Teaching and Teacher Education 122 (2023) 103982

Contents lists available at ScienceDirect

Teaching and Teacher Education

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

Research paper

Effects of cue availability on primary school teachers' accuracy and confidence in their judgments of students' mathematics performance

Sophie Oudman*, Janneke van de Pol, Tamara van Gog

Department of Education, Utrecht University, the Netherlands

A R T I C L E I N F O

Article history:
Received 7 February 2022
Received in revised form
1 November 2022
Accepted 7 December 2022
Available online 20 December 2022

Keywords:
Teachers' judgment accuracy
Cue use
Cue diagnosticity
Confidence in judgment accuracy
Formative assessment
Primary mathematics education

A B S T R A C T

We investigated how the accuracy of teachers' judgments of their students' performance on procedural mathematical tasks, as well as their confidence in that, can be improved. Thirty-three primary school teachers judged how their students ($N = 553$) would perform on a multiplication and division task, with and without having access to performance cues (i.e., students' performance on similar tasks completed one week earlier). When available, teachers mostly seemed to base their judgments on performance cues. Availability of performance cues improved teachers' judgment accuracy, resulted in higher confidence in their judgment accuracy, and increased awareness of their judgment (in)accuracy.

To optimally stimulate student learning, teachers need to provide ‘differentiated’ instruction; instruction that is adapted to students' current level of performance or understanding (Van de Pol, Volman, & Beishuizen, 2010; Parsons et al., 2018; Tomlinson et al., 2003). In order to make adaptive instructional decisions, teachers' judgments of their students' performance or understanding—also called monitoring judgments—need to be accurate (see for empirical studies: Van de Pol, Volman, Oort, & Beishuizen, 2014; Klug et al., 2013; see for review studies Urhahne & Wijnia, 2021; Thiede et al., 2019). In the present study, we focus specifically on the accuracy of primary school teachers' judgments of students' mathematical performance, which are predominantly inaccurate and typically too optimistic (Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018; Gabriele et al., 2016; Thiede et al., 2015, 2018, 2019; Zhu & Urhahne, 2018). This is problematic as it can result in instructional decisions that are too optimistic, and thus, not adapted to a student's actual needs (Urhahne & Wijnia, 2021). For instance, teachers who overestimate their students' performance may provide them with tasks that are too difficult or fail to provide additional instruction to those who need it. This results in suboptimal learning progress and increases the likelihood of students failing at subsequent tasks, which may have adverse motivational and emotional effects (Seegers & Boekaerts, 1993).

To encourage teachers to make accurate monitoring judgments and in turn adaptive instructional decisions, the importance of *Data Based Decision Making* (DBDM; Campbell & Levin, 2009; Schildkamp et al., 2017) and *formative assessment* (Black & Wiliam, 2009; Van der Kleij et al., 2015) is increasingly emphasized in educational policy. DBDM interventions focus mostly on using data at a ‘macro level’, such as standardized assessments that students complete a few times a year, to make educational decisions with instructional purposes but also school development and accountability purposes (Schildkamp et al., 2017). Formative assessment focuses more on the ‘micro level’, that is, on eliciting and using student performance data to inform instructional decisions in the classroom, both during and in between lessons (Black & Wiliam, 2009). In line with the latter, in the present study, we investigate whether providing teachers with students' mathematical performance data at the micro level helps them more accurately predict students' future task performance. Prior studies showed that providing teachers with information on students' performance on a related prior task (which they can use in addition to their knowledge of students' general characteristics, such as nationality or learning problems, when making judgments), does not necessarily

* Corresponding author. Department of Education, Utrecht University, P.O. Box 80.140, 3508 TC, Utrecht, the Netherlands.
E-mail address: v.s.oudman@uu.nl (S. Oudman).

https://doi.org/10.1016/j.tate.2022.103982

lead to more accurate judgments (Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018, Van de Pol, Van Gog, & Thiede, 2021). It is important to gain knowledge about which type of tasks, that are part of current mathematics education, would provide teachers with performance data that increase the accuracy of their judgments.

Teachers need to make accurate judgments of students' performance, and, in order to make effective follow-up decisions, they also need to be *aware* of their (in)accuracy (Gabriele et al., 2016). That is, when teachers are (rightfully) confident that their judgment is accurate, they will carry out appropriate instructional actions based on those judgments; when they are not confident that their judgment is accurate, they can first seek more information about a student's performance before taking instructional actions (Gabriele et al., 2016). Therefore, a second aim of the current study was to explore how manipulating the availability of performance data affects teachers' awareness of their judgment (in)accuracy.

1. Literature review

Before introducing our research questions, we will first explain (1) how we define teachers' judgment accuracy and how this is influenced by the use of *cues* (i.e., specific pieces of information that can be used to inform judgments; Koriat, 1997; Cooksey et al., 2007), (2) how using certain cues and ignoring others can improve teachers' judgment accuracy, and (3) the concepts of teachers' *confidence in* and *awareness of* their judgment (in) accuracy.

1.1. Teachers' judgment accuracy and cue use

To assess the accuracy of teachers' judgments of their students' performance, different measures can be used (cf. Urhahne & Wijnia, 2021, for a discussion of different measures). Because our ultimate goal is to help teachers make more adaptive instructional decisions, we are mainly interested in *absolute accuracy*, defined as the absolute discrepancy between a teacher's judgment (i.e., prediction of how a student performs on a task) and student's actual performance on that task. For example, consider a teacher who expects that a student will correctly solve four out of ten problems, but the student only solves two out of ten problems correctly. In this case, the absolute deviation is two problems (4 minus 2) on a scale from zero to ten, with values closer to zero indicating higher accuracy. Prior research has also frequently used *bias*, that is, the signed version of absolute accuracy, indicating whether and how much teachers over- or underestimate their students' performance. Because overestimation and underestimation cancel each other out when averaging scores, this measure does not always reflect the extent to which judgments are actually accurate when using it in regression analyses. *Relative measures*, such as rank components, have also frequently been used in prior research and can be useful in educational contexts, for example to gain insight into which students are *most* in need of additional support. However, it is possible to make a perfect rank order in terms of students' performance, while, for instance, overestimating the actual performance of all students. When it comes to tailoring instructional activities to individual students' needs, for example, when deciding which students need additional instruction or which students are ready for a more difficult task (cf. mathematics lesson books such as Baak et al., 2018 and Borghouts et al., 2019, or EDI, a widely applied teaching model: Hollingsworth & Ybarra, 2018), teachers should be able to accurately judge students' actual task performance.

1.1.1. How cue use affects judgment accuracy

Unfortunately, teachers' monitoring judgments of their students' performance or understanding are often inaccurate. An explanation for why this is the case can be found in Koriat's (1997) cue-utilization perspective on monitoring accuracy. According to this foundational theory, judgments are based on specific pieces of information (i.e., cues) that differ in the extent to which they are actually predictive of students' performance.

When judging how well a student will perform on a future test or task, teachers can use different types of cues (Thiede et al., 2019). In this study, we specifically focus on student cues and performance cues. *Student cues* are general characteristics of students. Student cues that are repeatedly reported in prior studies are students' general cognitive ability, nationality, SES, sex or gender, classroom engagement, conscientiousness, disability status, self-concept, and interest (e.g., Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018, Van de Pol, Van Gog, & Thiede, 2021; Cooksey et al., 2007; Furnari et al., 2017; Gortazar et al., 2022; Helwig et al., 2001; Hurwitz et al., 2007; Johnston et al., 2019; Kaiser et al., 2013, 2015; Meissel et al., 2017; Paleczek et al., 2017; Ready & Wright, 2011; Zhu & Urhahne, 2020). *Performance cues* consist of information about students' prior performance on the same skills or content that the teacher is judging (cf. Van de Pol, Van Gog, & Thiede, 2021). For instance, teachers can decide which students do not yet master multi-digit multiplication (and need additional instruction) based on formative assessments such as, (1) students' scores on a task or test about the same problem type, completed one or multiple days earlier, or (2) how well students can answer practice problems during the whole-class instruction (Hollingsworth & Ybarra, 2018; Thiede et al., 2015; 2018).

According to Koriat's (1997) work on cue-utilization, teachers' judgments are more accurate when the cues being used are more *diagnostic*, that is, predictive of students' actual performance (cf. Thiede et al., 2019). The diagnosticity and use of cues can be graphically displayed by means of the Lens Model developed by Brunswik (1955; Fig. 1), in which the analogy of a convex lens is used to display the relations between a judgment (in this study: judgment made by the teacher), cues, and the true state (in this study: student's performance). The teacher only "sees" the student's achievement (or other student related variables) through the "lens" of the cues (see Urhahne & Wijnia, 2021, for a review of literature applying the lens model to research on teaching). Thus, the key to improving teachers' judgment accuracy lies in fostering their use of more diagnostic cues.

1.2. Improving teachers' judgment accuracy: the effect of cue availability

Prior studies suggest that performance cues are generally more diagnostic than student cues (in mathematics: Thiede et al., 2019; in text comprehension: Van de Pol, Van Gog, & Thiede, 2021). In many studies on teacher judgments, teachers had only student cues and no performance cues available (e.g., Furnari et al., 2017; Paleczek et al., 2017). Hence, it is not surprising that their judgments were often inaccurate. Providing teachers with information from which they can derive performance cues, or focusing their attention on available information from which they can derive performance cues, might improve teachers' absolute judgment accuracy of their students' mathematical performance. Prior studies, however, did not find systematic evidence for this effect. In two studies, Thiede et al. investigated whether increased use of formative assessment practices—aimed at collecting performance cues to guide instructional decisions—improved primary school teachers' judgment accuracy with regard to mathematics. In one of the studies, Thiede et al. (2018) found that an intervention aimed at increasing teachers' use of formative assessment did not result in an improvement of teachers' relative accuracy and bias. In the other

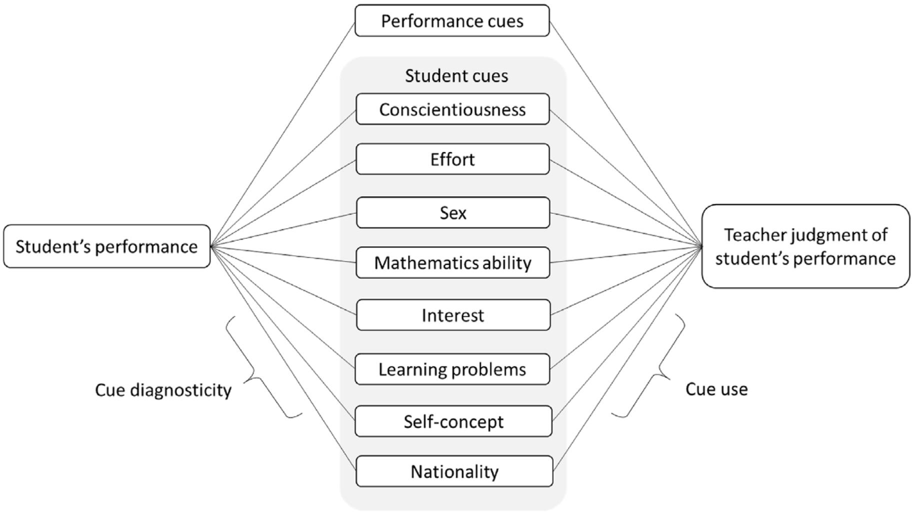


Fig. 1. Brunswik's lens model applied to the current study.
Note. Only cues included in the current study are displayed. The correlation between the cues and teacher judgment is an indication of teachers' cue use, the correlation between the cues and the student's actual performance is called cue diagnosticity.

study, Thiede et al. (2019) found that the observed frequency with which teachers used formative assessment practices during their mathematics lessons was related to teachers' relative judgment accuracy, but not to their bias. Neither of those studies measured absolute accuracy.

