Teaching and Teacher Education 76 (2018) 214–226

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Research paper

Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding

Sophie Oudman [a, *], Janneke van de Pol [a], Arthur Bakker [b], Mirjam Moerbeek [c], Tamara van Gog [a]

[a] Department of Education, Utrecht University, The Netherlands
[b] Freudenthal Institute, Utrecht University, The Netherlands
[c] Department of Methodology and Statistics, Utrecht University, The Netherlands

HIGHLIGHTS

- Teachers judged (predicted) students' performance on decimal magnitude test problems.
- They received student cues (names), answer cues (to prior practice problems), or both.
- Access to only answer cues was helpful in judging what students did not understand.
- Availability of both cues did not improve accuracy compared to only student cues.

ARTICLE INFO

Article history:
Received 7 February 2017
Received in revised form
13 February 2018
Accepted 14 February 2018
Available online 26 February 2018

Keywords:
Teacher judgment
Judgment accuracy
Cue utilization
Primary education
Mathematics education
Decimals

ABSTRACT

To gain insight into how teachers' judgment accuracy can be improved, we investigated effects of cue-type availability. While thinking aloud, 21 teachers judged their fourth grade students' ($n = 176$) decimal magnitude understanding. Sensitivity (correctly judging what students did understand) did not improve from availability of both answer cues (students' answers to prior practice problems) and student cues (knowledge of students triggered by knowing their names), and was lower when only answer cues were available, compared to only student cues. Specificity (correctly judging what students did not understand) was higher when only answer cues were available, compared to only student cues or both student and answer cues.

1. Introduction

To stimulate students' learning optimally, teachers need to provide adaptive instruction; that is, they have to tailor their explanations and instruction to a student's current level of understanding (e.g., Van de Pol, Volman, & Beishuizen, 2010). For teachers to be able to make adaptive instructional decisions, their judgments of their students' understanding need to be accurate (Klug, Bruder, Kelava, Spiel, & Schmitz, 2013; Südkamp, Kaiser, & Möller, 2012; Van de Pol, Volman, & Beishuizen, 2011). Prior studies have shown, however, that there is much room for improving teachers' judgment accuracy (see for a meta-analysis Südkamp et al., 2012). This especially applies to teachers' judgment accuracy of students' conceptual mathematical understanding (Thiede et al., 2015). Yet, research that gives insight into how teachers' judgment accuracy can be improved is scarce.

Therefore, the first aim of the present study was to investigate how teachers' judgment accuracy of students' conceptual mathematical understanding can be enhanced, by manipulating the availability of information that can be used while making a judgment. According to the cue-utilization approach, judgments are based on specific pieces of information (i.e., cues) that can be more or less predictive (i.e., diagnostic) of students' actual understanding (Brunswik, 1956; Koriat, 1997; Thiede, Griffin, Wiley, & Anderson, 2010; Van Loon, De Bruin, Van Gog, Van Merriënboer, &

* Corresponding author. Department of Education, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands.
E-mail address: v.s.oudman@uu.nl (S. Oudman).

https://doi.org/10.1016/j.tate.2018.02.007
0742-051X/

Dunlosky, 2014). The more predictive the cues being used, the more accurate a teacher's judgments of students' understanding will be. Manipulating which information is available will provide insight into which cues do and do not improve judgment accuracy. The second aim of the present study is to explore what cues teachers base their judgments on under the different cue-availability conditions, to gain more insight into their judgment process. This may ultimately aid the development of support tools to improve teachers' judgment accuracy.

1.1. Teachers' judgments of students' conceptual mathematical understanding

In their meta-analysis Südkamp et al. (2012) conclude that teachers' judgment accuracy, reflected by the correlation between teachers' judgments of students' performance in language and mathematics and students' actual test performance, was positive and fairly high (Fisher's z transformed correlation = .63), but that there is still much room for improvement. As in most studies on teachers' judgment accuracy, in the studies included in the meta-analysis teachers' judgments were measured by asking teachers for one global rating per student (e.g., ratings of students' reading performance or a prediction of the number of correct answers on a test) or student rankings (e.g., ranking of the students in their class from lowest to highest mathematical understanding). The accuracy of these global judgments reflects teachers' knowledge on students' overall performance, not how well they are able to judge what individual students do and do not understand within a domain. Item-specific judgments do reflect this latter type of knowledge, which is what teachers need in order to make adaptive instructional decisions, such as differentiating tasks or providing adequate instruction and feedback to individual students (Artelt & Rausch, 2014; Gabriele, Joram, & Park, 2016).

The few studies that did include item-specific judgments of students' mathematical understanding (Artelt & Rausch, 2014; Gabriele et al., 2016; Karing, Pfost, & Artelt, 2011) found average "hit rates" (i.e., the proportion of accurately judged items when judging all items of a test) between the 58% and 78%. Taking into account that a random item prediction has on average 50% chance of being accurate, as teachers only indicated whether an item was answered correctly or incorrectly by the student, 58% is just above chance (i.e., when guessing). The item-specific judgments included in these studies did not distinguish between students' procedural skills and conceptual understanding in mathematics. Thiede et al. (2015) did make this distinction and their findings indicate that especially judging student conceptual mathematics understanding is challenging; they found that the average judgment accuracy for students' conceptual mathematics understanding, as measured by the gamma correlation (computed across the students within a class), was only .20 after intervention (vs. gamma correlation = .66 for computational skills; 1 would mean perfect prediction). In sum, prior studies showed there is a need to improve the accuracy of teachers' (item-specific) judgments of students' conceptual mathematics understanding, but knowledge on how to do so is lacking.

1.1.1. Teachers' cue utilization

Teachers make numerous instructional decisions during everyday teaching practice that are based on judgments of their students' current level of understanding. The accuracy of these judgments could be influenced by several factors, such as teacher characteristics (e.g., their professional expertise) and characteristics of the test that is used to make judgments (e.g., the subject area; Südkamp et al., 2012). In the current study we especially focus on the specific pieces of information that teachers base their judgments on, typically referred to as cues (Brunswik, 1956; Koriat, 1997). For instance, a teacher's observation that a particular student flawlessly completed yesterday's assignment on decimal magnitude (i.e., a cue) can lead to the judgment that this student's understanding of decimal magnitude is excellent. In turn, the teacher may make the instructional decision that the student can skip today's exercises on decimal magnitude and continue with exercises on adding decimals.

Studies on what cues teachers actually use when judging students' understanding are scarce and differ strongly in their methodology. Whitmer (1982) interviewed elementary teachers about the information they commonly used when giving grades for mathematics and language. Webb (2015) also interviewed elementary teachers, directly after they made prospective predictions of how their students scored on a mathematics test, and asked them what they based their predicted scores on. Cooksey, Freebody, and Wyatt-Smith (2007) asked elementary teachers to think aloud while making retrospective judgments of students written texts (i.e., the teachers were provided with students' products). Cues that were frequently reported by the teachers in these three studies were students' prior performances in a specific subject; students' general cognitive abilities and learning disorders; students' problem solving skills; students' motivation and interest; students' effort and discipline; what content had been taught or practiced previously; and the difficulty of a specific domain or task. Apparently, teachers derive the cues they use to inform their judgments from different information sources, such as the content material or characteristics of the task at hand (i.e., task content cues), information about students' prior performances (i.e., answer cues), and more general information about the students (i.e., student cues).

Several studies have zoomed in on teachers' use of student cues, rather than on other cue types. Correlational studies showed that students' ethnicity, SES, classroom engagement, disability status, and social competency were predictive of the height of teachers' judgments of their students' literacy and mathematical understanding, even when controlling for students' actual performances (e.g., Furnari, Whittaker, Kinzie, & DeCoster, 2017; Hurwitz, Elliott, & Braden, 2007; Kaiser, Retelsdorf, Südkamp, & Möller, 2013; Paleczek, Seifert, & Gasteiger–Klicpera, 2017; Ready & Wright, 2011). This implies that teachers use these student cues while making judgments of students' understanding. Comparable conclusions can be drawn from two experimental studies by Kaiser et al. (2013) and Kaiser, Südkamp, & Möller (2017), in which participants directed pre-designed questions at fictional students and observed their responses in a simulated classroom environment. Next, participants indicated how many of those pre-designed questions they thought each of the students had answered correctly, reflecting teachers' judgments of students' mathematics or reading achievement. Students' engagement (operationalized as the probability of a simulated student volunteering to answer a question; Kaiser et al., 2013) and students' minority status (Kaiser et al., 2017) were significantly related to teachers' judgments of the amount of correctly answered questions. In another study, Kaiser, Möller, Helm, and Kunter (2015) compared the effects of different types of information on pre-service teachers' judgment accuracy of fictional students' mathematics grades. All teachers were provided with information on students' oral and written achievement in mathematics and some teachers were additionally provided with student characteristics such as students' self-concept and intelligence. From the significant correlation between teachers' judgments of the mathematics grade and the value of the presented student characteristics (i.e., gender, intelligence, and German dictation exercise grade) it can again be concluded that teachers probably used these student cues while making judgments of students' mathematics performance.

