Metacognition and Learning (2023) 18:295–312
https://doi.org/10.1007/s11409-023-09333-2

Knowledge about others' knowledge: how accurately do teachers estimate their students' test scores?

Mehmet Akif Güzel[1] · **Tahsin Oğuz Başokçu**[2]

Received: 10 May 2022 / Accepted: 11 January 2023 / Published online: 21 January 2023

Abstract
Besides learners' awareness of their knowledge, a growing number of studies also emphasise the importance of teachers' awareness of how well their students perform to adjust their teaching strategies accordingly. Therefore, proposing a multi-layered metacognitive regulatory model in teaching first, we investigated whether estimation type, item difficulty, and class performance affect teachers' judgment accuracies ([JAs], i.e., score estimations). Teachers (N=38) of 86 classes made item-by-item and overall estimations of their classes' test scores (N=2608 sixth-graders native in Turkish) at a PISA-equivalent mathematics test that was developed in the earliest phase of the current long-term research project. The results showed that teachers' item-by-item estimations were below their classes' actual performance, unlike their overall estimations. Teachers of low-performance classes were less accurate than those of high-performance classes. These teachers also showed the clearest underestimation for the easy questions, whereas teachers of high-performance classes overestimated their classes' scores for the difficult questions. This dissociation implied that the teachers 'must have' primarily used their perceptions about their classes (e.g., classes' existing performance) as a mnemonic judgment cue rather than item difficulty as an external cue when making their score estimations. The implications of the results were discussed in the light of existing literature and suggestions for prospective research were given.

Studies of metacognition principally investigate people's higher-order processes, such as knowledge about one's own cognition (i.e., knowledge), awareness of its level, or regulation of that knowledge (Flavell, 1979; Nelson, 1990). Though metacognition pertains to the meta-level processes of one's own cognition and regulation of it via monitoring and

✉ Mehmet Akif Güzel
akif.guzel@agu.edu.tr

Tahsin Oğuz Başokçu
tahsin.oguz.basokcu@ege.edu.tr

[1] Department of Psychology, Abdullah Gül University, 38040 Kayseri, Türkiye

[2] Department of Educational Sciences, Ege University, İzmir, Türkiye

controlling processes (e.g., Nelson & Narens 1994), it seems also reasonable to ask how accurate we are at knowing and even regulating *someone else's* knowledge. There may be various answers to the question arising, particularly the one regarding the regulation, from different fields or contexts. For instance, a person may initially assume and later continue monitoring the amount of knowledge a friend has on a topic during a conversation, then they may wish to control the amount of the information exchanged by arranging the details. Also, even any company wants to know the instant perceptions of their potential or existing customers about their brand(s) (i.e., monitoring) and wishes to take some initiative to boost their sales by enhancing positive attitudes towards their brand(s) (i.e., control). Amongst many possible instances, however, one of the most convenient groups to study the accuracy of one's knowledge about others' knowledge seems to be 'teachers' (see, e.g., Nelson et al., 1998; Tullis, 2018).

An improving line of research, namely teachers' judgment accuracies (JAs), has revealed that teachers' better awareness of their teaching effectiveness facilitates their students' academic achievement (Brendefur et al., 2016; Thiede et al., 2019). Teachers' JAs obtained for their students' cognitive abilities, academic achievement, and divergent thinking have extensively been investigated (see, e.g., Gralewski & Karwowski 2019; Machts et al., 2016; Südkamp et al., 2012 for meta-analytic studies). For instance, previous works have found that students' socioeconomic status and gender (Kaiser et al., 2017; Ready et al., 2011), their engagement in class (Kaiser et al., 2013), attention, impulsivity, hyperactivity (Jenkins & Demaray, 2016), past performance of the class (Hecht & Greenfield, 2002; Martínez et al., 2009), and even personality similarity between teachers and students (Rausch et al., 2016) play a critical role in teachers' judgment accuracies. Specifically, however, the most common finding in teachers' JAs has been the following: teachers mostly tend to overestimate their students' performance in standardised tests, implying that they don't perceive the difficulty of tasks or test items in the same way that their students do (Urhahne & Wijnia, 2021).