Whereas the two studies of Thiede et al. focused on formative assessment practices in general, Zhu and Urhahne (2018) investigated the effects of using a specific tool which provided teachers with information about students' performance from which they could derive performance cues: Primary school teachers were asked to use learner response systems (also referred to as clickers) approximately two times a week during their mathematics lessons. With the help of the learner response systems, teachers posed questions during whole-class instructions and received individual student's responses. This intervention improved teachers' relative judgment accuracy, bias, and absolute judgment accuracy: The absolute deviation between teachers' judgments and students' actual performance decreased from 10.43 to 4.14 items, on a test consisting of 25 items. Because the teachers in the Zhu and Urhahne (2018) study judged students' general mathematical skills (as was the case in the abovementioned studies of Thiede et al.), it remains unknown whether this intervention would also improve teachers' judgments of students' performance on specific mathematical tasks. This is important to establish because instructional decisions should be based on judgments of students' performance on the relevant tasks, and not on students' general mathematical performance (Baak et al., 2018; Borghouts et al., 2019; Hollingsworth & Ybarra, 2018; Stiggins & Chappuis, 2006). Moreover, as both the studies by Zhu and Urhahne (2018) and Thiede et al. (2018; 2019) did not provide information on the cues that teachers (presumably) used, it remains unclear whether the (lack of) increase in teachers' judgment accuracy in these studies was caused by (the absence of) improved cue use.

Studies that did measure teachers' cue use showed that giving teachers access to diagnostic performance cues did not necessarily improve their judgment accuracy; rather, the improvement seems to depend on the extent to which teachers simultaneously ignore less diagnostic student cues. For instance, Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018 manipulated the type of cues primary school teachers had available when making judgments of their students' conceptual understanding of decimals. The teachers had access to: (1) only students' names (i.e., student cues), (2) only anonymized student work from which performance cues could be inferred, or (3) both students' names and their work (i.e., both student and performance cues). Teachers could infer students' decimal (mis)conceptions (i.e., performance cues) by analyzing students' work. The teachers thought aloud while making judgments, to measure their cue use. The findings suggest that teachers were most accurate when only performance cues were available (although this was only true for judgments of what students *did not understand,* not for judgments of what students *did understand*). When both student and performance cues were available, teachers did use performance cues, yet were not more accurate and did not focus less on student cues, compared to when only student cues were available. These findings suggest that it is hard to ignore student cues when both student and performance cues are available.

Similar results were found by Van de Pol, Van Gog, & Thiede, 2021. They investigated the relation between secondary school teachers' self-reported cue-utilization and their judgment accuracy of students' text comprehension. Using non-diagnostic student cues, such as effort and intelligence, in addition to diagnostic performance cues (characteristics of diagrams completed by the students, e.g., number of correct relations), appeared to hamper teachers' absolute judgment accuracy. The findings also showed that teachers had difficulties with accurately inferring the performance cues: The teachers' judgment of the number of correct relations in a diagram completed by students deviated substantially from the actual number of correct relations in that diagram. When teachers' judgments of performance cues were inaccurate, their judgments of students' performance were also less accurate. Moreover, the findings indicated that regardless of whether the judgments of the performance cues were (in)accurate the teachers might have had difficulty with translating performance cues (e.g., number of correct causal relations in a diagram) into judgments (i.e., their estimates of the number of causal relations students would correctly recall on the posttest).

In summary, there may be three possible explanations for why providing teachers with performance cues does not always help improve their judgment accuracy: (1) when highly diagnostic performance cues are available, teachers do not merely use these performance cues but they also use less diagnostic student cues, (2) inferring performance cues from student work can be difficult, and (3) it is difficult to translate performance cues into judgments of students' performance. Therefore, the present study provided teachers with student work from which performance cues could easily be derived (as the scores are already provided) and that were highly aligned with the task teachers judged. In (Dutch) educational practice, teachers often have the opportunity to use this type of performance cue. For instance, students in the upper years of (Dutch) primary school work on basic procedural mathematics skills—addition, subtraction, multiplication, and division—on a weekly basis, sometimes as the main learning objective of a task and sometimes as part of a task with another main learning objective (Baak et al., 2018; Borghouts et al., 2019). These procedural mathematical tasks often contain performance information that is not difficult to infer: Tasks can unambiguously be scored in terms of number of problems answered (in)correctly. By means of formative assessment practices, teachers can elicit and use this information to infer cues that can inform their judgments and in turn, their instructional decisions. For example, over the course of two days, students work on a multiplication task. On the first day, they are introduced to the task and on the second day, they get to rehearse the task. Based on students' task scores on the first day, teachers could decide which students need additional instruction on the second day. Or, in the weeks after a monthly assessment, teachers could give additional instruction to students whose assessment performance (i.e., performance cues) indicate that they have not mastered a particular problem type.

We aim to investigate whether prompting teachers to use information from which performance cues can easily be inferred and are well aligned with the tasks that are to-be-judged, will positively affect their judgment accuracy. We also aim to investigate how prompting teachers to use this type of performance cue affects teachers' cue use. This might lead to insights on how to stimulate teachers to use more diagnostic cues, which can in turn foster their judgment accuracy. Moreover, when using performance cues, it is unknown whether it is best for teachers to ignore *all* student cues or whether there are specific student cues that can be of added value. Even if the diagnosticity of student cues is generally lower than that of performance cues, it is possible that some student cues may have added diagnostic value when used in combination with performance cues. If certain student cues do indeed add diagnostic value, then using these could possibly lead to more accurate teacher judgments than merely using performance cues.

1.3. Teachers' confidence in their judgment accuracy

In order to make effective instructional decisions teachers not only need to make accurate judgments of their students' performance (e.g., Van de Pol, Volman, Oort, & Beishuizen, 2014; Klug et al., 2013), but they also need to be *aware* of their judgment (in) accuracy (Gabriele et al., 2016). Teachers show awareness of their judgment (in)accuracy when they feel relatively more confident about more accurate judgments and relatively less confident about less accurate judgments. This is typically measured by asking teachers how confident they are that their judgments are accurate, directly after making a performance judgment (some studies refer to this as second-order judgments: e.g., Dunlosky et al., 2005). When teachers are aware of their (in)accuracy they are likely to make appropriate instructional decisions, either based on judgments that were accurate and in which they have confidence, or by obtaining more information on the judgments that were inaccurate and in which they have less confidence. When teachers are *not aware* of their (in)accuracy, they either feel confident about less accurate judgments, which may lead to inappropriate instructional decisions, or they lack confidence in accurate judgments, which prompts teachers to seek more information (costing time and effort) when in fact this is not necessary (Gabriele et al., 2016). Particularly the combination of less accurate judgments and high confidence can have negative consequences because, in those cases, teachers' instructional decisions are less likely to be tailored to students' needs. To the best of our knowledge, with the exception of Gabriele et al. (2016), teachers' awareness of their (in)accuracy with regard to their students' academic performance has not yet been investigated. Moreover, the measure used by Gabriele et al. (2016) does not allow us to answer to what extent teachers are aware of their judgment (in)accuracy. Furthermore, it remains unknown how the availability of performance cues affects teachers' confidence in, or awareness of, their (in)accuracy.

As teachers' instructional decisions may become more effective when the match between their judgment accuracy and the confidence in their judgment accuracy is higher, increasing teachers' confidence in their judgment accuracy seems desirable in a situation in which judgments also become more accurate. Teachers might feel more confident of the accuracy of the judgments that are based on performance cues than those that are not because teachers might expect that performance cues have high diagnostic value. This is supported by findings of Zhu (2019) who investigated which cues primary school teachers report being most important to base their judgments of students' achievement on. The teachers reported that, when judging students' achievement, relying on the performance of the last test (i.e., a performance cue) is three to five times more important than relying on grades of other subjects or text anxiety (i.e., student cues). However, teachers' perceived effectiveness of using formative assessment to inform their teaching varies and this affects their willingness to carry out formative assessments (for a review, see Yan et al., 2021). It could be that teachers with a less positive attitude towards formative assessment think that basing their instructional decisions on task-specific performance cues is not more effective than basing their decisions on other cues, such as students' general mathematics ability. For these teachers, making performance cues available might not increase their confidence in their judgment accuracy. An increase in confidence would be desirable if teachers' accuracy also increases when performance cues are available compared to when performance cues are not available (which we indeed expect, see section 1.2). However, an average increase in both teachers' judgment accuracy and confidence does not necessarily lead to increased *awareness* of their (in)accuracy. This is because awareness is defined as how well teachers can *distinguish* between their more accurate, and their less accurate, judgments in terms of confidence. It remains unknown how the *relation* between teachers' accuracy and their confidence—as indication of teachers' awareness of their (in)accuracy—changes after performance cues are made available to teachers.