1.1.2. Relation between teachers' cue utilization and judgment accuracy

As mentioned before, the more predictive the cues that the teacher uses in making judgments of a student's understanding, the higher the judgment accuracy will be (Brunswik, 1956; Koriat, 1997; Thiede et al., 2010; Van Loon et al., 2014). For example, the correctness of a student's conceptual expression during a class discussion might be more predictive for this students' actual understanding of the content material than the time the student spends on a task. Basing judgments on more predictive cues will lead to more accurate judgments, which in turn will lead to more adaptive instructional decisions (Klug et al., 2013; Südkamp et al., 2012; Van de Pol et al., 2011).

In the experimental study by Kaiser et al. (2015) teachers who were only provided with information on students' oral and written achievement in mathematics, made more accurate judgments of fictional students' mathematics grades than teachers who were additionally provided with student characteristics (i.e., students' engagement, minority status, gender, intelligence, and German dictation exercise grade). This finding suggests that student cues might be not predictive of students' actual understanding, as the availability (and therefore presumably, the use) of such cues resulted in less accurate judgments of students' understanding. It is an open question whether this would also apply when teachers make judgments of their own students (of whom they might—in addition to such non-predictive cues—also have knowledge that could be predictive).

Moreover, the question on what cues teachers should focus to increase the accuracy of their judgments has hardly been addressed to date. Research on *student* judgments of their own understanding has shown that redirecting students' attention to products of generative activities (see Fiorella & Mayer, 2015) improved students' judgment accuracy of their text understanding compared to students who were not encouraged to engage in such generative activities. The generating activities consisted of generating keywords (De Bruin, Thiede, Camp, & Redford, 2011), self-explanations (Griffin, Wiley, & Thiede, 2008), making diagrams (Van Loon et al., 2014), writing summaries (Thiede et al., 2010), or making concept maps (Thiede et al., 2010).

Teachers can also obtain cues from students' answers that result from written or oral generative activities (i.e., answer cues). Some evidence that focusing on answer cues improves teachers' judgment accuracy of students' mathematical understanding comes from a study by Thiede et al. (2015). They examined whether teachers' judgment accuracy of students' mathematical understanding was affected by involvement in a professional development program. This program stimulated teachers to focus more on student products that give insight into student thinking (e.g., by asking students to articulate their way of reasoning) during teaching. Teachers who took part in the program indeed made more accurate judgments of students' mathematical computational skills and conceptual understanding than teachers who did not participate in the training program. Nevertheless, judgment accuracy for students' conceptual understanding was still quite poor after participation (gamma correlation: .20; 1 would mean perfect prediction). Besides, it remains unclear whether it was indeed the increased focus on answer cues, or some other aspects of the 45-h training that caused the improvement in teachers' judgment accuracy (e.g., improved mathematical content knowledge).

1.2. Present study

Accurately judging students' conceptual mathematics understanding seems a challenging task that needs further investigation. In the present study, we experimentally investigated whether giving teachers access to cues with a high expected predictive value (i.e., answer cues)—compared to student cues—would improve teachers' judgment accuracy of students' conceptual understanding of decimal magnitude. More specifically, the first Research Question is whether teachers' judgment accuracy is affected when answer cues are available, additional to or instead of student cues, compared to when only student cues are available.

To answer this question, we experimentally manipulated the availability of the different cue types by providing teachers with a name of a student from their own class (student cues only), the anonymized answers on decimal magnitude practice problems of one of their students (answer cues only), or both the student's name and his/her answers (student + answer cues). We measured teachers' judgment accuracy of students' decimal magnitude understanding by comparing teachers' item-specific predictions of how well their students would perform on a decimal magnitude test with students' actual performance on such a test. The material (i.e., the first assignment consisted of practice problems and the second assignment consisted of items of which teachers had to predict students' performance) was created in such a way that analysis of students' answers on the first assignment could provide teachers with information on students' (mis)conceptions in the domain of decimal magnitude. An example of such a misconception is that students think of decimals as if they are whole numbers (e.g., 0.35 is greater than 0.8; see Durkin & Rittle-Johnson, 2015; Isotani et al., 2011).

With regard to the first Research Question we hypothesized that teachers' judgment accuracy would be lower in the name-only (i.e., student cues) condition than in both the answers-only (i.e., answer cues) and the name + answers (i.e., student + answer cues) condition, because answer cues can be expected to be more predictive of students' performance on the test assignment than the more general student characteristics on which teachers had to rely in the name-only condition (see Section 1.1.2). When judging their own students instead of fictional students (the latter was the case in the study by Kaiser et al., 2015), teachers may have knowledge about their students that could be predictive of students' actual understanding (e.g., knowledge on students' general conceptual mathematics understanding). However, knowing the student's name will likely also activate non-predictive student cues, whereas students' answers on practice problems are more directly associated with their understanding and as a result more predictive.

Given that simulated classroom research with fictional students showed that teachers' judgment accuracy was impaired when student characteristics were available (Kaiser et al., 2015), the second question we addressed is whether the accuracy of teachers' judgments of their own students' understanding is affected when only answer cues are available compared to when both student and answer cues are available. Focusing on student cues with presumably low predictive value might interfere with thorough or full analysis of students' answers on the first assignment. Hence, the second hypothesis we test in the present study is that teachers make more accurate judgments in the answers-only condition than in the name + answers condition.

As we expect that hypothesized differences in judgment accuracy between conditions would be due to differences in teachers' cue use across conditions (see Section 1.1.2), the second aim of this study is to explore differences in teachers' cue utilization between conditions. The third Research Question we addressed is: (How) do the cues that teachers use when making judgments of students decimal magnitude understanding differ across the name-only, name + answers and answers-only conditions?

2. Methods

2.1. Participants

2.1.1. Teachers

Twenty-one teachers (17 female) from 17 different primary schools in the Netherlands, teaching in fourth grade (i.e., Dutch "group six") volunteered to participate in this study. They were between 25 and 54 years old ($M_{age} = 36.34$, $SD = 8.99$) and had between three and 33 years of teaching experience ($M = 10.33$, $SD = 7.60$). They had been teaching their classes between two and five days a week ($M = 3.88$, $SD = 1.24$) from the beginning of the school year (i.e., end of August; data collection took place in October and November 2016). Six of them had been teaching the students in their class in a previous grade as well. Eight teachers had completed additional mathematics education courses after graduating from regular teacher training (e.g., on serious mathematics problems/dyscalculia or courses required to become the school's mathematics specialist).

2.1.2. Students

From the 454 students who attended the 21 participating fourth grade classes, 418 were included in the study (224 girls, $M_{age} = 9.55$, $SD = 0.42$). Students were excluded because of following special mathematics programs ($n = 20$), no parental consent to use students' data ($n = 11$), large portions of incomplete assignments ($n = 2$), or because their teachers accidentally saw their answers during task completion ($n = 3$). From this sample, three students in each condition (hereafter: "target students") were selected per teacher (i.e., nine in total per teacher). Based on students' test performance, a low, medium, and high performing student was selected per condition (see Section 2.5). Due to time restrictions (see Section 2.5) 13 of the target students (max. 2 per teacher) were dropped from the procedure. This resulted in a final sample of 176 students about whom the teachers made judgments (82 girls; $M_{age} = 9.59$, $SD = 0.42$; name-only condition: $n = 62$; name + answers condition: $n = 57$; answers-only condition: $n = 57$). At the time the study took place, decimal magnitude had not yet been taught (this is not done before the end of fourth grade), so the topic was new to almost all students.

