of the mathematical subject discussed in class. For the new rating system the category system CMDA had been adopted as a tool for the *video-based* interpretation of metacognitive and (negative) discursive activities. The decomposition of metacognitive-discursive instructional quality in several dimensions, and the construction of evaluation criteria (rating scales) which differentiate between the extent of metacognition on the one hand, and its role in supporting understanding of mathematics on the other hand, were the milestone in developing the rating system (for details see Nowińska, in print). Another challenge was to make the rating system practicable, i.e. to assure valid and reliable ratings despite the complexity and high inference of the conclusions needed for such ratings.

In the following, we first explain the design of the rating system. Second, we describe the methodology used to evaluate its reliability (G study) as well as the number of lessons per teacher needed to get reliable and valid score (D study). Third, the results are discussed with respect to the stability of metacognitive instructional features and its consequences for evaluating such practices.

**The design of the rating system RSMDQ**

Our decomposition of metacognitive-discursive instructional quality in seven dimensions is on the one hand based on research literature concerning relations between metacognition and learning gains (e.g. Mevarech & Kramarski, 2014; Depaepe et al., 2010), and on the other hand on the preliminary research work related to the category system developed by Cohors-Fresenborg and Kaune (e.g., Cohors-Fresenborg et al., 2010; Gretzmann, 2011). Furthermore, we analyzed more than 20 videotaped lessons to deepen our understanding of these dimensions and to develop a detailed description for each of them using the observed video actions. Each dimension consists of a guiding question (GQ) focusing the rater’s attention on the aspects to be analysed and evaluated, as well as of several answering categories. For each GQ, the answering categories describe reactions to discourse situations that differ in quality, and constitute a rating scale. The categories are ordered so that they reflect increasing quality of discourse with regard to the relevant aspects.

---

2 [www.mathematik.uni-osnabrueck.de/fileadmin/didaktik/Projekte_KM/Kategoriensystem_EN.pdf](www.mathematik.uni-osnabrueck.de/fileadmin/didaktik/Projekte_KM/Kategoriensystem_EN.pdf)
In the following, the seven dimensions of the metacognitive-discursive instructional quality, structured by the guiding questions and the answering categories, are presented briefly. This is necessary to understand the results concerning the reliability of the rating system, and the generalizability of the evaluation results. Due to the space limitation, for the guiding questions 4 and 5, which have five answering categories (constituting their rating scales) each, only three of these categories can be presented here. For a more detailed version of the instrument see Nowińska (in press).

**GQ 1:** How prevalent are metacognitive activities during class in interactions between teacher and students?

1: Very few metacognitive activities are practised in class and no effort is made to use these activities accurately to better explain and understand the subject of the discussion or to foster metacognitive skills of the learners. 2: Metacognition is carried out by the teacher and by the students and effort is made to explain and understand the subject of the discussion; teachers’ metacognitive activities dominate the discourse and its result. 3: The teacher decreases his or her own activities in favour of student activities and motivates the learners to practice metacognition. 4: The learners are autonomous in practicing metacognition, and they make effort to precisely explain and clarify the mathematical subject of the discussion.

**GQ 2:** How are metacognitive activities combined with justifications?

1: Justifications combined with metacognition are almost absent and it cannot be assumed that they are valued as being important in the culture established in the class, in particular as a means to build a coherent, precise, and elaborated mathematical discourse fostering students’ understanding of the mathematics. 2: Justifications combined with metacognition are practised in class but no effort is made to use them to build a coherent, precise, and elaborated mathematical discourse fostering students’ understanding of the mathematics. 3: Justifications combined with metacognition are practised and valued as being important in the class but the learners are not autonomous in explaining and justifying. 4: The learners practice metacognition and they are autonomous in explaining and justifying.

**GQ 3:** To what extent does the interplay of metacognitive and discursive activities foster students’ understanding of subject-specific issues discussed in class?

1: No productive use of metacognitive and discursive activities can be observed. 2: Only in the case of one single learner it can be observed that his or her metacognitive and discursive activities foster his or her understanding of the subject-specific issues discussed in the class. 3: Metacognitive and discursive activities foster students’ understanding of subject-specific issues discussed in class.