2. The present study

The present study aims to investigate 1) how prompting primary school teachers to use performance cues will affect their cue use when judging their students' performance on mathematical tasks and 2) how use of different (combinations of) cues affects their judgment accuracy, confidence in their judgment accuracy, and awareness of their judgment (in)accuracy. We specifically focus on procedural mathematical tasks (multiplication and division) that form a large part of the mathematics curriculum in the upper years of (Dutch) primary school (e.g., Baak et al., 2018; Borghouts et al., 2019). Teachers made judgments about how many of the six multiplication problems (e.g., 6 × 472) and six division problems (282 : 6) their students answered correctly: These were made under two conditions, one with and one without having access to performance cues. Performance cues consisted of the number of problems students answered correctly on similar tasks (problems with same solution procedure and difficulty, but different numbers) completed one week earlier. Thus, these performance cues were well aligned with the to-be-judged-tasks and did not require interpretation of student work (as the number of problems answered correctly was given).

Our first research question (RQ1) was: How does availability of performance cues, that can easily be derived from student work and are well aligned with the task performance teachers have to judge, affect teachers' judgment accuracy of students' mathematical task performance? We expected that when performance cues are available that teachers would make more accurate judgments, because (1) performance cues are generally diagnostic (Van de Pol, Van Gog, & Thiede, 2021; Thiede et al., 2019), (2) if the provided information does not require interpretation then inaccurate cue judgments might not appear and therefore do not hamper judgment accuracy, and (3) the more the cues are aligned with the to-be-made performance judgments, the easier it might be to translate the cue judgments into judgments of students' performance.

Second, we investigated how availability of performance cues affects teachers' use of student and performance cues, as indicated by the degree to which the cues predict teachers' judgments (RQ2). We would expect that teachers who are provided with information that does not require interpretation and is easy to translate into judgments, are less inclined to additionally use student cues when judging students' performance.

Third, in order to make accurate judgments, teachers should use highly diagnostic cues (Van de Pol, Van Gog, & Thiede, 2021; Thiede et al., 2019), and thus we need to have knowledge of which cues are diagnostic for students' performance on procedural mathematical tasks. Therefore, we investigated to what extent student and performance cues predict primary school students' performance on procedural mathematical tasks, as an indication of cue diagnosticity (RQ3). In line with Van de Pol, Van Gog, & Thiede, 2021 we expected performance cues to be more diagnostic than student cues, although the present study is concerned with a different school subject and age group.

Fourth, assuming the performance cues in the present study are indeed more diagnostic than student cues, and teachers actually use the performance cues, leading to more accurate judgments, then it is relevant to find out whether it is best for teachers to ignore *all* student cues when performance cues are available, or whether there is potential added value in using some student cues in combination with performance cues (RQ4). In other words, we explored if there are student cues that, on top of performance cues, increase the explained variance in students' task performance. We had no specific expectations regarding this question, given a lack of prior studies on cue diagnosticity with regard to primary school procedural mathematical tasks.

Fifth, it has been suggested that teachers not only need to make accurate judgments of students' performance, but that they also need to be *aware* of their (in)accuracy (see section 1.3). We explored how the availability of performance cues affects (a) teachers' confidence in their judgment accuracy, and (b) teachers' awareness of their judgment (in)accuracy (RQ5). Regarding RQ5a, we explored whether and to what degree teachers feel more confident when performance cues are available, compared to when performance cues are not available. The reason being that teachers might expect performance cues to be highly diagnostic. Note that an increase in teachers' *confidence* in their judgment accuracy does not necessarily lead to an increase in *awareness* of their (in)accuracy. Hence, we had no expectations regarding RQ5b about teachers' awareness of their (in)accuracy.

3. Method

3.1. Context of the present study: Dutch (mathematics) education

In the Netherlands, students go to primary school from age 4 to age 12. The average class size at the time of writing is 23 students (Rijksoverheid, n.d.). In the primary schools that participated in the current study, mathematics was taught by the teacher who also teaches most other subjects, as is the case in most Dutch primary schools. From age 6–7 onwards, students receive about 1 h of formal mathematics education, daily. At age 7–8 students start with multiplication and around age 9 with division. From then on, multiplication and division are covered about weekly in the mathematics curriculum until the end of primary school, periodically becoming more complex. Sometimes multiplication and division are the main learning objective of a task, sometimes it is part of a task with another main learning objective (Baak et al., 2018; Borghouts et al., 2019; SLO, 2021).

3.2. Participants

3.2.1. Teachers

Thirty-four teachers, teaching 9–10 year old students (Dutch grade 6), volunteered to participate in this study. Teachers were recruited via (1) an advertisement on social media, (2) contacting schools that participated in a prior study (Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018), and (3) the network of the first author who is also a primary school teacher. One teacher dropped out because they did not feel comfortable with completing the questionnaire about the student characteristics. The other 33 teachers (25 female) taught across 21 different primary schools in the Netherlands, ranged from ages 23–59 years old ($M = 37.71$, $SD = 12.10$), and had one to 39 years of teaching experience ($M = 12.33$, $SD = 10.18$). They taught their classes two to five days a week ($M = 4.24$, $SD = 0.94$). Data collection took place between January and May 2019. The teachers were teaching their students from the beginning of the school year, which, in the Netherlands, roughly spans from the end of August until half July, so they had known their students between 5 and 9 months. Eight of the teachers had also been teaching their class in a previous grade.[1] This study received approval from the ethics review board of the authors' institute. See for the dataset: Oudman, Van de Pol, & Van Gog, 2022c.

3.2.2. Students

Of the 777 students who participated, data from 552 students were included in the analyses of the multiplication task and 553 in the analyses of the division task. Data from 545 students were included in both the analyses for multiplication and division. Fig. 2 displays students' demographics and the number of students that had to be excluded and why.

3.3. Design

This study had a within-subjects design with two conditions: In the student-cue only condition, teachers made judgments and indicated their confidence in the accuracy of these judgments for 10 students, while being provided with the students' names (i.e., making student cues available). In the student + performance cue condition, teachers made judgments for 10 other students while being provided with the students' names and the students'

[1] The eight teachers who taught their class also in a previous grade did not make significantly more accurate judgments than the other 25 teachers, $p > .05$.

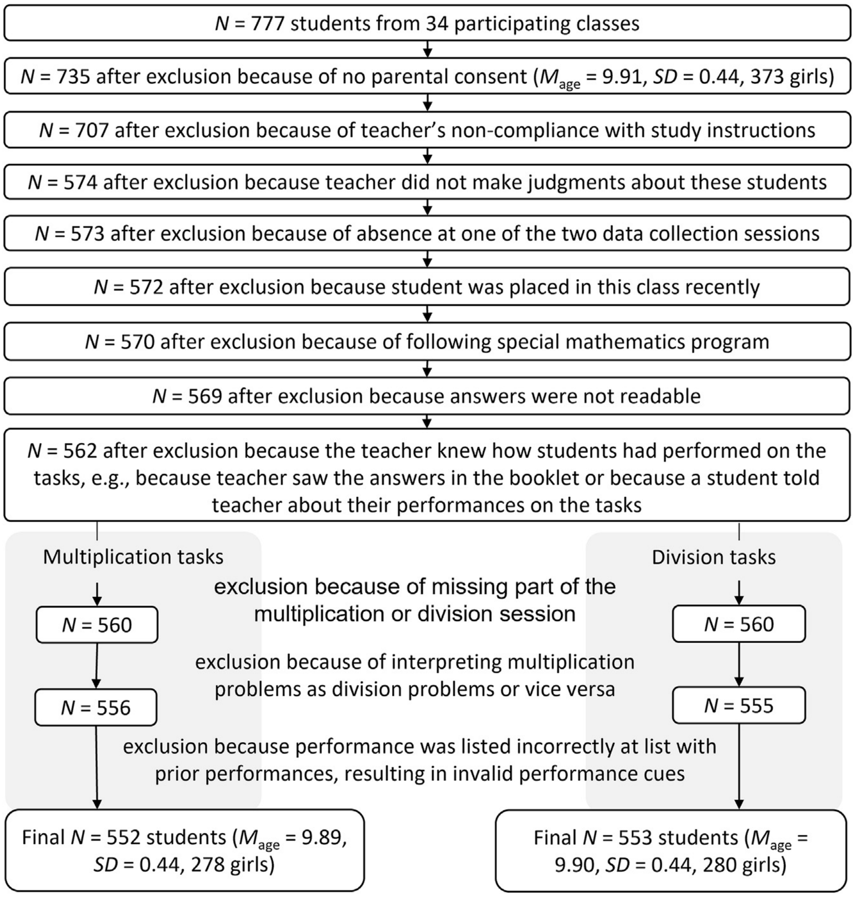

N = 777 students from 34 participating classes
N = 735 after exclusion because of no parental consent (M_age = 9.91, SD = 0.44, 373 girls)
N = 707 after exclusion because of teacher's non-compliance with study instructions
N = 574 after exclusion because teacher did not make judgments about these students
N = 573 after exclusion because of absence at one of the two data collection sessions
N = 572 after exclusion because student was placed in this class recently
N = 570 after exclusion because of following special mathematics program
N = 569 after exclusion because answers were not readable
N = 562 after exclusion because the teacher knew how students had performed on the tasks, e.g., because teacher saw the answers in the booklet or because a student told teacher about their performances on the tasks
Multiplication tasks
Division tasks
exclusion because of missing part of the multiplication or division session
N = 560
N = 560
exclusion because of interpreting multiplication problems as division problems or vice versa
N = 556
N = 555
exclusion because performance was listed incorrectly at list with prior performances, resulting in invalid performance cues
Final N = 552 students (M_age = 9.89, SD = 0.44, 278 girls)
Final N = 553 students (M_age = 9.90, SD = 0.44, 280 girls)

Fig. 2. Flowchart of Reasons for, and Number of, Excluded Students.
Note. Excluded students were removed from the dataset. Multivariate outliers were defined for each analysis separately and are still included in the numbers in this flowchart.

performance on similar tasks completed one week prior (i.e., making performance cues available).

3.4. Materials and measures

3.4.1. Students' performance

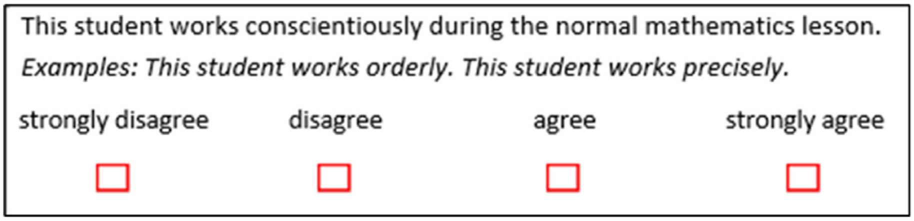
On two days that were exactly one week apart, students made parallel versions (i.e., with isomorphic problems that have the same solution procedure and difficulty, but different numbers) of a multiplication and division task. On both days, students answered six multiplication problems (single-digit multiplicands multiplied by 3-digit multipliers, e.g., 6×472) and six division problems (3-digit dividends divided by single-digit divisors, e.g., 282 : 6). Students received one point for each correctly answered problem, thus, per task performance scores ranged between 0 and 6. Students' task performance (i.e., how many problems they answered correctly) on day 1, were made available to teachers in the student + performance cue condition when making judgments on day 2.