2.2. Design

This study had a within-subjects design, with all 21 teachers making judgments of students' decimal magnitude understanding under three conditions: 1) name only (teachers were only provided with student names), 2) name + answers (teachers were provided with student names and students' answers to prior practice problems), and 3) answers only (teachers were provided with anonymized answers only). Teachers made judgments of three students per condition while thinking aloud.

2.3. Materials

Students were provided with instructions and assignments in the domain of decimal magnitude. Student (mis)conceptions in the domain of decimal magnitude are clearly defined. Five common and persistent misconceptions are: (1) thinking of decimals as if they are whole numbers (e.g., 0.35 is greater than 0.8 because 35 is greater than 8); (2) ignoring a zero that is in the tenths place (e.g., 0.08 is the same at 0.8); (3) assuming that adding a zero at the end of the decimal increases its magnitude (e.g., 0.30 is greater than 0.3); (4) viewing decimals less than one as being less than zero or more than one (e.g., 0.2 is less than 0); and (5) treating decimals as fractions thus thinking that numbers with more decimals are smaller (e.g., 0.852 is smaller than 0.3; see Durkin & Rittle-Johnson, 2015; Isotani et al., 2011).

2.3.1. Introductory video lesson

In an introductory video lesson, the topic of decimal magnitude was introduced to the students by explaining the place values of the tenths, hundredths and thousandths on a number line by connecting its meaning to fractions. No explicit attention was paid to specific misconceptions. Moreover, the study procedure was explained to the students in the video. The video had a total duration of 8:30 min. This video was created by the first author—who is also a primary school teacher—based on the most commonly used Dutch mathematics textbooks.

2.3.2. Student assignments

Students' answers on the first assignment (i.e., practice problems) functioned as a product of student generative activities that may give insight into student thinking. The first assignment consisted of 16 number line problems on decimal magnitude (nine multiple choice and seven open problems). The assignment was constructed such that each wrong answer was indicative of a particular misconception. For instance, when students placed 0.07 near the location of 0.7 on the number line, they were considered to hold the "ignoring the zero in the tenths place" misconception (the complete assignment, including indication of the misconceptions, is provided as online supplementary material; Appendix B). For some items, multiple answer options indicated the same misconceptions. For other items, different answer options indicated a different misconception. In total, each misconception could become evident four or five times. The items were based on examples from earlier research about student misconceptions in the field of decimal magnitude (Adams et al., 2014; Durkin & Rittle-Johnson, 2012, 2015; Rittle-Johnson, Siegler, & Alibali, 2001).

The second assignment (i.e., test problems) consisted of 17 decimal magnitude problems; five number line problems and 12 word problems (15 multiple choice and two open problems; all included as online supplementary material; Appendix B). Because the format of the items differed substantially between the first and second assignment, teachers had to use their interpretations of student thinking (i.e., students' misconceptions) when making judgments in the name + answers and answers-only condition (i.e., they could not directly translate correctness of an item in the first assignment into a judgment on the correctness of a particular item in the second assignment). The items in the second assignment were also based on examples from earlier research about student misconceptions in the field of decimal magnitude (Adams et al., 2014; Durkin & Rittle-Johnson, 2012, 2015; Rittle-Johnson et al., 2001). In total, each misconception could become evident three to five times. Two of the items were considered to assess students' overall understanding of decimal place values. Students' performance on the second assignment was scored by assigning one point for each correct answer (min = 0, max = 17).

Correlations between each of the misconceptions at the first assignment and the same misconception at the second assignment (measured by the number of errors indicative of the misconception) were significant and ranged from low to moderate ($r_{whole\ number} = .51$, $r_{ignoring\ zero\ in\ tenths\ place} = .49$, $r_{fraction} = .41$, $r_{outside\ 0\ and\ 1} = .27$, $r_{zero\ at\ end\ makes\ bigger} = .52$, for all $p < .001$), meaning that the answer cues on the first assignment have (modest) predictive value for performance on the second assignment.

2.4. Teachers' judgments

Teachers were asked to make item-specific judgments about the performance on the second assignment of the nine target students

from their classroom; they saw the 17 test items and indicated for each item whether they thought that the student had answered it correctly or incorrectly (see Fig. 1). For the three students in the name-only condition, teachers had to make these item-specific judgments knowing only the name of the students. For the three students in the name + answers condition, teachers were provided with students' names and students' answers on the practice assignment. Finally, for the three target students in the answers-only condition, teachers had to make the judgments seeing only students' answers on the practice assignment. Note that the completed practice assignment did not trace back to specific students, since students all used the same pencil type and were instructed not to write on the test sheets except for marking the place on the number line or answer option they thought was correct.

Most prior studies treat teachers' judgment accuracy as a single process through which teachers both judge what students do understand and do not understand (e.g., Kaiser et al., 2015; Südkamp et al., 2012; Thiede et al., 2015). In line with recent studies on students' own judgment accuracy we applied a two-process model, focusing separately on judgments of what students do understand (called "sensitivity" or "certainty") and judgments of what students do not understand (called "specificity" or "uncertainty"; cf. Rutherford, 2017; Schraw, Kuch, & Gutierrez, 2013). Knowing what students understand seems necessary for teachers to be able to anchor instructions and tasks to concepts and procedures already mastered by students; knowing what students do not understand (i.e., which misconceptions they have or where gaps in their knowledge lie) seems, for instance, necessary to give adequate additional instruction. Instruction will only foster conceptual change when it addresses the specific (mis)conceptions held by students (Prediger, 2008). Modeling both sensitivity and specificity allows examination of the potentially different processes surrounding accurate teacher judgments.

Sensitivity was calculated by first counting the number of items that were answered correctly by a student and were judged accurately (i.e., judged as correct) by the teacher. This number was then divided by the total number of items answered correctly by that student. For instance, when a student answered 10 of the 17 items correctly and the teacher judged five of the 10 correctly answered items accurately as being correctly answered, the sensitivity value was 0.5. Specificity was calculated by dividing the number of items that were answered incorrectly and were judged accurately by the teacher as being incorrectly answered, by the total number of items answered incorrectly. Consequently, teachers received a score between 0 and 1 for sensitivity and specificity for each student, with 0 indicating that none of the correct (sensitivity) or incorrect (specificity) items was judged accurately, and 1 indicating that all of the correct (sensitivity) or incorrect (specificity) items were judged accurately. For 11 of the 176 students (6%) we could not compute a specificity measure because they answered all items correctly. Because those data were Missing Not At Random (MNAR), listwise deletion or methods such as multiple imputation could not be applied (Van Buuren, 2012). We decided to assign these students value 1 for specificity (the maximum value), since these students had zero incorrect answers and teachers as a matter of fact judged 100% of this number of incorrect answers also as incorrect. To check whether this decision affected the results, we additionally conducted the main analyses with listwise deletion (another method to deal with missing data) of the students who answered all items correctly. Listwise deletion led to the same pattern of results as those presented in Section 3.1.1.

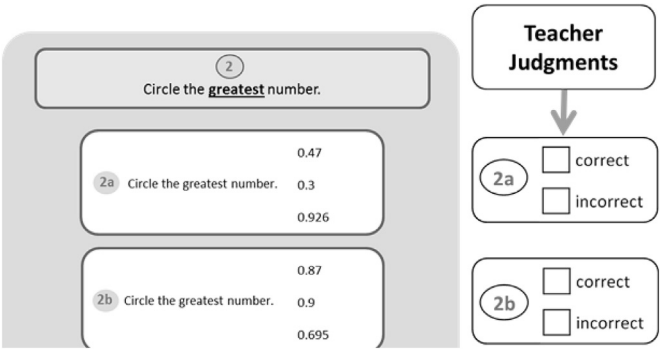

Fig. 1. Fragment of teachers' judgment task (translated from Dutch).