Despite the existence of factors that may account for teacher JAs or reveal the correlations between these miscellaneous factors and teachers' JAs, previous work still seems to fall behind in explaining how accurately teachers estimate their students' actual test performance from an explicitly metacognitive perspective. Moreover, it has been a while since a fruitful working model has arrived at the research venue of teachers' JAs that could guide the related studies. This model, entitled 'the metacognitive monitoring and control of students and teachers' (Fig. 26.1, p. 682) and is based on Nelson and Narens' (1994) metacognitive framework, has been proposed by Thiede et al., (2019) recently. Though it is beyond the scope and aim of the present research to exhaustively redefine it, we first revisit this working model by suggesting some additions to it and then outline the present study in the proceeding sections.

A multi-layered Metacognitive Regulatory Model in Teaching

As Nelson & Narens (1994) suggested in their metacognitive framework, a flow of information is presumed to occur between the meta- and object-levels where the meta-level (metacognition) monitors and so is aware of the knowledge accumulated at the object-level (cognition). The meta-level consequently is presumed to control the object-level via several behaviours that are guided by a predetermined goal to be achieved (Mazzoni et al.,

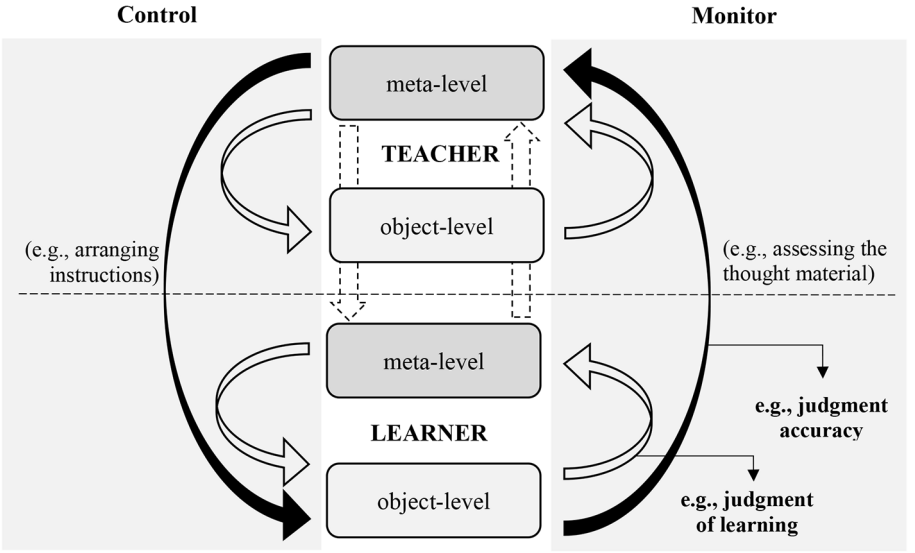

Figure 1 A multi-layered metacognitive regulation model in teaching
Note: This model is a readdressed version of Thiede et al., (2019), which is based on Nelson and Narens' Metacognitive Framework (1990; 1994). Arrows on the left show the control processes and those on the right present the monitoring processes. Dashed arrows show the bidirectional influence between teachers' and learners' meta-levels, regarding to metacognitive goals of teaching and learning objectives, respectively

1990; Son & Metcalfe, 2000). In the working model of Thiede and his colleagues (2019), which is also based on Nelson and Naren's conceptualisation, there exist two regulatory processes: students' metacognitive regulation of their own knowledge and teachers' regulation of their students' knowledge. However, in the working model shown in Fig. 1, we propose to include three separate metacognitive regulation processes instead of two. These processes are as follows: teacher's metacognitive regulatory process of their own object-levels -which is not explicitly present in Thiede et al.'s model-, their students' regulation of their own knowledge, and teachers' regulation of their student's knowledge (bold arrows in Fig. 1).[1] Since previous research has shown that teachers' better usage of metacognitive, educational, and instructional strategies, as well as enhancement of their knowledge regarding how to be effective teachers and the subjects they teach, affect students' achievement (e.g., Darling-Hammond 2000; Harris & Sass, 2011; Nye et al., 2004), we believe that incorporating teachers' regulation of their own knowledge into the model would further complete what Thiede and his colleagues portrayed in their model.

[1] The model can also involve -at least at a theoretical level- a fourth possible regulation, which is regarding the learners' metacognitive regulation of their teachers and is also not mentioned in Thiede and his colleagues' model (2019). However, we did not explicitly include this theoretically possible regulation loop that would emerge from the learner's meta-level to the teacher's object-object mainly because students do not have a clearly defined supervisory role in the educational settings just like their teachers. Nevertheless, learners may influence the teachers' meta-levels (i.e., their metacognitive goals regarding their teaching objectives) by, for instance, course evaluations and/or various feedbacks; therefore, we show this as a bidirectional influence in the model rather than a fourth regulatory loop (see dashed arrows between learner's and teacher's meta-levels in Fig. 1).