3.4.2. Teachers' judgment accuracy

Per student, teachers were provided with the six multiplication or division items of day 2 and answered the question "How many of these six multiplication/division problems do you think this student answers correctly within 12 min[2]?" on a 7-point scale ranging from 0 to 6. Based on students' task performance one week earlier, ten students per teacher were selected per condition, so that students with comparable scores were equally divided across the two conditions, within each class. This resulted in comparable means and variances of students' prior performance across conditions. When a class consisted of more than 20 students, we optimized the sample regarding the variability in student performance within each class (i.e., we avoided selecting students with similar scores as much as possible). When a class consisted of 20 students or less, teachers made judgments about all their students.

We analyzed teachers' *absolute accuracy*, determined by the absolute difference between the judged and actual performance (regardless of whether it was positive or negative), ranging from 0 to 6, with values closer to zero indicating more accurate judgments (Schraw, 2009; Urhahne & Wijnia, 2021). To allow for comparison of our findings to other studies using a different measure, we also report descriptive statistics of two other measures of teachers' judgment accuracy: Bias and the rank component. *Bias* was computed by subtracting students' actual performance from their judgment and ranges from −6 to 6, with values below zero indicating underestimation and values above zero indicating overestimation. The closer the values are to zero, the lower teachers' overestimation or underestimation of their student's performance is (Schraw, 2009; Urhahne & Wijnia, 2021). The *rank component* indicates how well teachers can accurately rank their students in terms of their performance. This measure was determined by the correlation between teachers' judgments

[2] Students who have automated the procedures would need less than 10 min, based on the opinion of two mathematics experts and three experienced teachers, teaching 9–10 year olds.



This student works conscientiously during the normal mathematics lesson.

Examples: This student works orderly. This student works precisely.

strongly disagree	disagree	agree	strongly agree
□	□	□	□

Fig. 3. Example item of teacher questionnaire about student cues.

and students' actual performance, thereby accounting for the non-independence of observations within classes, by applying multilevel regression models. The rank component ranges from 0 to 1, with values closer to 1 indicating higher accuracy (Schraw, 2009; Urhahne & Wijnia, 2021).

3.4.3. *Confidence judgments*

Directly after teachers made a judgment of a student's performance, they made a confidence judgment by answering the question “How confident are you about the previous judgment?”, on a 6-point Likert scale, ranging from “very unconfident” (1) to “very confident” (6).

3.4.4. *Cue measures*

In the present study, performance cues consisted of students' performance on the multiplication and division task completed one week earlier, ranging from 0 to 6 per task. Student cues were measured using a teacher questionnaire, in which most cues were measured by one item and from the teachers' perspective: We assume that teachers use their own perception of student cues to base their judgments on (cf., studies that also used one-item teacher reports to measure teachers' cue use: Helwig et al., 2001; Kaiser et al., 2013; Zhu & Urhahne, 2020). The teacher-perceived student cues consisted of conscientiousness (during mathematics lessons), effort (during mathematics lessons), sex, interest in mathematics, general mathematics ability, nationality, presence of learning problems, and self-concept (students’ confidence in their mathematical skills). See Fig. 3 for an example item, Appendix Table 4 for a list of student cue measures, and Appendix Table 5 for the descriptive statistics per cue.

The data for this study were collected in the context of a larger project that also included student characteristics as perceived by teachers, for use in other studies. Students' intelligence was included in this questionnaire with the intention of use in the present study. However, we removed students' intelligence from the analyses to prevent multicollinearity because the correlation with mathematics ability was very high (0.82). We additionally performed the analyses with intelligence instead of general mathematics ability as predictor and this led to similar conclusions as those reported here. Parents’ educational level was also included in the questionnaire with the intention of use in the present study, but was removed from the analyses because teachers could not report this variable with certainty for most students.

3.4.5. *Other measures*

After making their judgments teachers indicated whether they had information about students' performance on the multiplication and division task on day 1 or 2, other than stemming from the performance cues they received (e.g., because they accidently saw students’ answers in a booklet or because students told them about their performance on the tasks). The reported students were deleted from the analyses (Fig. 2).

After completing the student cue questionnaire, teachers answered a question about the extent to which tasks that were related to the multiplication and division tasks were part of their curriculum in the past week.[3]

3.5. *Procedure*

Data collection took place on two separate lesson days with exactly one week in between. On both days, the student and teacher session took place at the same time and lasted between 45 min and 1 h. At least two weeks prior to the first day of the data collection, parents were informed and given the opportunity to object to their children's participation or use of their children's data.

3.5.1. *Students*

The student procedure was the same on day 1 and 2, except that isomorphic problems were used. After a short introduction by the experimenter, all students received a booklet and pen and completed the multiplication task, for which they had 12 min. It was emphasized that there was no need to hurry (as mentioned above, students who automated the procedures would need less than 10 min). When students finished the task in less than 12 min, they were instructed to read the (fiction) books they kept in their drawers. After 12 min, the experimenter gave the instruction that students who had not yet finished all problems should stop working. After the tasks, the students answered several questions[4] that were used for the larger project that the present study is part of. Finally, the same procedure was repeated for the division problems.

3.5.2. *Teachers*

The teacher data collected on day 1 were not used in the present study, but in another study of the larger project (Oudman et al., 2022b). During the session on day 2, teachers were provided with a laptop, a list of names of students they had to make judgments about, noise-canceling headphones, and a covered list with the students' performance of one week earlier (i.e., performance cues). They sat in or close to their classroom so that they could not see their students working (as students were working on the tasks of day 2), yet would be able to intervene if an incident would occur in

[3] This question was added because the multiplication and division tasks used in the present study were part of the regular curriculum. When these (or related) tasks would have been covered in the curriculum in the past week, teachers would have had more knowledge about students' multiplication and division skills (i.e., kind of performance cues) than when these (or related) tasks were not part of the past week curriculum. Adding this variable as predictor to the analyses did not change the significance of the results.

[4] Students rated their invested effort, made a monitoring and regulation judgment (indicating their need for intervention regarding the type of problems they just completed, such as additional practice or instruction), and rated their feeling of confidence in the accuracy of the monitoring and regulation judgments they just made. The monitoring and regulation judgments were used in the study of Oudman et al. (2022a).

the classroom that required their attention (which was not the case). For each selected student, teachers completed a 1) multiplication performance judgment, 2) subsequent confidence judgment, 3) division performance judgment, and 4) subsequent confidence judgment. When teachers finished making judgments about the students in the student-cue only condition, they uncovered the list with students' prior performance and made the same four judgments for the students in the student + performance cue condition. After making the judgments, teachers indicated whether they had additional information about their students' performance on day 1 or 2 (see section 3.4.5). Next, they completed the questionnaire about student cues for students in both conditions. We assumed that teachers' perceptions of student characteristics would be influenced less by teachers' judgments of students' performance than teacher judgments by thinking about student characteristics. Finally, teachers answered the question about the past week's curriculum (see section 3.4.5).

3.6. Analyses

All analyses for the multiplication and division task were performed separately because teachers' judgment accuracy could vary along with the subject matter (Kolovou et al., 2021). We performed multilevel regression analyses 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 as each teacher participated with one class (level 2). All fixed effects were tested at the student level. We used the maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality. Full output for all analyses, including intercepts and random effects, are presented in Online Appendix B. To answer RQ1, about the effect of availability of performance cues on teachers' judgment accuracy, teachers' absolute accuracy was regressed on condition (student-cue only condition, vs. student + performance cue condition).

RQ2 and 3 concern teachers' cue use and cue diagnosticity respectively and were analyzed in line with prior studies about how student characteristics relate to teacher judgments (e.g., Furnari et al., 2017; Meissel et al., 2017; Paleczek et al., 2017; Ready & Wright, 2011). RQ2 (cue use) was answered by regressing teacher judgments on the student cues as perceived by teachers. RQ3 (cue diagnosticity) was answered by regressing students' performance on the student cues as perceived by teachers. Table 2 (multiplication) and Table 3 (division) display the explained variance in teachers' judgments (in case of cue use) and in students' performance (in case of cue diagnosticity) by each cue, *including* shared explained variance by other cues we measured. The lens models in Fig. 4 (multiplication) and Fig. 5 (division) display the explained

Table 1
Descriptive statistics of students' performance, teachers' judgment and confidence variables.

		Multiplication		Division	
		Student cues only *n* = 279[a]	Student + performance cues *n* = 272[a]	Student cues only *n* = 277[a]	Student + performance cues *n* = 275[a]
Variable	Range	*M (SD)*			
Student performance	0 to 6	3.76 (2.10)[b]	3.88 (2.11)[b]	2.94 (2.50)[b]	3.07 (2.47)[b]
Judgment	0 to 6	3.69 (1.73)	3.95 (1.91)	2.96 (1.82)	3.04 (2.18)
Absolute accuracy	0 to 6	1.58 (1.44)	1.14 (1.17)	1.77 (1.42)	1.08 (1.11)
Bias	−6 to 6	−0.07 (2.23)[c]	0.08 (1.68)[c]	0.02 (2.31)[c]	−0.03 (1.59)[c]
Rank component[d]	0 to 1	0.41 (0.08)***	0.65 (0.04)***	0.55 (0.05)***	0.78 (0.03)***
Confidence in accuracy	1 to 6	4.31 (0.95)	4.69 (0.93)	4.06 (1.04)	4.66 (0.97)
Accuracy awareness[e]	0 to 1	−0.17 (0.05)***	−0.17 (0.06)**	−0.21 (0.06)***	−0.38 (0.04)***

***$p \le .001$, **$p \le .01$, *$p \le .05$.

[a] For the variables Absolute accuracy, Confidence in accuracy, and Accuracy awareness the samples sizes were somewhat smaller (max. 7% smaller), because multivariate outliers were removed to answer the research questions.

[b] These two values do not differ significantly from each other, $p > .05$.

[c] These values do not differ significantly from zero, $p > .05$.