2.5. Procedure

The study procedure consisted of a student and a teacher part. The student part took place during a normal lesson day and lasted about 45 min. Students were informed that they would see an introductory video and make some tasks on the novel topic of decimal magnitude. They were also informed that they would not receive a grade for their work, but were encouraged to try their best on the assignments. Then, the introductory video was shown, after which students individually worked on the first and second assignment. Their teachers were present during the lesson, but had been instructed not to help students, answer questions, or look at students' answers (to prevent them from obtaining specific knowledge of some students' decimal magnitude understanding). Researchers only answered student questions that were not related to the mathematical content. After the both assignments had been completed, the student work was collected and the second assignment of each student was scored by the researchers. To ensure that teachers judged students with varying understanding of decimal magnitude, we divided all students within one class into three groups. The groups were based on the expected amount of decimal magnitude misconceptions students held, as represented by their performance on the second assignment. We applied the following distinction: high score = 14–17 (expected to hold no or one misconception), medium score = 10–13 (expected to hold two misconceptions), and low score = 0–9 (expected to hold more than two misconceptions). Nine target students per teacher were then selected: three low, three medium, and three high scoring students. In each judgment condition, teachers would encounter one student with a high, one with a medium, and one with a low score. When there were not enough students in a score category, a student from another category with the nearest score was selected (e.g., when there were not enough students in the high category a student from the medium category with a score of 13 was placed in the high category). The average test scores of the selected students were comparable across conditions; $M_{\text{name-only}} = 10.68$ ($SD = 3.61$), $M_{\text{name+answers}} = 10.79$ ($SD = 3.72$), and $M_{\text{answers-only}} = 10.79$ ($SD = 3.80$).

After the students went home, the teacher part started. Teachers first completed the first and second assignment themselves to become familiar with the assignments. Then, they judged the target students' performance on the second assignment, by condition: first for the three students in the name-only condition, then for the three students in the name + answers condition, then for the three students in the answers-only condition. Although the order of low/medium/high performing students was randomized within the conditions, the order of conditions was fixed to avoid that teachers in the name-only and name + answers conditions would be triggered to use other cues (e.g., related to observed student performance on other mathematics tasks) than they would normally do. Note that data from a prior study on text comprehension judgments (Van de Pol, de Bruin, van Loon, & van Gog, 2017), showed no

signs of a learning effect (i.e., teachers' judgment accuracy does not increase as they gain more experience with making judgments). Prior to the name-only and name + answers condition, the teachers were familiarized with the condition-specific judgment procedure in a practice phase with five test items that they had to judge for one of their non-target students. Because a pilot study had shown that teachers became less alert after one hour of judging, the judgment phase was ended after one hour. When the researchers, in the beginning of the judgment process, noticed that a participant was relatively slow, they decided to drop a student from the name + answers and/or answers-only condition (see for numbers Section 2.1.2 or Table 2).

Teachers were asked to think aloud while making the judgments, to gain insight into teachers' cue utilization (cf. Cooksey et al., 2007). The participants were prompted to continue thinking aloud when they were silent for five seconds or more, but were not asked for clarifications or elaborations as this might interfere with the cognitive processes involved in making the judgments (Ericsson & Simon, 1993; Van Someren, Barnard, & Sandberg, 1994). Research has shown that thinking aloud does not affect marking processes (which are presumably closely related to judgment processes; Crisp, 2008) or change the course or structure of thought processes in general (Ericsson & Simon, 1993; Van Someren et al., 1994). Although think-aloud protocols can slow down the process and probably do not reflect *all* of a person's thoughts, they do provide more information on cognitive processes than most other methods such as prospective interviews or self-reports (e.g. Ericsson & Simon, 1993; Van Someren et al., 1994). The 21 think-aloud protocols were audio-recorded, transcribed, and anonymized.

2.6. Data analysis

2.6.1. Analyses of accuracy differences across conditions

To investigate the effects of availability of only student cues, only answer cues, or both student and answer cues on teachers' judgment accuracy (Research Question 1 and 2) we performed a multilevel regression analysis in Mplus version 8 (Muthén & Muthén, 1998–2017). To account for the nested data structure with students (level 1) clustered in classes and thus in teachers (level 2), the "Complex" function in Mplus was used with maximum likelihood estimation with robust standard errors (MLR). A regression analysis with two outcome measures (i.e., sensitivity and specificity) was applied, because sensitivity and specificity correlated significantly with each other, $r_{zero\text{-}order} = -.353$, $p < .001$. The predictor variable condition was added using dummy coding.

2.6.2. Coding the think-aloud protocols

In order to investigate which cues teachers used (Research Question 3) we analyzed their think-aloud data. The 176 think-aloud transcripts were coded to identify the cues teachers reported while making their judgments.

First, to ensure a systematic segmentation procedure independent of coding categories, we defined a unit of analysis as "a sentence or part of a compound sentence that can be regarded as meaningful in itself, regardless of the meaning of the coding categories" (Strijbos, Martens, Prins, & Jochems, 2006, p. 37). A subsample of six transcripts (two from each condition, randomly selected) were independently segmented by two coders (the first author and a research assistant). The proportion agreement was determined from the perspective of each coder serving as an upper and lower bound of the 'true' agreement (cf. Strijbos et al., 2006). The proportion agreement had a lower bound of 88.9% and an upper bound of 90.4%, both above the threshold of 80% (cf. Strijbos et al., 2006). The coders respectively segmented the transcripts into 422 and 415 segments. In case of disagreement, the coders reached consensus on the segmentation through discussion.

The transcripts were coded in three steps. The final coding scheme including descriptions and examples can be found in Appendix A. For the first step of open coding 12 transcripts (4 from each condition) were used, across which 64 different codes were identified. Next, we divided these codes into categories, resulting in a coding scheme of 26 categories. We then checked whether these categories sufficed by applying this coding scheme to 6 further transcripts (2 from each condition). To check the interrater reliability, the two coders independently coded 10% of the transcripts (18 transcripts, 6 from each condition, in total 1124 segments). The interrater reliability was sufficient to high ($\kappa = .79$, agreement = 81.9%; Landis & Koch, 1977). In case of disagreement, the coders reached consensus on the coding through discussion.

For the analyses, these 26 categories were aggregated into six main categories: Content, Student, Answers, Student*Content, Teacher, and Miscellaneous. The first four categories refer to the information sources teachers presumably used in their judgments; these four were included in the analyses. The Content category included codes of statements related to curriculum content and material content (e.g., statements about what was or was not yet taught in the curriculum thus far, or about item characteristics of the first or second assignment). The Student category included codes assigned to statements about student characteristics (e.g., statements related to students' general cognitive ability or students' motivation). One particular interesting subcategory of the Student category was "fabricated student", assigned to statements occurring in the answers-only condition implying that teachers had an idea about the identity of the student or tried to guess the identity. Codes within the Answers category were based on students' answers on the practice assignment (e.g., statements related to a student's performance on one item or a group of items). Codes within the Student*Content category could be based on the student, the content material or the answers, or a combination of these, but always reflected an interaction between the student and the content (e.g., statements referring to a decimal misconception held or a strategy used by a student).

The two remaining categories (i.e., Teacher and Miscellaneous) included codes that were irrelevant for answering our research questions. The Teacher category included statements about teachers' emotions or meta-thoughts about the judgment process. The Miscellaneous category included all other irrelevant codes (e.g., unclear statements).

When teachers received the same code on multiple sequential segments, for example: "I think it took student x quite a while. I saw that it took her a long time", we would count this code only once. Hence, each segment was additionally coded with regard to repetitions. When one of the 26 codes from the coding scheme was repeated within one completed argumentation of a teacher (i.e., describing a student before starting with the judgments, or when analyzing a student's answer on one item or judging one item) this was also coded as "repetition". Repetitions were excluded from the frequency statistics as presented in the Results Section and also excluded from the analyses. The reliability of applying the repetition codes was determined by independently coding the repetition dimension of 6 transcripts (2 from each condition, in total 511 segments) that were already segmented and coded with codes from the coding scheme by one of the coders. The interrater reliability for the repetition dimension was very high ($\kappa = .93$, agreement = 97.5%).

After reliability was checked, the rest of the data (including the data that was used for developing the coding scheme) was segmented and coded definitively. The two coders each coded half of the data. In cases of doubt about segmenting or what code to

apply the coders reached consensus on the coding through discussion. After the coders each coded the transcripts of three teachers, they calibrated by independently coding three segmented transcripts and discussing the cases of disagreement until consensus, before continuing with the next three teachers.