As shown in Fig. 1, teachers monitor learners' knowledge via various assessment tools, such as exams, homework, projects, etc., and control the students' knowledge levels through instructions, such as by modifying their instruction methods. Students are also presumed to monitor and control their knowledge based on their learning objectives. Additionally, teachers' monitoring ability of their students' knowledge can be indexed by several metacognitive judgments, such as how accurately the teachers judge the level of their students' performance or knowledge, e.g., their classes' exam scores. Specifically, however, the current multi-layered working model adjoins JAs as a direct product of the monitoring process of teachers (the broken arrow on the right-hand side in Fig. 1) thereby conveying the teaching effectiveness as well. On the other hand, teacher JA and its later effects on teaching effectiveness (e.g., instruction) and subsequently on students' academic achievement are shown separately in Thiede and his colleagues' model (2019). This might have been needed by them just to lay out the relation between teachers' monitoring and their control processes in separate illustrations only (e.g., teachers' judgments and instructions, respectively) although 'judgment accuracy' and its subsequent effect, namely, 'effectiveness of teaching', already refer to these very same processes (teachers' judgments and instructions, respectively). We, therefore, suggest that it is a clearer conceptualisation of the metacognitive regulation processes in teaching provided that a multi-layered model involves the following regulations 'together': (a) teachers' *own* metacognitive regulatory processes -unlike Thiede and his colleagues' model does not clearly involve-; (b) students' own metacognitive regulatory processes; and (c) teachers' monitoring and control processes of their classes that are operated on their students individually and/or on their classes as a whole (see, e.g., Carr 2010; Duffy et al., 2009; Zimmerman, 2008 for the regulatory processes of separate parties involved in teaching, i.e., learners & teachers). Albeit this revisited version is on the teachers' judgment accuracies just like Thiede and his colleagues described, it is also possible to see the learners' particular metacognitive judgments in the model. For instance, students' immediate or delayed score estimations of their test performance may provide evidence of how accurately they monitor their knowledge (Basokcu & Guzel, 2022; Prohaska, 1994; Svanum & Bigatti, 2006) just like those measured in the studies of judgments of learning ([JOL]; e.g., Nelson & Dunlosky 1991; Dunlosky & Metcalfe, 2009a; Rhodes, 2016).

The revisited model may naturally not be exhaustive of all plausible factors involved in any educational context and it cannot be considered as the ultimate one. Nevertheless, we believe that it captures the metacognitive process of two distinct parties, teachers and learners separately, along with teachers' metacognitive regulation of their teaching activity through monitoring and controlling students' object-levels. Again, as was suggested by Nelson and Narens' metacognitive framework (1994), the meta-level is presumed to operate both monitoring and control processes, though these processes are implemented separately. This multi-layered model also assumes that a perfect match between teachers' score estimations and their students' actual scores does *not* necessarily ensure a perfect instructional activity (i.e., control) although it seems highly likely to observe both or neither of them together just like what is expected in Thiede and his colleagues' model (2019). Teaching effectiveness, in short, seems to depend on how well teachers' metacognitive strategies are clearly set and loyally maintained, which should undeniably be towards enhancing their students' academic achievements (see, e.g., Brendefur et al., 2016; also see, e.g., Gabriele et al., 2016 on how maintaining a teaching goal affects learners' achievement).

Overview of the study

We investigated teachers' judgment accuracies and tested whether item difficulty, class performance, and estimation type account for these judgments. As we are aware, none of the previous works on teachers' JAs has tested these three factors together. Additionally, we herein defined this accuracy similar to the absolute accuracies and so measured them as the divergence between teachers' score estimations and their classes' actual scores that are obtained from a given test (see, e.g., Urhahne & Wijnia 2021 for further details on how absolute and relative JAs are measured), though JAs have been operationalised and measured somewhat variably in different kinds of literature, such as in cognitive psychology and education (Dunlosky & Metcalfe, 2009b; Nelson & Dunlosky, 1991; Thiede et al., 2015). Unlike the previous studies that measured teachers' absolute accuracies, the current study did not ask teachers to estimate the scores of a particular group of students in their classes or each student individually, but it asked teachers to estimate their whole class(es)' performance (i.e., expected 'average exam score of the class').