[d] Correlation between judgment and performance, accounted for the non-independence of observations within classes, followed by Standard Error between brackets.

[e] Correlation between teachers' absolute accuracy and confidence in accuracy, accounted for the non-independence of observations within classes, followed by Standard Error between brackets.

Table 2
Standardized regression coefficients, indicating cue diagnosticity and cue use for the multiplication task, including shared explained variance by the other cues.

Cue	Diagnosticity β (*SE*)[a]			Use β (*SE*)[b]	
	Total sample	Student cues only	Student + performance cues	Student cues only	Student + performance cues
Performance cues					
Multiplication	0.65 (0.03)***	0.63 (0.04)***	0.67 (0.04)***	–	0.93 (0.01)***
Division	0.42 (0.04)***	0.45 (0.06)***	0.43 (0.06)***	–	0.54 (0.04)***
Teacher-perceived student cues					
Conscientiousness	0.42 (0.04)***	0.42 (0.05)***	0.44 (0.05)***	0.43 (0.05)***	0.41 (0.06)***
Effort	0.33 (0.04)***	0.40 (0.05)***	0.31 (0.04)***	0.47 (0.05)***	0.30 (0.05)***
Sex (boy/girl)[c]	0.13 (0.05)**	0.11 (0.07)	0.18 (0.07)**	−0.19 (0.07)**	0.02 (0.08)
Mathematics ability	0.41 (0.04)***	0.38 (0.06)***	0.46 (0.06)***	0.81 (0.03)***	0.65 (0.05)***
Interest	0.36 (0.05)***	0.36 (0.07)***	0.38 (0.06)***	0.63 (0.05)***	0.49 (0.06)***
Learning problems	−0.26 (0.05)***	−0.30 (0.08)***	−0.25 (0.06)***	−0.33 (0.07)***	−0.27 (0.07)***
Self-concept	0.33 (0.05)***	0.33 (0.07)***	0.34 (0.06)***	0.61 (0.04)***	0.44 (0.06)***
Nationality[d]	0.02 (0.04)	−0.02 (0.07)	0.09 (0.05)	−0.14 (0.07)*	−0.02 (0.07)

***$p \le .001$, **$p \le .01$, *$p \le .05$.

[a] Standardized regression coefficients from multilevel regression models in which a single cue predicts students' performance.

[b] Standardized regression coefficients from multilevel regression models in which a single cue predicts teachers' judgments.

[c] This was an open question, but teachers only gave these two answers.

[d] See for coding Appendix Table 5.

Table 3
Standardized regression coefficients, indicating cue diagnosticity and cue use for the division task, including shared explained variance by the other cues.

Cue	Diagnosticity β (*SE*)^{a}			Use β (*SE*)^{b}	
	Total sample	Student cues only	Student + performance cues	Student cues only	Student + performance cues
Performance cues					
Multiplication	0.41 (0.04)***	0.44 (0.06)***	0.46 (0.05)***	–	0.53 (0.05)***
Division	0.77 (0.03)***	0.79 (0.03)***	0.79 (0.03)***	–	0.94 (0.01)***
Teacher-perceived student cues					
Conscientiousness	0.28 (0.03)***	0.32 (0.04)***	0.27 (0.06)***	0.42 (0.06)***	0.30 (0.06)***
Effort	0.34 (0.03)***	0.40 (0.06)***	0.32 (0.05)***	0.47 (0.05)***	0.32 (0.05)***
SexSex (boy/girl)^{c}	−0.04 (0.05)	−0.04 (0.08)	−0.01 (0.06)	−0.19 (0.07)**	−0.04 (0.07)
Mathematics ability	0.53 (0.04)***	0.50 (0.05)***	0.59 (0.04)***	0.79 (0.03)***	0.68 (0.05)***
Interest	0.44 (0.04)***	0.45 (0.05)***	0.45 (0.05)***	0.65 (0.05)***	0.50 (0.05)***
Learning problems	−0.24 (0.04)***	−0.28 (0.06)***	−0.21 (0.05)***	−0.35 (0.07)***	−0.28 (0.06)***
Self-concept	0.47 (0.04)***	0.43 (0.05)***	0.51 (0.05)***	0.63 (0.05)***	0.53 (0.05)***
Nationality^{d}	−0.06 (0.05)	−0.01 (0.06)	−0.05 (0.07)	−0.15 (0.06)*	−0.07 (0.06)

Note. See Table 2 for explanations.

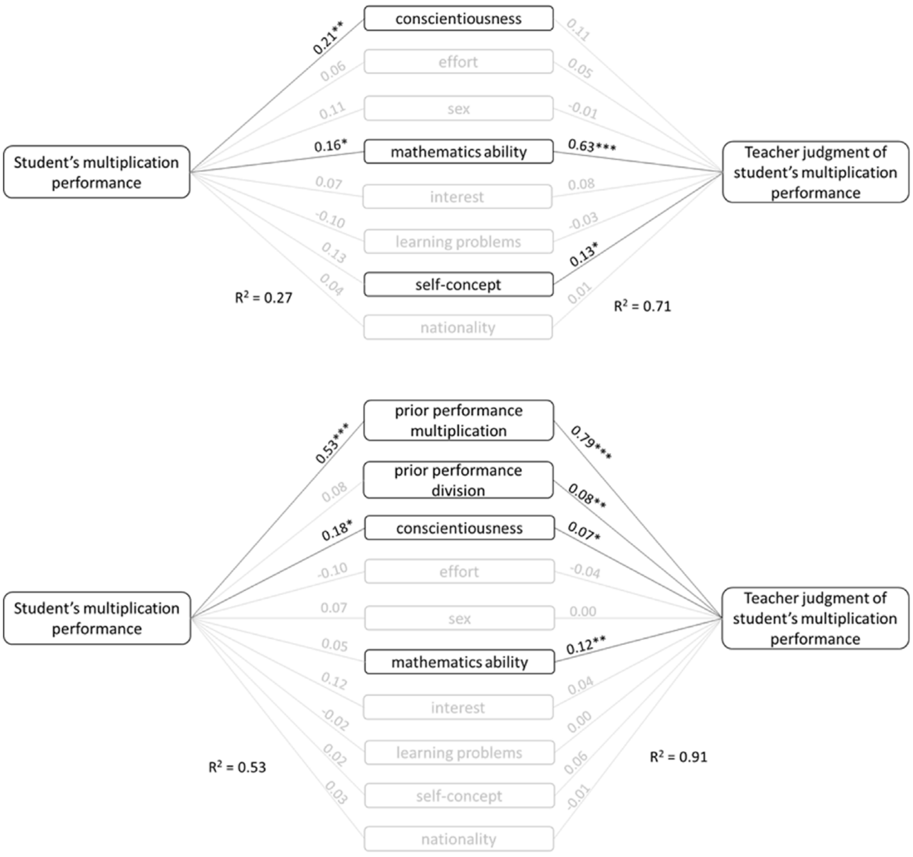


Fig. 4. Lens models of teacher judgments of student's multiplication performance when only teacher-perceived student cues were available (upper model) and when student and performance cues were available (bottom model)
Note. Standardized regression coefficients on the left side of the model represent diagnosticity, standardized regression coefficients on the right side represent teachers' cue use. R^2 is the explained variance in students' performance and teachers' judgments respectively, by all cues in the model. $^{***}p \leq .001$, $^{**}p \leq .01$, $^{*}p \leq .05$.

variance in teachers' judgments and in students' performance by each cue, *excluding* shared explained variance by other cues we measured. Consequently, many coefficients representing diagnosticity and use are significant in Tables 2 and 3, but not significant in the lens models (Figs. 4 and 5), because the explained variance by these cues (almost) entirely overlaps with that of other cues. While we focus our analyses on the lens models, we do feel it is important to present the data in Tables 2 and 3 because these cues can still be diagnostic and teachers may still use these cues (although probably less than cues that are still significant in the lens models), even though the lens models might not give that impression.

To answer RQ4, about the potential added value of student cues for teachers' judgment accuracy, we compared the explained variance in students' performance by the performance cues only (i.e., students' prior performance on similar multiplication and division tasks) with the explained variance in students' performance by both the performance and student cues.

To answer RQ5a, about the effect of availability of performance cues on teachers' confidence in their judgment accuracy, we regressed teachers' confidence on condition (student-cue only condition, vs. student + performance cue condition). To answer RQ5b, about the effect of availability of performance cues on

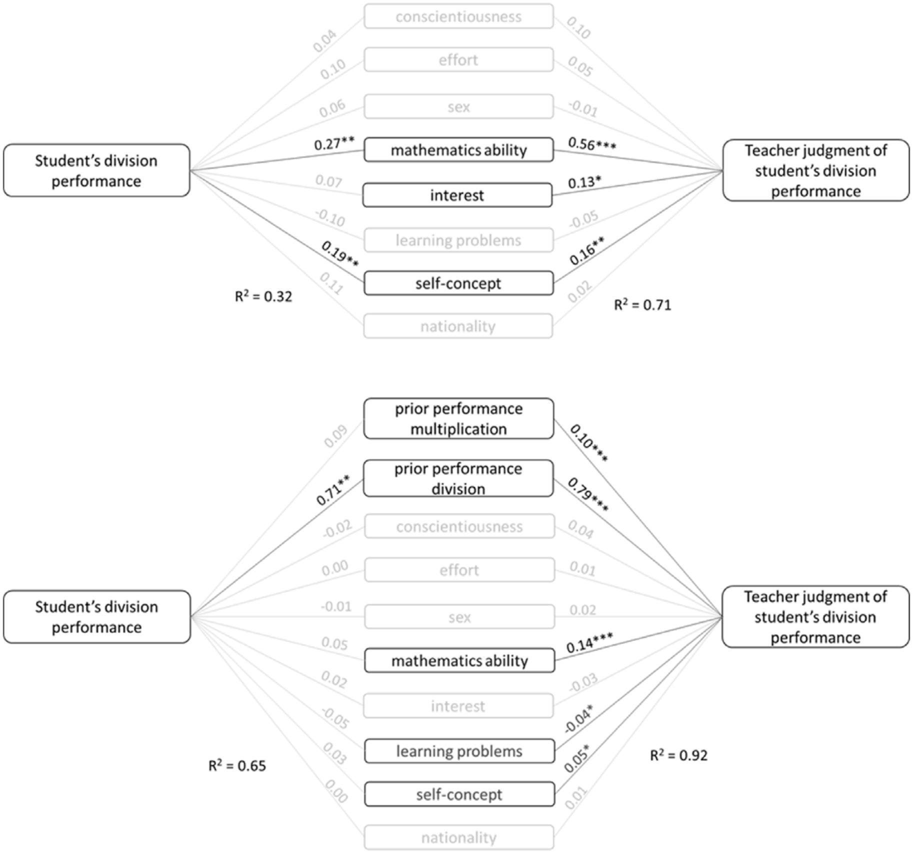


Fig. 5. Lens models of teacher judgments of student's division performance when only teacher-perceived student cues were available (upper model) and when student and performance cues were available (bottom model)
Note. See Fig. 4 for explanation.

teachers' awareness of their (in)accuracy, we analyzed the effect of the interaction term between condition and teachers' accuracy on teachers' confidence in their accuracy. A significant interaction term would mean a stronger relation between teachers' accuracy and their confidence in one of the conditions, compared to the other condition, suggesting a difference in teachers' awareness of their (in)accuracy across conditions.