2.6.3. Analyses of differences in cue utilization across conditions

To investigate the effects of the availability of only student cues, only answer cues, or both student and answer cues on teachers' cue utilization (Research Question 3), we performed multilevel regression analysis comparable to one conducted to investigate the accuracy differences (see Section 2.6.1). Instead of sensitivity and specificity the average frequencies of the four relevant main categories per student (excluding repetitions) were included as outcome measures.

3. Results

3.1. Teachers' sensitivity and specificity

Table 1 displays a cross tabulation of teachers' item specific judgments and students' actual item performances, including student and teacher totals. Table 2 displays teachers' average sensitivity and specificity values per condition. Teachers on average judged 8.15 (75%), 7.63 (69%), and 7.09 (64%) of students' correctly answered items as correct (i.e., sensitivity), and 2.63 (41%), 3.37 (48%), and 4.16 (64%) of students' incorrectly answered items as incorrect (i.e., specificity) in the name-only, name + answers, and answers-only condition, respectively.

3.1.1. The effect of cue-type availability on teachers' sensitivity and specificity

Table 3 displays the results of a multilevel analysis on sensitivity and specificity including condition as predictor. Regarding our first Research Question, we tested whether teachers' sensitivity and specificity was higher in the name + answers and answers condition, than in the name-only condition. Comparison of the name-only and the name + answers condition did not show significant differences in teachers' sensitivity and specificity in those conditions (sensitivity: $p = .066$; specificity: $p = .120$). Thus, gaining access to students' answers on practice problems did not significantly improve teachers' judgments of what students did and did not understand compared to when teachers could solely rely on their general knowledge of the students. Comparison of the name-only and the answers-only condition showed a significant difference in sensitivity ($p = .014$), but not in the expected direction: sensitivity was higher in the name-only condition. The regression coefficient shows that teachers' sensitivity increased with 0.11 when teachers made judgments in the name-only condition compared to the answers-only condition and that the effect size (0.48) was small to medium (cf. Cohen, 1992). In line with our hypothesis, though, teachers' specificity was higher in the answers-only than in the name-only condition ($p < .001$). Thus, teachers were more accurate at indicating what students did understand, but less accurate at indicating what students did not understand when they could only rely on general knowledge of their students (triggered by access to students' names) than when they could only rely on students' anonymized answers on practice problems. The regression coefficient shows that teachers' specificity increased with 0.24 when teachers made judgments in the answers-only condition, compared to the name-only condition and that the effect size (0.73) was medium to large.

Regarding the second Research Question, contrary to our hypothesis, the analysis showed that teachers' sensitivity in the answers-only condition did not differ significantly from the name + answers condition ($p = .265$). In line with our hypothesis, however, teachers' specificity in the answers-only condition was significantly higher than in the name + answers condition ($p < .001$). Thus, the teachers were better able to indicate what students did not understand, when they could only see students'

Table 2
Mean sensitivity and specificity values per condition.

Condition	N	Sensitivity (*SD*)[a]	Specificity (*SD*)[a]
Name-only	62	.75 (.20)	.41 (.31)
Name + answers	57	.69 (.22)	.48 (.32)
Answers-only	57	.64 (.24)	.64 (.31)

[a] min. = 0, max = 1.

Table 3
Parameter estimates from a multilevel analysis on teachers' sensitivity and specificity.

Effects	*B*	*SE B*	Cohen's *d*	*p*
Sensitivity				
Name-only vs. Name + Answers	0.06	0.03	0.27	.066
Name + answers vs. Answers-only	0.05	0.04	0.20	.265
Name-only vs. Answers-only	0.11	0.04	0.48	.014*
Specificity				
Name-only vs. Name + Answers	−0.08	0.05	−0.24	.120
Name + answers vs. Answers-only	−0.16	0.04	−0.49	<.001*
Name-only vs. Answers-only	−0.24	0.07	−0.73	<.001*

Note. *This effect significantly differed from zero when applying Bonferroni correction for multiple hypotheses testing, using an alpha level of 0.05/3 = 0.017.

Table 1
Cross tabulation of teachers' item-specific judgments and students' actual test assignment scores, including student and teacher totals.

	Student correct (*SD*)	Student incorrect (*SD*)	Total Teacher (*SD*)
Name-only			
Teacher correct	8.15 (3.88)	3.69 (2.53)	11.84 (3.66)
Teacher incorrect	2.53 (2.28)	2.63 (2.75)	5.16 (3.66)
Total Student	10.68 (3.61)	6.32 (3.61)	10.77[a] (2.78)
Name + Answers			
Teacher correct	7.63 (3.92)	2.84 (1.74)	10.47 (3.68)
Teacher incorrect	3.16 (2.23)	3.37 (3.14)	6.53 (3.68)
Total Student	10.79 (3.72)	6.21 (3.72)	11.00[a] (2.41)
Answers-only			
Teacher correct	7.09 (4.11)	2.05 (1.69)	9.14 (4.00)
Teacher incorrect	3.70 (2.84)	4.16 (3.30)	7.86 (4.00)
Total Student	10.79 (3.80)	6.21 (3.80)	11.25[a] (2.87)

Note. Numbers represent the absolute number of items judged as, or answered correctly/incorrectly.
[a] Average number of items (answered correctly and incorrectly by students) that teachers judged accurately.

answers on practice problems (i.e., anonymized) than when they knew the name of the student who produced these answers. The regression coefficient shows that teachers' specificity increased with 0.16 when teachers made judgments in the answers-only condition, compared to the name + answers condition and that the effect size (0.49) was small to medium.

3.2. Cues reported by teachers

In Table 4, all cues reported by teachers are displayed, including frequencies and proportions, excluding the segments coded as repetition. In the description of the results we only focus on the relevant codes. Fig. 2 shows a frequency distribution of the main categories across conditions. In the name-only condition, teachers reported most cues from the Student*Content category ($M = 9.23$, $SD = 5.90$). (Mis)conception was the most frequent code of all relevant codes in this condition ($M = 3.48$, $SD = 3.42$). In the name + answers condition, Student*Content was also the most frequent main category ($M = 14.46$, $SD = 7.34$). In this condition, item performance was the most frequent code ($M = 8.47$, $SD = 5.70$). In the answers-only condition, Answers was the most frequent main category ($M = 12.95$, $SD = 6.55$). As in the name + answers condition, item performance was the most frequent code ($M = 10.07$, $SD = 5.58$). Although one might not

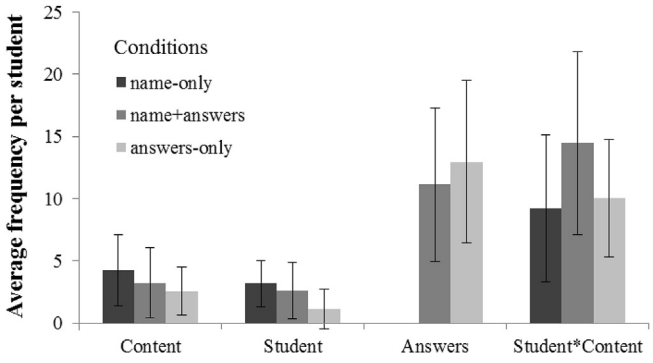

Fig. 2. Effects of cue availability on the frequency of cues reported by teachers. Error bars indicate standard deviations.

Table 4
Average frequencies and proportions of codes and main categories per student.