Teachers who participated in the study were first asked to browse the test booklet thoroughly, then they were asked to estimate the number of students in their classes who could answer the questions correctly. They made these estimations for each item separately (item-by-item estimation) and they predicted the mean score that their class would obtain from the test as well (overall estimation). We expected that teachers would base their judgments on some relevant and available cues, such as how difficult the items were and how successful the class was known by the teachers at solving such questions, and so restricted the factors to item difficulty and class performance. Teachers of high-performance students (i.e., high-scorers) could competently monitor their classes' academic performance; therefore, their score estimations were expected to be more accurate than those of low-performance classes. However, the teachers of low-performance students were expected to obtain more accurate estimations particularly for the difficult questions than the easier ones, since these students would already obtain lower scores for the difficult questions and their teachers would in turn yield closer predictions for these questions. Based on the earlier common finding that teachers have an overestimation tendency of their students' performance at standard tests (Urhahne & Wijnia, 2021), we expected that teachers would yield a clearer overestimation tendency when making overall estimations for the whole test (the expected average score that the whole class would obtain from the test) than making item-by-item estimations (number of students in the class who were expected to solve the given question correctly) since item difficulty would not be a salient cue as when the estimations made for each question separately.

Method

Participants

The current long-term research project involved a chain of studies on metacognitive performance (e.g., monitoring and score estimations) of sixth-grade mathematics students and their teachers. Therefore, as a participant recruitment strategy, we focused on recruiting the students first since we expected that their teachers would then be contacted to ask for

their volunteer. For this purpose, all students were selected from public schools and the same province only where their teachers are appointed based on their nationwide exam scores (i.e., public personnel selection exam: for teachers) thereby rendering the teachers' experiences, ages, and teaching efficacy comparable. The schools varied concerning their students' socio-economic status (e.g., A, B, & C, where A refers to the highest SES) and students' national exam results that were obtained in the previous year. The Ministry of Education, which collects and keeps all the related data itself, provided the research team with the features of the schools in terms of these two features, SES and average national exam scores. According to this data, the student population consisted of 448 state primary schools in 30 districts of Izmir, Türkiye covering 1822 classes and 45,069 sixth-grade students in total (note that the sixth graders in the Turkish Education system strictly cover only those ageing between 11 and 12). The minimum sample size meeting 99% of the confidence level and 2% margin error for the population was calculated as 3809 (Thompson, 2012). The project team selected the to-be-reached schools that were clustered in terms of their SES and average national exam scores with a stratified random sampling in a way that the proportions of these clusters existing in the selected sample were identical to the proportions of such clusters existing in the whole population. The project team initially contacted the designated schools' administrations and the school administrations wrote to the parents of these students to get their consent.[2] Eventually, 2832 sixth-grade students (1410 male & 1422 female) from 15 elementary schools, of which the average student number was 189 (range = 90–239), volunteered to participate in the study.[3]

The teachers who participated in the study were the mathematics teachers of a sample of sixth-grade students who took the test. Out of 67 mathematics teachers whose classes took the test, 38 of them (10 male & 28 female; age: $M = 37.3$, $SD = 0.4$) who were teaching sixth-grade students in 86 classes ($N = 2608$) in these selected schools volunteered to participate in the study. Amongst the teachers who took part in the study, 13 of them had one section, 13 of them had two sections, five of them had three sections, three of them had four sections, and four of them had five sections of sixth-grade mathematics classes. The average class size of the teachers was 30.33 ($SD = 5.97$). Since the teachers were working in state schools only and so the appointment criteria are strictly defined by state regulations, the selected teachers had comparable educational backgrounds and training. The teachers who took part in the study had at least four years of experience ($M = 11.63$ years, $SD = 6.77$) and they were teaching sixth-grade mathematics to their sections six hours per week for at least eight months, thereby assuring the teachers were already cognisant of their classes' existing performance in mathematics.

[2] The school administrators' and parents' collaborations were so high that the number of parents who did not give consent was almost none.

[3] Being a part of this long-term research project, a separate study on the metacognitive monitoring performance of these students measured with type-2 signal detection theory's calculation method along with their pre-test and immediate post-test score estimations has been reported elsewhere (see, Basokcu & Guzel 2022 for this earlier report). Since the present paper is primarily on teachers' JAs, we didn't repeat any of the findings regarding the students' monitoring and score estimation performance in this current report. Therefore, we would like to refer the readers who would be curious about how sixth-grade mathematics learners behave metacognitively before, during, and immediately after the test is completed to the earlier report.