3.6.1. *Missing cases and outliers*

When data were missing because students or teachers did not complete a question (this applied to 0.2–9.2% per variable), data were deleted list-wise in the analysis. For each multilevel model we analyzed, zero to 39 cases (a maximum of 7.1% of the data) were identified as multivariate outliers. We were mainly interested in the results without outliers to avoid drawing conclusions that were potentially affected by extreme cases in our data. For the sake of transparency we also ran the analyses with the inclusion of the outliers. This only lead to a difference in statistical significance for the analysis of the multiplication task of RQ5b; thus, we reported both effects for this analysis, with and without outliers.

4. Results

4.1. *Descriptive statistics*

Table 1 shows descriptive statistics for the performance, judgment, and confidence variables. In order to enable comparison with prior and future studies reporting on teachers' bias and the rank component, Table 1 also includes these two measures, in addition to teachers' absolute accuracy (the measure used in the analyses). The intraclass correlation coefficient (ICC) for the main variables, reflecting the amount of between-teacher variability compared to the total amount of variability (both between and within teachers), were as follows: For teachers' absolute accuracy 2.5% for multiplication and 2.6% for division; For teachers' confidence in their accuracy 15.9% for multiplication and 13.2% for division. Thus, the largest part of the variability within the main variables resided at the within-teacher (i.e., student) level. For teachers' confidence in their accuracy, differences between teachers were more pronounced than for absolute accuracy.

4.2. *Effect of availability of performance cues on teachers' judgment accuracy (RQ1)*

In line with our hypothesis, teachers' judgments of students' multiplication and division performance were *more accurate* when both *student and performance cues* were provided than when only student cues were provided: See means of absolute accuracy in Table 1. The increase in accuracy was significant for both tasks (Multiplication: $B = -0.43$, $p \leq .001$; Division: $B = -0.69$, $p \leq .001$; Online Appendix B). Teachers' judgment accuracy increased with 0.33 standard deviations for multiplication and with 0.52 standard deviations for division when making judgments with access to both student and performance cues, compared to only student cues. The effect size in terms of f^2 is 0.03, indicating a small effect: 0.02 is the criterion for a small effect, 0.15 for a medium effect, 0.35 for a large effect (Cohen, 1988).

4.3. Effect of availability of performance cues on teachers' cue use (RQ2)

As shown in the lens models (Figs. 4 and 5), when *only student cues were available*, the uniquely explained variance in teachers' judgments by teacher-perceived *students' general mathematics ability* was the largest for both the multiplication and division task and at least three times larger than the variance explained by all other cues we measured. This suggests that when only student cues were available and teachers made judgements of students' performance, that they predominantly used their perceptions of students' general mathematics ability.

When *both student and performance cues were available*, the uniquely explained variance in teachers' judgments by *students' prior performance on a similar task* (i.e., performance cues) was the largest for both multiplication and division, and at least five times larger than the variance explained by all other cues we measured. This suggests that when both student and performance cues were available and teachers made judgments of students' performance, that they predominantly used performance cues (i.e., students' prior performance on the relevant task). It should be noted that teachers did not always copy students' prior performance when making their judgments, as the correlations between the performance cues and teachers' judgments were high but not perfect (0.93 for multiplication and 0.94 for division; Tables 2 and 3). In summary, when performance cues are made available, teachers' cue use may shift from predominantly using their perception of students' general mathematics ability to predominantly using the provided performance cues (Figs. 4 and 5).

In the student-cue only condition, all teacher-perceived student cues accounted for 71% of the variance in teachers' judgments for both multiplication and division; in the student + performance cue condition, all teacher-perceived student cues and performance cues accounted for 91% of the variance in teachers' judgments for multiplication and 92% for division (Figs. 4 and 5). The effect size in terms of f^2 are 2.45 in the student-cue only condition (both tasks) and 10.11 (multiplication) and 11.50 (division) in the student + performance condition, indicating exceptionally large effects (0.35 is the criterion for a large effect; Cohen, 1988). This suggests that the cues we measured give a fair indication of the cues teachers actually used, in both conditions.

4.4. Diagnosticity of student and performance cues (RQ3)

When teachers *only had their perceptions of student cues available*, the uniquely explained variance in students' performance (representing diagnosticity) by teacher perceived *students' general mathematics ability* (for both tasks), *conscientiousness* (for multiplication), and *self-concept* (for division) were significant (Figs. 4 and 5). When *both performance cues and teacher-perceived student cues were available*, only the diagnosticity of *students' prior performance on a similar task* (i.e., performance cues; for both multiplication and division) and the student cue *conscientiousness* (for multiplication) were significant. Hence, students' general mathematics ability has 'unique' diagnostic value when only teacher-perceived student cues are available, but not when performance cues are made available. The performance cues were at least three times more diagnostic than the teacher-perceived student cues. In the student-cue only condition, all teacher-perceived student cues accounted for 27% of the variance in students' performance for multiplication and 32% for division; in the student + performance cue condition, all teacher-perceived student cues and performance cues accounted for 53% of the variance in students' performance for multiplication and 65% for division (Figs. 4 and 5). The effect size in terms of f^2 are 0.37 (multiplication) and 0.47 (division) in the student-cue only condition and 1.13 (multiplication) and 1.86 (division) in the student + performance condition, indicating large effects (Cohen, 1988). This suggests that the cues we measured, together, fairly predict students' performance, especially in the student + performance cue condition.

4.5. Potential added value of student cues for teachers' judgment accuracy (RQ4)

In the student + performance cue condition, the explained variance in students' performance *by students' prior performance* on similar multiplication and division tasks was 48% for multiplication and 64% for division. When *adding the teacher-perceived student cues* as predictors of students' performance, the explained variance increased only 5% (i.e., to 53%) for multiplication and only 1% (i.e., to 65%) for division (Figs. 4 and 5). The effect size in terms of f^2 for this increase are 0.06 for multiplication, a small effect, and 0.01 for division, a trivial effect (Cohen, 1988).

4.6. Effect of availability of performance cues on teachers' confidence in their judgment accuracy (RQ5a)

Teachers felt *significantly more confident* of their judgment accuracy for both the multiplication and division task when they had *access to performance cues*, than when they did not have access to performance cues (Multiplication: $B = 0.37$, $p \leq 0.001$; Division: $B = 0.59$, $p \leq .001$; Online Appendix B). Their confidence on average increased from "somewhat confident" when only student cues were available to "confident" when performance cues were also available (Table 1).

When teachers made judgments with access to both teacher-perceived student cues and performance cues, their confidence increased with 0.39 standard deviations for multiplication and with 0.57 standard deviations for division, compared to when they only had access to teacher-perceived student cues. The effect sizes in terms of f^2 are 0.05 for multiplication and 0.10 for division, indicating small effects (Cohen, 1988), and thus a small increase in confidence, when performance cues are made available.

4.7. Effect of availability of performance cues on teachers' awareness of their (in)accuracy (RQ5b)

The correlation between teachers' judgment accuracy and the confidence in their accuracy, analyzed separately per task and condition, were negative and significant (Table 1). This suggests that teachers were *aware of their (in)accuracy*, as they felt more confident of more accurate judgments than of less accurate judgments and vice versa. This was the case for both tasks and both conditions. When their accuracy increased with 1 point on the 7-point scale, their confidence increased between 0.10 and 0.30 on a 6-point scale (Online Appendix B). The effect sizes in terms of f^2 are 0.03 (multiplication, both conditions; small effect), 0.05 (division, student cues only; small effect), and 0.17 (division, student + performance cues; medium effect; Cohen, 1988).

To test how the availability of performance cues (i.e., condition) affected teachers' accuracy awareness (RQ5b), we analyzed the effect of the interaction term between condition and teachers' accuracy on teachers' confidence in their accuracy. For both tasks, the relation between teachers' accuracy and their confidence was significantly stronger when teacher-perceived student cues and performance cues were available, compared to when only student cues were available (i.e., the interaction term was significant and negative; $B_{multiplication} = -0.10$, $p = .048$[5]; $B_{division} = -0.16$; $p = .003$; Online Appendix B). This suggests that teachers were on average somewhat *more aware of their (in)accuracy* when *both teacher-perceived student cues and performance cues were provided* compared to only student cues.

5. Discussion

The present study investigated (1) how prompting teachers to use performance cues affects primary school teachers' cue use when judging their students' performance on procedural mathematical tasks and (2) how the use of different (combinations of) cues affects their judgment accuracy, confidence in their judgment accuracy, and awareness of their judgment (in)accuracy.

5.1. Teachers' judgment accuracy and cue use (RQ1 and 2)

In line with our hypothesis, we showed that giving teachers access to performance cues—that can easily be derived from student work and are well aligned with the task performance teachers have to judge —in addition to their perceptions of student cues to which they always have access, positively affected teachers' absolute judgment accuracy of their students' mathematical task performance (RQ1). It should be noted that the effect was small, which can perhaps be explained in two ways. First, the teachers were already fairly accurate when they only had access to their perceptions of student cues: their judgments deviated 26% for multiplication and 30% for division from students' actual performance (Table 1). Second, the diagnosticity of the performance cues was high but not close to perfect; we return to this issue in section 5.4.