Assigned Codes	name-only		name + answers		answers-only	
	mean #	%	mean #	%	mean #	%
Relevant codes[a]						
Content						
Item characteristics	2.74	0.16	2.67	0.08	2.37	0.09
Curriculum	0.29	0.02	0.12	<0.01	0.05	<0.01
Instruction this lesson	1.19	0.07	0.44	0.01	0.14	0.01
Total Content	*4.23*	*0.25*	*3.23*	*0.10*	*2.56*	*0.10*
Student						
General cognitive	0.66	0.04	0.39	0.01	0.04	<0.01
Math general	0.60	0.04	0.63	0.02	0.19	0.01
Other math domain	0.19	0.01	0.12	<0.01	0.02	<0.01
Effort and work regulation	1.00	0.06	0.65	0.02	0.26	0.01
Affective	0.40	0.02	0.40	0.01	0.09	<0.01
Class behavior	0.02	<0.01	0.02	<0.01	0.00	0.00
Background	0.02	<0.01	0.00	0.00	0.00	0.00
Gender	0.00	0.00	0.00	0.00	0.05	<0.01
Student other	0.31	0.02	0.40	0.01	0.07	<0.01
Fabricated student	0.00	0.00	0.00	0.00	0.39	0.01
Total Student	*3.19*	*0.19*	*2.61*	*0.08*	*1.11*	*0.04*
Answers						
Item performance	0.00	0.00	8.47	0.27	10.07	0.38
Overall test performance	0.00	0.00	2.63	0.08	2.88	0.11
Total Answers	*0.00*	*0.00*	*11.11*	*0.35*	*12.95*	*0.49*
Student*Content						
Understanding decimals	3.35	0.20	4.53	0.14	3.09	0.12
Strategy	1.95	0.12	1.42	0.05	0.35	0.01
(Mis)conception	3.48	0.21	7.72	0.25	5.98	0.22
Student guessed	0.24	0.01	0.35	0.01	0.35	0.01
Comparison other student	0.19	0.01	0.44	0.01	0.26	0.01
*Total Student*Content*	*9.23*	*0.55*	*14.46*	*0.46*	*10.04*	*0.38*
Total relevant codes	*16.64*		*31.40*		*26.65*	
Irrelevant codes[b]						
Teacher						
Affective teacher	0.21	0.01	0.86	0.01	0.39	0.01
Meta process teacher	1.65	0.04	2.12	0.04	1.98	0.04
Guessing	0.23	0.01	0.25	<0.01	0.19	<0.01
Total Teacher	*2.08*	*0.05*	*3.23*	*0.05*	*2.56*	*0.05*
Miscellaneous						
Judgment	15.06	0.38	15.65	0.26	14.96	0.29
Other	5.23	0.13	7.28	0.12	5.32	0.10
Unclear	0.98	0.02	1.68	0.03	1.56	0.03
Total Miscellaneous	*21.27*	*0.53*	*24.61*	*0.42*	*21.84*	*0.43*
Total all codes	*40.00*		*59.25*		*51.05*	

[a] Proportions reflect proportion of "total relevant codes".
[b] Proportions reflect proportion of "total all codes". Repetitions are excluded.

expect student cues to be reported at all in the answers-only condition, teachers sometimes ($M = 1.11$, $SD = 1.62$) reported the "fabricated student" code (i.e., statements implying that teachers had an idea about the identity of the student or tried to guess the identity). Teachers occasionally also assigned characteristics to these fabricated students, and as a result also other student codes (e.g., "effort and work regulation") were reported sometimes in the answers-only condition.

3.2.1. Cue differences across conditions

The third Research Question was whether, and if so how, the cues reported by teachers differed across the conditions. Table 5 and Fig. 2 display the frequency differences of the main categories across conditions. In describing the results, we focus on the significant differences. Content cues were reported most in the name-only condition, in which they were reported significantly more often than in the answers-only condition ($B = 1.66$, $d = .62$, $p < .001$), but not significantly more than in the name + answers condition ($B = 1.00$, $d = .38$, $p = .042$; not significant as Bonferroni correction for multiple hypotheses testing was applied, using an alpha level of $.05/3 = .017$). This suggests that teachers use more cues related to curriculum content and material content when only student names are available, compared to when they have access to anonymized students' answers on practice problems.

As one would expect, student cues were reported significantly less frequently in the answers-only condition than in the name + answers condition ($B = -2.09$, $d = -1.00$, $p < .001$) and the name-only condition ($B = -1.51$, $d = -.72$, $p < .001$). Vice versa, as one would expect answer cues were reported significantly more in the answers-only ($B = 12.95$, $d = 1.86$, $p < .001$) and name + answers ($B = 11.11$, $d = 1.44$, $p < .001$) conditions than in the name-only condition, in which teachers did not have access to answer cues. Even though answer cues were reported most in the answers-only condition, this was not significantly more often than in the name + answers condition ($B = 1.84$, $d = .24$, $p = .022$). Likewise, student cues were not reported significantly more often in the name-only condition than in the name + answers condition ($B = 0.58$, $d = .28$, $p = .132$), suggesting that teachers did not rely less on their general knowledge about students when they had access to students' practice answers in addition to their names.

Student*Content cues were reported most in the name + answers condition and this differed significantly from both the name-only ($B = 5.25$, $d = .81$, $p < .001$) and answers-only condition ($B = 4.42$, $d = .68$, $p < .001$).

Table 5
Parameter estimates from a multilevel analysis on the frequency of assigned codes.

Effects	B	*SE*	Cohen's *d*	*p*
Content				
Name-only vs. Name + Answers	1.00	0.50	0.38	.042
Name + answers vs. Answers-only	0.67	0.39	0.25	.086
Name-only vs. Answers-only	1.66	0.47	0.62	<.001*
Student				
Name-only vs. Name + Answers	0.58	0.38	0.28	.132
Name + answers vs. Answers-only	1.51	0.35	0.72	<.001*
Name-only vs. Answers-only	2.09	0.32	1.00	<.001*
Answers				
Name-only vs. Name + Answers	−11.11	1.22	−1.44	<.001*
Name + answers vs. Answers-only	−1.84	0.80	−0.24	.022
Name-only vs. Answers-only	−12.95	1.34	−1.68	<.001*
Student*Content				
Name-only vs. Name + Answers	−5.23	1.03	−0.81	<.001*
Name + answers vs. Answers-only	4.42	0.99	0.68	<.001*
Name-only vs. Answers-only	−0.81	0.91	−0.13	0.366

Note. *This effect significantly differed from zero when applying Bonferroni correction for multiple hypotheses testing, using an alpha level of 0.05/3 = 0.017.

4. Discussion

The first aim of the present study was to investigate whether teachers' judgment accuracy of students' conceptual mathematics understanding would be affected by manipulating the availability of students' names, answers on a prior practice assignment, or both. This would lead to the (un)availability of certain cues on which teachers' judgments could be based. Teachers' judgment accuracy was measured by teachers' item-specific judgments of what students do understand (sensitivity) and judgments of what students do not understand (specificity) within the domain of decimal magnitude.

Our first hypothesis was that teachers' sensitivity and specificity would be higher when having access to students' answers on a practice assignment (considered to be predictive of students' actual understanding and therefore, of their performance on the test assignment) compared to when having access to only student names (which would result in activation of cues that we expected to have low predictive value). Our second hypothesis was that teachers' sensitivity and specificity would be higher when having access to only students' answers, compared to when having access to both students' answers and names. Contrary to our hypotheses, teachers' ability to indicate what students did understand (sensitivity) was not higher when students' answers on prior practice problems were available; it was even significantly lower when only students' answers were available, compared to when only names were available. Partly in line with the hypotheses regarding specificity, teachers were better able to judge accurately what students did not understand when they had access to their answers on prior practice problems, but only when they did not know who the students were (increase of .24 on a scale from 0 to 1). Although these findings show that the types of cues that are available to teachers may affect their judgment accuracy (mainly in terms of specificity), it does not tell us which cues teachers used exactly.

Therefore, the second aim of our study was to explore how teachers' cue use differed depending on the information types that were available (i.e., student cues, answer cues or both). The analyses of teachers' think-aloud data, recorded while they made judgments, showed that teachers in all conditions used cues related to the content of the task at hand and curriculum content. Not surprisingly, when teachers had only access to student names, they made no use of information on students' answers (which they did not have access to). Surprisingly, however, when teachers did not have access to student names, but only to students' answers, they still made some use of student cues (although significantly less than when student names were available). Teachers hypothesized, for instance, from which student the answers were ("fabricated student cues") and even assigned features to the anonymous students, such as having sloppy habits, having low concentration, being clever, or being uncertain. Another finding that we did not anticipate was that teachers also used cues that reflected an interaction between the student and the content material (e.g., statements referring to a decimal misconception held or a strategy used by a student) of which it was mostly unclear whether these cues were derived from the student and/or the answers and/or the content material.

The differences in teachers' cue use across conditions can explain the differences in teachers' specificity, as we discuss in Section 4.2. First, however, we discuss the findings regarding sensitivity.