Materials

Mathematics Test. An 11-item test was developed to assess the mathematical abilities of the sixth graders. Seven of the questions were multiple-choice, three of them were short-answer, and one of them was a true-false question, where the latest one involved three sub-question items.[4]

The test was developed equivalently to the PISA-mathematics section and involved algebraic statements, area calculations, geometrical objects, and volume calculations that were written appropriately for the sixth grade. The test in the current study was developed in the earliest phase of this research project with the participation of 556 sixth-grade students (265 male & 291 female) who were selected from the same student population in Izmir and with the same participant selection and recruitment procedure.[5] This test contained several question items created by the researchers together with the original PISA questions to determine whether the items created would be measuring the same constructs of the original PISA test items. In this test development phase of the project, which showed that the created items measured the same constructs of the original PISA test equally well, the mean item difficulty of the test was found 0.35 (range = 0.06 to 0.76), the mean discrimination index was 0.46 (range = 0.21 to 0.65), the mean point-biserial correlation was 0.52 (range = 0.33 to 0.73), and Kuder-Richardson-20 was 0.71. These analyses showed that the developed test had high reliability (Gronlund, 1993). Item-Response Theory (IRT) analyses (i.e., 2-parameters logistic model) showed the mean of threshold b-parameter, indexing item difficulties, was 0.70 (range=-1.71 to 2.86), and the mean of the slope 'a' parameter, measuring item discrimination, was 1.25 (range = 0.44 to 3.26). These analyses confirmed that the test's reliability was high, the test's items were difficult, and the items had high item discrimination (Lord, 2012; also see, Basokcu & Canpolat 2018 for the full participants' characteristics and further psychometric properties of the developed test).

Procedure

The students completed the test on the same day and at the same time in 15 elementary schools. The Ministry of Education officers and the teachers who were working on this day were responsible for managing the data collection process. The officers dispatched the material to the schools and collected them back from the teachers who were working on the test-taking day and were appointed by the schools' administrations to run the exam (note that the teachers who made score estimations did not run the exam themselves). Students who did not wish to participate in the study were asked to stay in the class and to read silently whilst others were being tested and so asked to hand in the test empty once the exam was over. The invigilators also informed the students on how to code their responses on the optical answer sheets, the duration of the test (45 min), and the scoring rules. They also asked the students

[4] The students made algebraic calculations on the test booklet to reach the correct answers, e.g., volume calculation, for the short-answer questions. Hence, only the 'final answers' given to the short-answer questions were scored in terms of their accuracies. Also, as indicated in the test booklet, all of the three sub-items in the true-false question must have been answered correctly to count this true-false question was accurately responded.

[5] The teachers in the present study confirmed that neither themselves nor their students who took part in the current study had already taken part in the test development phase of the research project earlier as an invigilator or as a participant.

not to use any calculator and not to check their course materials, such as course books or notes, when being tested.

After signing the written informed consent, the mathematics teachers of each class that took the exam were provided with the test booklet along with a separate form to write down their predictions right after the exam was over in the staffroom (note that the teachers had no contact with the students earlier or meanwhile). The teachers were first informed of the number of students who took the test in their section(s) since some students might have been absent on the test-taking day or did not wish to complete the test even though they had started to solve it.[6] Additionally, the students and teachers were not instructed that the questions varied in terms of difficulty. The teachers estimated the scores of their own class or classes for each question, referring to the item-by-item estimations with the following question: 'how many students in your class (section X) might have solved this question correctly?'. They also estimated the average correct number of questions that the whole class might have solved, which referred to the overall estimation, via responding to this question: 'what is the average score you think the whole class (section X) will obtain from this test?'. Being a self-paced one, the study lasted between 20 and 30 min for the teachers to complete.