The finding that teachers' accuracy significantly increased when performance cues were provided differs from the results of two prior studies (Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018; Van de Pol, Van Gog, & Thiede, 2021) which showed that giving teachers access to performance cues, in addition to student cues, did not necessarily lead to more accurate judgments. This difference in findings can presumably be explained by differences in the type of performance cues provided: In the present study, the provided information did not require interpretation (i.e., in prior studies, teachers had to interpret students' answers, whereas here, they were provided with the number of problems the students answered correctly on a similar prior task) and the tasks that the performance cues originated from were highly aligned with the to-be-judged-tasks (i.e., the earlier completed problems were isomorphic to the to-be-judged problems). This might have made it easier for teachers to use the performance cues—and ignore less diagnostic student cues—when making judgments about students' performance. This is supported by our findings regarding the second research question: When both teacher-perceived student cues and performance cues were available, the teachers hardly used student cues in addition to the performance cues.

The findings regarding the second research question also indicate that, when performance cues were not available, teachers seemed to predominantly base their judgments on their perception of students' general mathematics ability. Students' general mathematics ability can be seen as a more global proxy of performance cues and it might not be surprising that, when performance cues on similar tasks are not available, teachers seem to have the tendency to base their judgments on their knowledge of students' general performance in the relevant subject. However, teachers' perceptions of students' mathematics ability were not much more diagnostic than other (low diagnostic) student cues, and substantially less diagnostic than the performance cues. This stresses how important it is that teachers collect task specific performance cues to base their judgments on.

5.2. Cue diagnosticity (RQ3 and 4)

That teachers in the present study seemed to mainly use performance cues when these were available seems a good decision, as the performance cues were much more diagnostic than the teacher-perceived student cues (including the student cue general mathematics ability; RQ3). Although we do not have data on this, it is possible that teachers knew that the performance cues would be more diagnostic than student cues. This would be in line with the finding of Zhu (2019) that teachers reported that, when making judgments of their students' achievement, relying on last test performance (i.e., a performance cue) is much more important than relying on grades of other subjects or test anxiety (i.e., student cues).

Even if student cues are less diagnostic than performance cues, it could have been possible that some teacher-perceived student cues would have had added diagnostic value when used in combination with performance cues. However, our findings with regard to the fourth research question suggest that use of teacher-perceived student cues (or at least the student cues we measured), in addition to using performance cues, would have little if any added value for the accuracy of teacher judgments.

5.3. Teachers' confidence in and awareness of their (in)Accuracy (RQ5)

Finally, we explored teachers' awareness of their judgment (in) accuracy. This can be important for student learning, because when teachers are aware of their (in)accuracy they are more likely to make appropriate instructional decisions, either based on their judgments that were accurate, or by seeking more information about students' performance when their judgments were inaccurate (cf. Gabriele 2016). Teachers show accuracy awareness of their judgment (in)accuracy when they feel relatively more confident about more accurate judgments and relatively less confident about less accurate judgments.

As we expected, teachers felt more confident of their judgment accuracy when performance cues were available, than when only teacher-perceived student cues were available (RQ5a), and rightly so, as their accuracy was also higher. Interestingly, teachers' confidence in their accuracy only increased slightly and on average came close to ‘confident’ but not ‘very confident’. Teachers might have known that the performance cues were more diagnostic than student cues, but also that the diagnosticity of the performance cues was not close to perfect.

An average increase in teachers' *confidence* in their judgment accuracy and in their *accuracy* does not necessarily lead to an increase in teachers' *awareness* of their (in)accuracy. The present study was the first to explore whether or not teachers were aware of their judgment (in)accuracy with regard to students' performance, and found positive results: Teachers indeed showed some awareness of their (in)accuracy, for both the multiplication and division task and when performance cues were or were not available. Moreover, teachers' accuracy awareness was positively affected by the availability of performance cues (RQ5b): Teachers were somewhat more aware of their (in)accuracy when performance and teacher-perceived student cues were available, compared to when only student cues were available. This finding could mean that when teachers use diagnostic performance cues,

[5] When outliers were still included in this sample, the interaction effect was not significant for the multiplication task: $B = -0.05$, $p = .299$.

teachers' instructional decisions about procedural mathematical tasks are not only more accurate, but also more effective. For example, when teachers use performance cues instead of only their perceptions of student cues, they could be more likely to act upon accurate judgments that they are more confident about, and more likely to seek additional information when their judgments are inaccurate (Gabriele et al., 2016). Of course, this should be confirmed by future research.

5.4. Limitations and future research

One limitation of this study is that we did not directly measure teachers' cue use, but did so by means of correlations between teachers' judgments and measures of (teacher-perceived) cue values. The explained variance in teachers' judgments by the cues we measured was high (above 0.70 when only student cues were available and above 0.90 when student and performance cues were available). This suggests that the cues we measured give a fair indication of the cues teachers actually used. As we did not directly measure cue use, it is possible that teachers did not actually use the cues that are indicated by our findings, but instead used cues that are related (both conceptually and correlational) to the cues we measured. For instance, as mentioned in section 3.4.4, teachers' perceptions of their students' intelligence and mathematics ability are highly related, so concluding which of the two variables they actually use is not possible via correlational research. Future research investigating the effect of interventions on teachers' cue use and their accuracy can include more direct measures of teachers' cue use via think aloud protocols or questionnaires about which cues they used. As teachers' perceptions of the student cues can differ from the actual cue values, for instance as measured by student questionnaires (Van de Pol, Van Gog, & Thiede, 2021), future studies could also investigate whether teachers' judgment accuracy would improve when teachers would more accurately judge the student cues. Additionally, these future studies could incorporate a check on whether the order of measuring judgments and cue use matters: We measured teachers' perceptions of cues after teachers made the judgments, but would the findings change when it was measured in the opposite order?

A question raised by our findings is how the diagnosticity of performance cues can be further improved. In the current study, the diagnosticity of the performance cues ranged between 0.63 and 0.79 (including shared explained variance by other cues), which is fairly high, but not close to perfect. If the diagnosticity of performance cues can be increased, the accuracy of teachers' judgments might also improve further. It would be interesting in future research to look for factors that influence the diagnosticity of performance cues, such as the type of task or the time between the task on which performance cues were collected and the to-be-judged tasks (i.e., a week in the present study). Future research could also attempt to measure to what extent cue diagnosticity differs across students and whether this can be explained by specific student characteristics. For instance, in theory, cues like effort or interest might be diagnostic when they are high or low in a student, but less diagnostic when they are medium/moderate.

Another question we cannot answer based on our data, is to what extent teachers were aware of the diagnosticity of the cues they used. As discussed earlier, our findings that teachers mostly used performance cues when these were available and hardly used student cues, might suggest that they were aware of the higher diagnosticity of performance cues. However, we do not know this for certain. Teachers' beliefs about and awareness of cue diagnosticity might influence their judgment accuracy and confidence in their accuracy. It would be valuable in future research to interview teachers about their thoughts on cue diagnosticity and how this affects their cue use, judgments, and confidence in their accuracy.

Finally, an important question is to what extent our findings would generalize. First, our findings apply to performance cues that can easily be inferred from student work and are well aligned with the to-be-judged-tasks. Of course, even when teaching procedural tasks as the ones in the present study, teachers can also engage in interpreting students' strategy use to inform their instructional decisions. However, teachers often lack time (Schildkamp et al., 2017) and it is very time-consuming to analyze students' strategy use each time a student does not master a task. Using students' task scores to make quick decisions on which students do and do not master a task is an efficient method that can be alternated with more in-depth analyses of students' strategy use. While teachers commonly have access to 'quick' cues, these cues are not always available, for example when starting with a task that is new to the students. Future research could investigate for different type of tasks that are used in educational practice, and that vary to the degree to which they require interpretation, (1) whether they contain diagnostic performance cues, (2) whether teachers are able to use these performance cues and ignore less diagnostic cues, and (3) how this affects teachers' judgment accuracy and accuracy awareness. Future studies should also investigate further whether teachers can be trained to use performance cues that are more difficult to interpret (cf. those used in prior research; Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018, Van de Pol, Van Gog, & Thiede, 2021) and at the same time ignore student cues. For instance by asking teachers to make judgments based on vignettes after which they receive feedback about their use of student and performance cues. Lastly, it is an open question whether our findings would replicate in a larger sample of teachers and schools and whether they generalize to other age groups within and beyond primary school.

6. Conclusions and practical implications

Formative assessment does not necessarily lead to accurate teacher judgments (Thiede et al., 2018, 2019) and this might be caused by the type of performance cues that are used by the teachers. Our findings indicate that the use of diagnostic performance cues that can easily be inferred from student work and are highly aligned with the task performance teachers have to judge (e.g., students' prior performance on similar tasks) improves teachers' judgment accuracy. It might seem obvious that teachers are able to use this type of performance cue and that this increases their judgment accuracy, but this was not a given when looking at prior research (Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018, Van de Pol, Van Gog, & Thiede, 2021). Moreover, it was unknown whether teachers would also *ignore* less diagnostic student cues (as they continued to use those in prior studies), which indeed seemed to be the case. Furthermore, the present study showed that teachers are already somewhat aware of their (in)accuracy, in that they feel relatively more confident about more accurate judgments and relatively less confident about less accurate judgments, and that teachers' confidence in and awareness of their (in)accuracy can be positively affected by using performance cues, all of which is important for adaptive teaching.

Our findings suggest that encouraging teachers to use short formative assessment practices, which are relatively easy to implement, might help them to more accurately evaluate their students' performance and needs. For instance, ending a mathematics lesson by asking students to solve a problem that represents the main learning objective of that lesson, and show their answers on mini-whiteboards they hold up (Wiliam, 2011), might provide teachers with the kind of easy to interpret performance cues that were also used in the present study, and help them to get a quick overview of which students might need additional interventions. When working with online learning systems, teachers can be encouraged to base their instructional decisions on the information on students' performance as shown in the teacher dashboards (which teachers do not necessarily consult; Molenaar & Knoop-van Campen, 2017). To help design and implement the most effective interventions, future research should test which tasks within the current (mathematics) curriculum do and do not provide performance cues that increase teachers' accuracy. This knowledge can then be included in teacher professional development programs aimed at improving formative assessment practices and teachers' judgment accuracy.

Data availability

The data used for the analyses are openly available in online depository (see Oudman et al., 2022c).

Acknowledgements

This research was supported by funding from the Dutch Ministry of Education, Culture and Science, (grant number OCW/PromoDoc/1065001). The authors would like to thank the participating teachers and students for their time and effort.