4.1. Effects of cue availability on sensitivity of teachers' judgments

The finding that the sensitivity of teachers' judgments was higher when they had only students' names available compared to

when they had only answer cues available was in contrast with our hypothesis (i.e., we expected that teachers' sensitivity would be higher when only answer cues were available). Rather than indicating that teachers made the most accurate item-specific judgments in the name-only condition, however, this finding probably reflects teachers' tendency to be more positive about their students' performance when they knew which student they were judging. Note that prior research, where the situations were comparable to our name-only condition, also showed that teachers generally overestimate their students (Artelt & Rausch, 2014; Klug, Bruder, & Schmitz, 2016). When teachers in the present study only knew the student's name, they judged, on average, almost 3 items more as having been answered correctly than when they only had students' answers on practice problems available. Given that students answered approximately two-thirds of the test items correctly, this means that when teachers would just randomly have assigned their "correct" judgments to the test items, the chance of judging a correct answer as correct was substantially higher in the name-only than in the answers-only condition.

4.2. Effects of cue availability on specificity of teachers' judgments

Specificity was higher, meaning that teachers were better able to accurately judge what students did *not* understand and would get wrong on the test assignment, when teachers had access to students' answers on prior practice problems, but only when they did not know who the students were. This finding may be explained by differences in teachers' cue use across conditions. We expected that having access to students' answers would result in more accurate judgments, because teachers would focus less on student cues and more on answer cues, the latter being presumably more predictive of students' actual understanding (see Section 1.1.2). Indeed, teachers reported using significantly more answer cues when answers were available compared to when answers were unavailable (i.e., name-only condition), but they did not use *more* answer cues in the answers-only compared to the name + answers condition, so this cannot explain the higher specificity in the answers-only condition compared to the name + answers condition. Teachers also used fewer student cues when names were not available (i.e., in the answers-only condition), which is an unsurprising finding. More interestingly, however, having access to students' answers in addition to their names, did not result in the use of *fewer* student cues than in the name-only condition. Findings of Kaiser et al. (2015) already indicated that teachers' judgments of fictional students' mathematics grades were impaired when being provided with student characteristics in addition to information on students' oral and written mathematics achievement. Our findings suggest that when teachers make judgments of their own students, focusing on student cues (triggered by access to student names) in addition to the answer cues may also interfere with adequately using the answer cues. The following quote of a teacher, taken from the name + answers condition in our study, illustrates that even though relevant cues (i.e., answer cues) were available, teachers may erroneously disqualify the relevant cues based on their knowledge about the student (i.e., student cues): "She places 0.13 …. ah that's interesting, she places it behind the one [teacher analyzes a student's practice problem]. So then she thinks … Well, that's sloppiness. I shouldn't take this one into account."

Another potential explanation for why specificity was higher in the answers-only condition than in the name-only condition might lie in differences in the use of content cues (i.e., cues related to curriculum content and material content). The findings show that teachers used significantly fewer content cues in the answers-only than in the name-only condition. According to Thiede et al. (2015), content cues are not predictive of students' actual understanding, leading to inaccurate teacher judgments. The same seems to apply to accuracy of students' own judgments: use of content cues led to less accurate judgments of their own text understanding (Thiede et al., 2010). In sum, making less use of student and content cues, and more use of answer cues, can explain why teachers' judgments of what students did not understand are most accurate when only having access to students' answers.

4.3. Limitations and future research

This study has several limitations. The decimal magnitude assignments included multiple-choice answers. The advantage of multiple choice was that it allowed us to construct the test in such a way that all potential misconceptions could be detected (not only the dominant ones). Unfortunately this also meant that students could correctly guess answers. This may have led to the relatively high test scores (with students correctly answering approximately two-thirds of the test items), which in turn may have affected the sensitivity measure of judgment accuracy (see also Section 4.1), and may explain why the answer cues in the present study had only modest predictive value for students' actual understanding. Another potential limitation is that item format (e.g., number lines vs. asking to circle a number) might affect predictive value. Because there was an imbalance between the number of number line and other tasks we could not reliably examine judgment accuracy by task type in the present study. Future research could further investigate if the specific format of the items would influence their predictive value, even if they test the same conceptual knowledge.

Nevertheless, even though they had only modest predictive value, teachers made more accurate specificity judgments when having access to the answer cues compared to only student names or to both answers and names. Hence, if answer cues can be defined in such a way that they have higher predictive value in future research, this can be expected to lead to even more accurate judgments and might provide teachers with useful tools that they (can) use in class when teaching mathematics (and other subjects) to monitor students' understanding and provide adaptive support.

Future research might as well consider including measures of teachers' knowledge of students' misconceptions, since more knowledge of misconceptions might lead to more accurate judgments (cf. Ostermann, Leuders, & Nückles, 2017). Finally, our sample was relatively small, even with a within-subjects design, so including a larger sample of teachers and students, would be desirable in future research.

4.4. Conclusions

As prior research indicated (Kaiser et al., 2015, 2017, 2013; Furnari et al., 2017; Hurwitz et al., 2007; Paleczek et al., 2017; Ready & Wright, 2011) teachers' knowledge of general student characteristics plays a major role in teachers' judgment processes. We examined how giving teachers access to students' answers on practice problems, additional to or instead of their general knowledge of specific students (triggered by access to students' names), affected teachers' judgment accuracy. The findings suggest that giving teachers access to the answers, in addition to knowledge of their students, does not make teachers focus less on student characteristics, and a result, does not significantly improve teachers' accuracy of students' decimal magnitude understanding. Giving teachers access to students' answers only (i.e., instead of their knowledge about students), seems to be especially effective for judging what a student does not yet understand. Our study shows that applying the cue-utilization approach in research on teachers' judgments may be a promising way to identify starting points for interventions for improving teachers' judgment accuracy, which

ultimately may foster the quality of teachers' instructional decisions.

Acknowledgments

This work was supported by the Dutch Ministry of Education, Culture and Science (grant number OCW/PromoDoc/1065001). The authors would like to thank Susan Ravensbergen for her help with data collection and analysis.

Appendix A. Coding scheme for the think-aloud transcripts

Codes per main category	Description	Example
Content		
Item characteristics	Statements about characteristics or features of the items in the assignment(s), such as difficulty or physical appearance of the numbers, answer options, or problem type. N.B. Statements related to students' (mis)conceptions do not belong to this category.	… because of the pyramid form of the answers … … this (item) is mean …
Curriculum	Statements about what was or was not yet taught in the curriculum thus far. N.B. Statements related to video instruction do not belong to this category.	Decimals are new to them. … we taught this (milliliters) somewhat.
Instruction this lesson	Statements related to the extent to what students paid attention to or remembered the video instruction prior to making the assignments.	Well, she pays attention to that kind of videos. She did get that explanation.
Student		
General cognitive	Statements related to students' cognitive ability or skills, in general or not specifically related to mathematics, such as students' intelligence, language skills or learning disorders.	It is a clever girl. He has dyslexia.
Math general	Statements related to students' general math ability.	Actually, he is quite good in math. … is one of my weak math students …
Other math domain	Statements related to students' skills in a specific mathematical domain, other than decimals, such as fractions, money, and geometry.	She is strong with fractions. There are some gaps in geometry, time, money. *Teacher is referring to a specific student.*
Effort and work regulation	Statements related to students' effort and regulation during working, such as speed, concentration, sloppiness and carefulness of working.	I think this has a lot to do with concentration. She is going to think really hard.
Affective	Statements related to students' emotions, motivation, and attitude, such as confidence, interest and stress.	… and so student x thinks … oh exciting! … and student x likes it …
Class behavior	Statements related to students' general classroom behavior, not specifically related to working.	That one has ADHD.
Background	Statements related to students' background characteristics and home conditions, such as SES and characteristics of the parents.	… because he is an immigrant … … or his parents do not speak much Dutch …
Gender	Statements related to a student's gender.	Well, I think, this is a he. Why do I assume this is a she?
Student other	Other statements about students that do not fall into one of the other categories. These are mostly very general statements.	That one is very unpredictable. Ok, that is a nice one. *Teacher is in both examples referring to a specific student.*
Fabricated student	Statements occurring in the answers-only condition implying that teachers had an idea about the identity of the student or tried to guess the identity.	I think somehow this is student x. I just immediately have a student in my mind.
Answers		
Item performance	Statements related to a student's performance on one item or a small group of items (max. five) in the first assignment, unrelated to students' strategy, understanding or (mis)conceptions.	He answered 1a correctly. It goes well half of the time. *Statement referring to student's performance on a subtask in the first assignment.*
Overall test performance	Statements related to the students' overall performance on the first assignment, unrelated to students' strategy, understanding or (mis) conceptions.	Here with practicing she is doing really bad. *The teacher is referring to the first assignment.* Well, her answers are so inconsequent.
Student*Content		
Understanding decimals	Statements related to students' prior knowledge or general understanding of decimal numbers.	No, this one does not really get how it works behind the decimal point. [… she chooses this one correctly], because this looks like what she knows …
Strategy	Statements related to the strategy or approach used by students, such as how to determine a position of a number on the number line or whether students use the strategy of adding digits to make two numbers equal of length.	… because she will puzzle on that number line like “four, oh that's less than five”. I think he will just add a digit.
(Mis)conception	Statements related to the specific decimal magnitude (mis) conceptions students might have.	… locating 0.08 as 0.8 … She sees 70 and 7 and thinks 70 is bigger than 7.
Student guessed	Statements reflecting that a teacher thinks a student guessed which answer is correctly, but that the student does not actually understand the content.	Maybe he guesses one correctly in this task … Or it is a coincidence (that the student made this item correctly) …
Comparison to other student	Statements referring to comparison of the student that is being judged to another student. N.B. We included this code in the Student*Content category, since teachers compared the students on characteristics related to the content, such as their understanding of decimals or misconceptions.	Student x also did that wrong … Well, when I assess this as correctly for the others [I should do it certainly for her].