Data analytic plan

Before running the analyses, we first calculated the teachers' item-by-item estimations ($P_{E.item}$) and their overall estimations ($P_{E.overall}$) along with the number of students who solved the items correctly ($P_{A.item}$ & $P_{A.overall}$, respectively). Estimated and actual scores were converted into percentages (e.g., [teacher's estimation of how many students can solve the question correctly / class size] x 100) since class sizes were not equal. Item-by-item judgment accuracies and overall judgment accuracies (JA_{item}, $JA_{overall}$) were measured in terms of the divergence of the teachers' estimated number ($P_{E.item}$) and their estimated scores ($P_{E.overall}$) from the actual number of students who solved the questions correctly and the whole classes' actual scores, respectively ($P_{A.item}$ & $P_{A.overall}$). Therefore, the following formulae were used to calculate the divergences: '$JA_{item} = P_{E.item} - P_{A.item}$', and '$JA_{overall} = P_{E.overall} - P_{A.overall}$'. The subtractions yielded either a minus or a plus sign unless 'P_E' and 'P_A' values were even. If the subtraction is not equal to '0', the directions were either towards underestimation or overestimation ('-' & '+' values, respectively).

Second, the items were clustered into three item-difficulty categories before running the analyses: easy, moderately difficult, and difficult questions. Frequency analyses (i.e., number of students who solved the item correctly out of the total number of students who took the test) revealed that the test involved two easy (range = 0.70 to 1.00), three moderate (range = 0.31 to 0.70), and four difficult questions (range = 0.00 to 0.30). The IRT difficulty parameter showed the categories dispersed similarly to how the items were clustered (Gronlund, 1982; Haladyna, 2004).

Lastly, the classes were categorised into three performance groups based on their average test scores (i.e., the average number of correct responses): low-, medium-, and high-performance groups. The mean performance of all classes was $M = 3.36$ ($SD = 0.68$). *Z*-statistics allowed us to determine the cut-off scores for the performance groups. As a result of this

[6] Absenteeism on the test-taking day was almost none and the number of students who didn't wish to complete the test was negligible.

statistic, 23 classes constituted the low-performance group (M=2.66, SD=0.30), 41 classes set up the medium-performance group (M=3.28, SD=0.21), and the remaining 22 classes constituted the high-performance group (M=4.23, SD=0.61). The analyses considered the above-mentioned item-difficulty categories and were conducted between the teachers of low-, medium-, and high-performance classes. The necessary criteria to use parametric tests, i.e., having random assignment, using interval data, and revealing normal distributions, were met; therefore, analysis of variance (ANOVA) and Pearson's product-moment correlation (r) statistics were used for the data analyses.

Results

Estimation type

Item-by-item estimations

A 3(item difficulty: easy, moderate, difficult) x 3(class performance: low-, medium-, high-performance) mixed ANOVA was run to detect the effects of item difficulty and class performance on teachers' directional JAs that were obtained for each question individually (item-by-item estimations). Item difficulty was a within- and class performance was a between-subjects factor. The results showed that the item difficulty main effect was significant; $F(1.81, 150)=526.37$, $p<.001$, $\eta^2=0.86$. Judgment accuracies for easy, moderate, and difficult questions were significantly different from each other in the following way: whereas the teachers underestimated their students' actual performance for easy questions (M=-34%, SEM=2) and moderate questions (M=-11%, SEM=1.8), they overestimated the students' performance for difficult questions (M=10%, SEM=1.5); see Fig. 2. The class performance main effect was also significant; $F(2, 83)=4.357$, $p=.016$, $\eta^2=0.10$. Though the teachers of medium-performance group (M=-11%, SE=2.2) did not differ from the remaining groups, the underestimation degree of the low-performance classes' teachers was significantly higher (M=-18%, SE=3) than the divergence of those teachers who were teaching high-performance classes (M=-5%, SE=3). The results also showed a significant interaction effect between item difficulty and class performance on teachers' directional accuracies; $F(3.61, 150)=5.943$, $p<.001$, $\eta^2=0.13$; see Fig. 2. Post-hoc comparisons revealed that the estimations made for easy questions were significantly lower than those for moderate as well as difficult questions in each teacher group. The teachers of the low-performance classes had significantly clearer underestimation (-44%) than the teachers of high-performance classes (-22%) whilst the teachers of medium-performance classes did not differ significantly from any other groups in terms of estimations made for easy questions (-36%). For the moderate questions, each group had comparable estimations. The pattern, however, conversed for the difficult questions. The teachers of high-performance classes yielded significantly clearer overestimation (14%) than those of low-performance classes (5%) whilst the teachers of medium-performance group did not differ from any other teacher groups (12%); also see, Table 1 for the teachers' estimated scores and their classes' actual scores.