**GQ 4:** To what extent does discursivity contribute to build a coherent discourse, in particular to make teacher’s and learners’ explanations, argumentations, conceptions and ways of thinking comprehensible and understandable for others?

1: No effort is made to practice discursivity. 2 […] 3: The teacher and the learners practice discursivity but no effort is made to build a coherent discourse. 4 […] 5: The learners are autonomous in practicing discursivity and make effort to build a coherent discourse.
**GQ 5:** To what extent does negative discursivity hinder the reciprocal understanding in class and the understanding of subject-specific issues (tasks, tools, methods, or ways of reasoning)?

1: The negative discursivity hinders the reciprocal understanding in class and the understanding of subject-specific issues. 2 […] 3: The negative discursivity affects the local clarity of the discourse but it does not hinder the understanding of subject-specific issues. 4 […] 5: The negative discursivity does not have any negative consequences for the discourse.

**GQ 6:** To what extent are metacognitive and discursive activities used to build coherent and stringently guided discourse units called ‘debates’?

1: No debates are observed in the discourse. 2: Only very short, not elaborated debates can be observed and they are guided by the learners. 3: Debates are guided by the teacher and the learners make no effort to build elaborate debates in their interactions. 4: The learners build elaborated debates.

**GQ 7:** To what extent are metacognitive and discursive activities used to foster learners’ understanding in the case that challenging and complex issues are discussed in class?

1: No challenging and complex issues are discussed in class. 2: A challenging, complex issue is posed in class but no metacognitive and discursive efforts are made to elaborate and clarify it. 3: The teacher makes metacognitive and discursive effort to clarify a challenging and complex issue but the learners’ negative discursivity hinders the understanding of it. 4: The teacher and the learners make metacognitive and discursive effort to elaborate and clarify a challenging and complex issue.

To assure 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 teacher’s and student’s contribution. 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 called category line (for more details see Nowińska, in print). The category line includes all codes set by the rater, and differentiates between codes for teacher and student activities. It serves as 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 this step is to get insight in the kind and quality of each single metacognitive and (negative) discursive activity, and to prevent the rater from rash and inadequate ratings. In the second step, the rater uses the category line and the video-transcript to evaluate the lessons by means of seven rating scales.

Furthermore, three raters (students at the end of their master study course in mathematics education) participated in an intensive rater training which was almost 6 months long (with meetings at least once a week for at least 3 hours). They were trained how to use the rating system, and obligated to justify their decisions regarding the coding and the final evaluation; the videos and transcripts used in this training were not used in the subsequent study aimed to assess the reliability of the ratings. The obligation allowed the trainer to discuss in detail the answers given by the raters, to discuss the reasons for differences between them, and to provide each rater with detailed feedback.
Methodology

To assess the reliability of ratings, 24 mathematics lessons (6 teachers/classes with 4 lessons per teacher/class) were videotaped. 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 discourse 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-minuts long discourse 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.

Given the research purpose of our study, generalizability theory was used (Shavelson & Webb, 1991; for an application to the instructional context, see Praetorius et al., 2012). The reliability of the ratings was measured with the relative G-coefficient. It indicates to what extent the ratings of the metacognitive-discursive instructional quality provided by the three independent raters indicate a sufficient generalizability (which can be interpreted similarly to reliability in classical test theory). For this coefficient the same quality criteria apply as for the reliability coefficient in classical test theory: G-coefficient ≥ 0.7 is needed for a satisfactory reliability. Generalizability studies (G studies) allow a decomposition of 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. In doing so, for example the variability of each dimension of the metacognitive-discursive quality between lessons of a specific teacher/class can be investigated. 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 be the seven dimensions.

Results

The results of the G-studies indicated satisfactory reliability of ratings concerning six out of the seven dimensions of the instructional metacognitive-discursive quality – dimensions 1 to 6 (it means dimensions given by the guiding questions 1 to 6), for which the relative G-coefficient varied between 0,78 and 0,98 (Table 1). The ratings concerning dimension 7 were not reliable, with a relative G-coefficient of 0,38.