Appendix

Table 4
Explanation of Student Cue Measures

Cue	Question (translated from Dutch)	Answer options (translated from Dutch)	Source on which question is based
Conscientiousness	This student works conscientiousness during the normal mathematics lesson. *Examples: This student works orderly. This student works precisely.*	Strongly disagree to Strongly agree[a]	Big Five conscientiousness scale (Goldberg, 1992)
Effort	This student shows effort during normal mathematics lessons. *Examples: this student works hard; this student pays attention.*	Strongly disagree to Strongly agree[a]	Cf. Helwig et al. (2011)
General mathematics ability	This student is in general strong in mathematics	Strongly below average/Below average/ Average/Above average/Strongly above average	Cf. Helwig et al. (2011)
Interest	This student is generally interested in mathematics.	Strongly disagree to Strongly agree[a]	Cf. Karing (2009)
Learning problems	Does this student have learning problems (no diagnosis needed)?	No learning problems/Dyslexia/ Dyscalculia/ADHD/ADD/Autism/Language delay/Other, namely …,	Cf. Van de Pol, Van Gog, & Thiede, 2021
Nationality	What is the country of Birth of this student/the mother of this student/the father of this student?	The Netherlands/Another Country, namely: …	Cf. Driessen et al. (2015), Van de Pol, Van Gog, & Thiede, 2021
Self-concept	This student generally feels confident about their mathematical skills. *Examples: this student is convinced that he/she performs well on mathematical tasks and tests; this student knows that he/she can master the mathematics skills that he/she needs to learn.*	Strongly disagree to Strongly agree[a]	Perceived self-efficacy scale (Marsh et al., 2006)
Sex	Before the start of the experiment teachers were asked to provide the experimenter with a list of student names and their sex.	Open question.	

[a] Scale: Strongly disagree, disagree, agree, strongly agree.

Table 5
Descriptive Statistics of the Cue Values for the Multiplication and Division Task

Cue	Range	Multiplication			Division		
		Total sample	Student cues only	Student + performance cues	Total sample	Student cues only	Student + performance cues
Performance cues		*M (SD)*					
Multiplication	0 to 6	3.36 (2.14)	3.27 (2.14)	3.45 (2.13)	3.35 (2.13)	3.26 (2.15)	3.46 (2.11)
Division	0 to 6	2.65 (2.38)	2.57 (2.37)	2.74 (2.39)	2.65 (2.37)	2.57 (2.36)	2.73 (2.39)
Student cues							
Conscientiousness	1 to 4	2.90 (0.79)	2.85 (0.79)	2.95 (0.80)	2.90 (0.80)	2.84 (0.79)	2.95 (0.81)
Effort	1 to 4	3.18 (0.67)	3.16 (0.67)	3.20 (0.66)	3.18 (0.67)	3.16 (0.67)	3.19 (0.67)
Sex	boy/girl[a]	49.6/50.4%	50.0/50.0%	49.0/51.0%	49.4/50.6%	50.8/49.2%	47.8/52.2%
Mathematics ability	1 to 5	3.27 (1.10)	3.21 (1.12)	3.32 (1.08)	3.26 (1.09)	3.21 (1.12)	3.30 (1.06)
Interest	1 to 4	3.00 (0.71)	2.96 (0.69)	3.03 (0.73)	2.99 (0.71)	2.96 (0.69)	3.02 (0.73)
Learning problems	no/yes	75.5/24.5%	74.0/26.0%	77.0/23.0%	75.0/25.0%	73.4/26.6%	76.5/23.5%
Self-concept	1 to 4	2.83 (0.79)	2.78 (0.80)	2.89 (0.77)	2.82 (0.79)	2.77 (0.80)	2.88 (0.78)
Nationality[b]	1 to 5	0.22 (0.73)	0.24 (0.77)	0.21 (0.70)	0.22 (0.73)	0.24 (0.77)	0.21 (0.69)

[a] This was an open question, but teachers only gave these two answers.
[b] Coded as follows: (0) student, mother and father born in Western country, (1) student and mother or father born in W, (2) student born in W, mother and father not, (3) student not born in W, mother and father born in NL, (4) student, mother and father not born in W (it did not occur that student was not born in W, mother or father born in W).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tate.2022.103982.

References

Baak, G., Boon, B., Bosma, G., Van der Brink, M., Cornelissen, F., Druif, D., … Wynia, F. (2018). *Getal & ruimte junior handleiding groep 6*. Noordhoff.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Borghouts, C., Buter, A., & Gool, A. (2019). *Pluspunt 4 handleiding groep 6*. Malmberg.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193–217. https://doi.org/10.1037/h0047470

Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability, 21*(1), 47–65. https://doi.org/10.1007/s11092-008-9063-x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

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

Driessen, G., Elshof, D., Mulder, L., & Roeleveld, J. (2015). *Cohortonderzoek cool 5-18: Technisch rapport basisonderwijs, derde meting 2013/14. ITS*. Amsterdam: Radboud Universiteit Nijmegen/Kohnstamm Instituut.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*(4), 551–565. https://doi.org/10.1016/j.jml.2005.01.011

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

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/10403590.4.1.26

Gortazar, L., de Lafuente, D. M., & Vega-Bayo, A. (2022). Comparing teacher and external assessments: Are boys, immigrants, and poorer students undergraded? *Teaching and Teacher Education, 115*, Article 103725. https://doi.org/10.2139/ssrn.3779872

Helwig, R., Anderson, L., & Tindal, G. (2001). Influence of elementary student sex on teachers' perceptions of mathematics achievement. *The Journal of Educational Research, 95*(2), 93–102. https://doi.org/10.1080/00220670109596577

Hollingsworth, J. R., & Ybarra, S. E. (2018). *Explicit direct instruction (EDI): The power of the well-crafted, well-taught lesson*. SAGE Publications.

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

Johnston, O., Wildy, H., & Shand, J. (2019). A decade of teacher expectations research 2008-2018: Historical foundations, new developments, and future pathways. *Australian Journal of Education, 63*(1), 44–73. https://doi.org/10.1177/0004944118824420

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

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Padagogische Psychologie, 23*(34), 197–209. https://doi.org/10.1024/1010-0652.23.34.197

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

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A. K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education, 100*, Article 103298. https://doi.org/10.1016/j.tate.2021.103298

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

Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*(4), 311–360. https://doi.org/10.1207/s15327574ijt0604_1

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

Molenaar, I., & Knoop-van Campen, C. (2017). Teacher dashboards in practice: Usage and impact. In E. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data driven approaches in digital education: 12th European conference on technology enhanced learning* (pp. 125–138). Springer. https://doi.org/10.1007/978-3-319-66610-5_10.

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

Oudman, S., Van de Pol, J., Bakker, A., Moerbeek, M., & Van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76*, 214–226. https://doi.org/10.1016/j.tate.2018.02.007.

Oudman, S., van de Pol, J., & Van Gog, T. (2022a). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning, 17*(1), 213–239. https://doi-org.proxy.library.uu.nl/10.1007/s11409-021-09281-9.

Oudman, S., van de Pol, J., van Loon, M., & Van Gog, T. (2022b). Primary School Teachers' Judgments of their Students' Monitoring and Regulation Skills [Manuscript submitted for publication]. *Department of Education*. Utrecht University. https://doi-org.proxy.library.uu.nl/10.1007/s11409-021-09281-9.

Oudman, S., Van de Pol, J., & Van Gog, T. (2022c). *Effects of cue availability on primary school teachers' accuracy and confidence in their judgments of students' mathematics performance*. Location. https://osf.io/zv8en/?view_only=8cdea0c6a8314eac8cf6eb889307e628.

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

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., … Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research, 88*(2), 205–242. https://doi.org/10.3102/0034654317743198

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

Rijksoverheid, (n.d.. Hoe is de groep van mijn kind op de basisschool samengesteld?. https://www.rijksoverheid.nl/onderwerpen/basisonderwijs/vraag-en-antwoord/hoe-zijn-de-groepen-in-het-basisonderwijs-bo-samengesteld.

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement, 28*(2), 242–258. https://doi.org/10.1080/09243453.2016.1256901

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Seegers, G., & Boekaerts, M. (1993). Task motivation and mathematics achievement in actual task situations. *Learning and Instruction, 3*(2), 133–150. https://doi.org/10.1016/0959-4752(93)90012-O

SLO. (2021). TULE rekenen/wiskunde. https://www.slo.nl/thema/meer/tule/rekenen-wiskunde/.

Stiggins, R. J., & Chappuis, J. (2006). What a difference a word makes: Assessment for learning rather than assessment of learning helps students succeed. *Journal of Staff Development, 27*, 10–14. http://downloads.pearsonassessments.com/ati/downloads/What-a-difference-a-word-makes.pdf.

Thiede, K. W., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., & Osguthorpe, R. D. (2018). Improving the accuracy of teachers' judgments of student learning. *Teaching and Teacher Education, 76*, 106–115. https://doi.org/10.1016/j.tate.2018.08.004

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., Oswalt, S., Brendefur, J. L., Carney, M. B., & Osguthorpe, R. D. (2019). Teachers' judgments of student learning of mathematics. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 678–695). Cambridge University Press. https://doi.org/10.1017/9781108235631.027.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., … Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2–3), 119–145. https://doi-org.proxy.library.uu.nl/10.1177/016235320302700203.

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, Article 100374. https://doi.org/10.1016/j.edurev.2020.100374

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.proxy.library.uu.nl/10.1007/s10648-010-9127-6.

Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2014). Teacher scaffolding in small-group work: An intervention study. *Journal of the Learning Sciences, 23*(4), 600–650. https://doi-org.proxy.library.uu.nl/10.1080/10508406.2013.805300.

Van de Pol, J., Van Gog, T., & Thiede, K. (2021). The relationship between teachers' cue-utilization and their monitoring accuracy of students' text comprehension. *Teaching and Teacher Education, 107*, Article e103482. https://doi.org/10.1016/j.tate.2021.103482

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice, 22*(3), 324–343. https://doi.org/10.1080/0969594X.2014.999024

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice, 28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042

Zhu, C. (2019). *Understanding the formation and improving the accuracy of teacher judgment [Doctoral dissertation, Universität Passau]*. https://opus4.kobv.de/opus4-uni-passau/frontdoor/index/index/docId/738.

Zhu, C., & Urhahne, D. (2018). The use of learner response systems in the classroom enhances teachers' judgment accuracy. *Learning and Instruction, 58*, 255–262. https://doi.org/10.1016/j.learninstruc.2018.07.011

Zhu, C., & Urhahne, D. (2020). Temporal stability of teachers' judgment accuracy of students' motivation, emotion, and achievement. *European Journal of Psychology of Education, 36*, 319–337. https://doi.org/10.1007/s10212-020-00480-7