(*continued*)

Codes per main category	Description	Example
Teacher		
Affective teacher	Statements reflecting teachers' affective experiences during the process, including statements about hope and astonishment.	… because I just hope she knows this … That is really frustrating. *Teacher is referring to her own emotions, not to those of a student.*
Meta process teacher	Statements related to teachers' meta thinking about the judgment process.	This is very hard. Or do I have to many high expectations?
Guessing	Statements reflecting that teachers do not know why they make certain judgments.	Well, I don't know why …. Then I am going to guess a bit …
Miscellaneous		
Judgment	Statements reflecting the mere prediction of a teacher about the students' correctness of a test item. N.B. When another code is also applicable to the segment that code is dominant and assigned instead of the judgment code.	She answers 2b incorrectly. … so I think he will choose the right answer option here.
Other	This code is assigned when another code does not apply, but when it is clear what a statement means. For example, a teacher reads aloud an item or poses a question to the researcher.	Well, larger than, smaller than …. Let's have a look at these answers …
Unclear	This code is assigned when it is not clear what a teacher's statement refers to. This mostly applies to incomplete statements.	And then he will … Well, indeed you see …

Appendix B. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.tate.2018.02.007.

References

Adams, D. M., Mclaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., et al. (2014). Computers in human behavior using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior, 36*, 401–411. https://doi.org/10.1016/j.chb.2014.03.053.

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments: When and for what reasons? In S. Krolak-Schwerdt, S. Glock, & M. Bohmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 229–248). Rotterdam, The Netherlands: Sense Publishers.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401–434. https://doi.org/10.1080/13803610701728311.

Crisp, V. (2008). The validity of using verbal protocol analysis to investigate the processes involved in examination marking. *Research in Education, 79*(1), 1–12. https://doi.org/10.7227/RIE.79.1.

De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. https://doi.org/10.1016/j.jecp.2011.02.005.

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction, 22*(3), 206–214. https://doi.org/10.1016/j.learninstruc.2011.11.001.

Durkin, K., & Rittle-Johnson, B. (2015). Diagnosing misconceptions: Revealing changing decimal fraction knowledge. *Learning and Instruction, 37*, 21–29. https://doi.org/10.1016/j.learninstruc.2014.08.003.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. London, England: MIT Press.

Fiorella, L., & Mayer, R. E. (2015). Eight ways to promote generative learning. *Educational Psychology Review, 28*, 717–741. https://doi.org/10.1007/s10648-015-9348-9.

Furnari, E. C., Whittaker, J., Kinzie, M., & DeCoster, J. (2017). Factors associated with accuracy in prekindergarten teacher ratings of students' mathematics skills. *Journal of Psychoeducational Assessment, 35*, 410–423. https://doi.org/10.1177/0734282916639195.

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers ' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49–60. https://doi.org/10.1016/j.learninstruc.2016.06.008.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*(1), 93–103. https://doi.org/10.3758/MC.36.1.93.

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly, 22*(2), 115–144. https://doi.org/10.1037/1045-3830.22.2.115.

Isotani, S., Adams, D., Mayer, R. E., Durkin, K., Rittle-Johnson, B., & McLaren, B. M. (2011). Can erroneous examples help middle-school students learn decimals? In D. Kloss, C. Gillet, D. Crespo García, R. M. Wild, & F. Wolpers (Eds.), *Proceedings of the 6th European conference on technology enhanced learning, 2011* (pp. 181–195). Palermo, Italy: Springer-Verlag Berlin Heidelberg, 2011.

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift Fur Erziehungswissenschaft, 18*(2), 279–302. https://doi.org/10.1007/s11618-015-0619-5.

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001.

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology, 109*(6), 871–888. https://doi.org/10.1037/edu0000156.

Karing, C., Pfost, M., & Artelt, C. (2011). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? *Journal for Educational Research Online, 3*(2), 119–147. Retrieved from http://search.proquest.com/openview/6acf3177c9040cd6b1d4245dfee5cc2a/1?pq-origsite=gscholar.

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*(1), 38–46. https://doi.org/10.1016/j.tate.2012.10.004.

Klug, J., Bruder, S., & Schmitz, B. (2016). Which variables predict teachers diagnostic competence when diagnosing students' learning behavior at different stages of a teacher's career? *Teachers and Teaching, 22*(4), 461–484. https://doi.org/10.1080/13540602.2015.1082729.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 3*(2), 363–374. Retrieved from http://www.jstor.org/stable/2529786?seq=1#page_scan_tab_contents.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Ostermann, A., Leuders, T., & Nückles, M. (2017). Improving the judgment of task difficulties: Prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education*. https://doi.org/10.1007/s10857-017-9369-z. Advance online publication.

Paleczek, L., Seifert, S., & Gasteiger–Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students: A multilevel analysis. *Psychology in the Schools, 54*(3), 228–245. https://doi.org/10.1002/pits.21993.

Prediger, S. (2008). The relevance of didactic categories for analysing obstacles in conceptual change: Revisiting the case of multiplication of fractions. *Learning and Instruction, 18*(1), 3–17. https://doi.org/10.1016/j.learninstruc.2006.08.001.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335–360. https://doi.org/10.3102/0002831210374874.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346–362. https://doi.org/10.1037/0022-0663.93.2.346.

Rutherford, T. (2017). Within and between person associations of calibration and

achievement. *Contemporary Educational Psychology, 49*, 226–237. https://doi.org/10.1016/j.cedpsych.2017.03.001.
Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007.
Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. (2006). Content analysis: What are they talking about? *Computers & Education, 46*(1), 29–48. https://doi.org/10.1016/j.compedu.2005.04.002.
Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627.
Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., … Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49*, 36–44. https://doi.org/10.1016/j.tate.2015.01.012.
Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331–362. https://doi.org/10.1080/01638530902959927.
Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC press.
Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., Van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007.
Van Someren, M. V., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical approach to modelling cognitive processes*. London: Academic Press.
Van de Pol, J., de Bruin, A., van Loon, M., & van Gog, T. (2017, August). The effect of cue availability on students' and teachers' judgment accuracy. In *Paper presented at the 17th biennial conference of the European Association for Research on Learning and Instruction, Tampere, Finland*.
Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6.
Van de Pol, J., Volman, M., & Beishuizen, J. (2011). Patterns of contingent teaching in teacher-student interaction. *Learning and Instruction, 21*(1), 46–57. https://doi.org/10.1016/j.learninstruc.2009.10.004.
Webb, M. B. (2015). *Exploring the correlation between teachers' mindset and judgment accuracy to reveal the cues behind teachers' expectations*. Doctoral dissertation. Boise, MT: Boise State University.
Whitmer, S. P. (1982, March). A descriptive multimethod study of teacher judgment during the marking process. In *Paper presented at the annual meeting of the American Educational Research Association: The many publics of education and educational research, New York City, NY*.