Based on the rating data, the variance in the ratings was decomposed in variance components attributable to teacher/class (t), lessons nested in teachers (l:t), raters (r), the interaction between teachers and raters (r×t), and the combination of the interaction of raters and lessons (nested in teachers), confounded with measurement error (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 is 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 in order to get satisfactory reliable ratings for dimension 7.

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

<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>

The ratio of t to l:t, which describes the stability of the seven dimensions between lessons of an individual teacher/class, indicates partly very high stability (for the dimension related to GQ 1) and partly very low stability (for the seventh dimension related to GQ 7).

To determine how many lessons per teacher/class are necessary to measure metacognitive-discursive instructional quality, 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 number of three, as it was actually given in the study. Figure 1 illustrates the results of the D study for each of the seven dimensions.

Figure 1: Relative G coefficients for D studies with 1-10 lessons per teacher/class for GQ 1–7

To obtain relative G coefficients greater than 0.7, one lesson would be needed for the dimensions related to GQ 1 as well as GQ 5; 2 lessons would be needed for the dimensions related to GQ 3,
GQ 4, and GQ 6; and 3 lessons for the dimension related to GQ 2. Thus, this level of reliability could be achieved in each of the six dimensions (GQ 1 to GQ 6) with 3 lessons per teacher/class, whereas 5 lessons per teacher/class would be needed for the reliability greater than 0.8 in all these dimensions. Due to the high amount of the variance attributable to rater bias in case of the dimension given by the GQ 7, no satisfactory reliability of rating concerning this dimension of the metacognitive-discursive quality can 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 (RSMDQ). For this purpose, seven dimensions of metacognitive-discursive quality had been developed. Despite the high inference of conclusions needed to assess instructional quality in 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 be explained, among others, with the intensive rater training, and with the two-steps procedure of the rating process. Both aspects hindered the rater from a superficial discourse analysis, and forced well reasoned scoring.

No satisfactory reliability could be achieved for the seventh dimension. The ratings provided by the three raters show that the reason for the not satisfactory reliability is high rater bias concerning the meaning of “complex subject-specific” issues being in the centre of the seventh dimension. The variance decomposition indicates that an additional rater training would be needed to get reliable ratings for this dimension. However, the question arises which content and methods would be needed to achieve a substantial improvement of the raters’ competencies required for reliable rating of this dimension. Further investigations are necessary to approach these research questions.

The seventh dimension without a satisfactory reliability plays an important role rather in a long-term evaluation of the metacognitive-discursive quality in a class. In general, discussing a complex subject-specific issue causes more negative discursivity and more teacher-guided debates than in case of an easier problem. The seventh dimension may explain to what extent the low scores for the metacognitive-discursive quality achieved in other dimensions are related to the complexity of issues discussed in the class.

For the purpose of a reliable evaluation of the metacognitive-discursive quality the rating system can be used without the seventh dimension. Based on results from our G-studies, such a reduced form of the rating system can be considered as a reliable tool. This statement regards the ratings of one single lesson given by three independent raters, after an intensive rater training. Removing the seventh dimension from the rating system would lead to the loss of an explanatory function of ratings achieved with the other six dimensions, but it would not distort the scoring given to dimensions 1 to 6. It is worthily to investigate whether this loss could be set off by ratings provided for the seventh dimension by experts with deep understanding of mathematics.

Several things of crucial importance for research and improvement concerning metacognition in mathematical discourse can be learned from our D studies. First, the seven dimensions of the
metacognitive-discursive instructional quality vary in their stability between lessons in a particular class. The highest variability could be determined for the dimension concerning metacognitive activities combined with justifications, and the lowest for this concerning the extent to which metacognition is practised in a class, in interactions between the teacher and the students. Further research is needed to investigate the relations between the dimensions, between them and the mathematical achievements, and to deepen our understanding of mechanism influencing the effectiveness of metacognition in promoting understanding in mathematics. Second, three lessons per teacher/class and three raters would be needed for reliable evaluation of the metacognitive-discursive quality with regard to six dimensions of this construct in a particular class in order to approach the mentioned research questions. Thus, the results of our G- and D-analyses are also of practical relevance for further research on metacognition in teaching and learning mathematics.

References