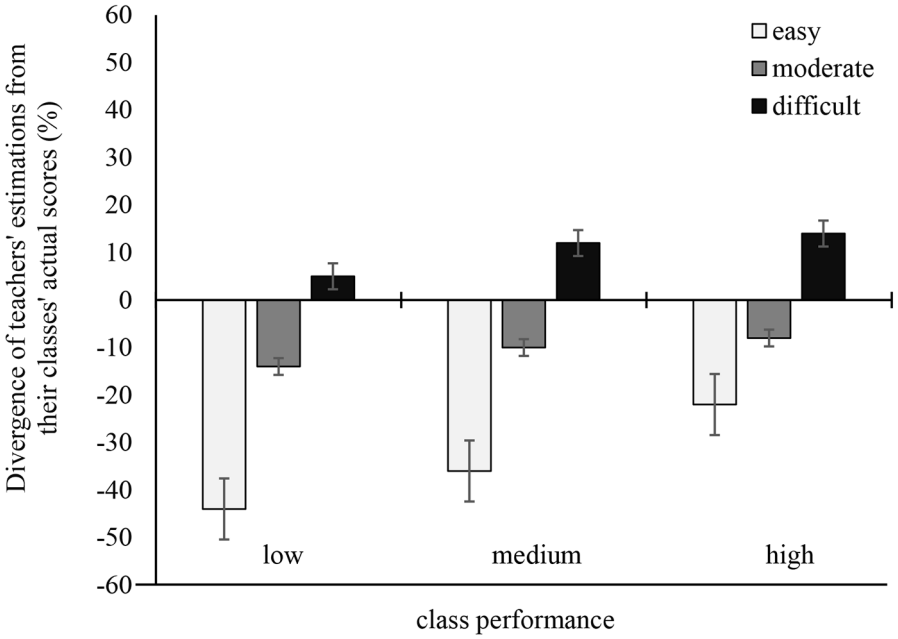

Figure 2 *Teachers' directional judgment accuracies with respect to item difficulty (easy, moderate, & difficult) and class performance (low-, medium-, & high-performance class)*
Note: The horizontal line crossing the level '0' is where the average actual scores of the classes are (see Table 1 for the actual scores). Therefore, the values below '0' indicate their teachers' underestimation tendencies and those above '0' indicate the teachers' overestimation tendencies. Standard errors are shown with error bars

Overall estimations

A one-way ANOVA was run to reveal whether class performance affected the teachers' directional judgment accuracies (note that the item difficulty factor cannot be observed when making overall estimations). The results showed that class performance did not affect the accuracies; $F(2, 72) = 0.007$, $p = .993$. As shown in Table 1, the teachers of low-performance, medium-performance, and high-performance classes tended to overestimate actual scores of their classes' performance with less than a 10% divergence ($M = 8.2\%$; $SEM = 3.6$; $M = 8.7\%$, $SEM = 2.3$; $M = 8.4\%$, $SEM = 3$, respectively). Although the teacher groups varied in their item-by-item estimations, they were similarly accurate in their JAs (i.e., diverged upwards from the actual scores comparably) when they made overall estimations.

Class performance

A 2(estimation type: item-by-item vs. overall) x 3(class performance: low-, medium-, high-performance) mixed ANOVA was run to detect the effects of estimation type, being a within-subjects factor, and class performance, a between-subjects factor, on the directional judgment accuracies (note that overall and item-by-item estimations were analysed by excluding the item difficulty variable since overall estimations do not take item dif-

Table 1 *Teachers' Item-by-item and Overall Estimations and Their Classes' Actual Performance (%) with Respect to Item Difficulty (Easy, Moderate, Difficult) and Class Performance (Low-, Medium-, High-performance classes); and Pearson's r Correlations Between the Estimated and the Actual Scores*

Estimation type	Class performance (n=teacher & student)	Item-difficulty [a]	Estimated score (%)	Actual score (%)	r
Item-by-item	low (n=23 & 692)	easy	22.8	67.2	0.23
		moderate	17.1	31.3	0.09
		difficult	16.3	11.4	0.46*
	medium (n=41 & 1321)	easy	40.8	76.6	−0.35*
		moderate	31.5	41.4	0.04
		difficult	26.5	14.8	0.16
	high (n=22 & 595)	easy	64.2	86.6	0.21
		moderate	47.9	55.6	0.21
		difficult	36.3	22.3	0.43*
Overall [b]	low (n=16)		38.2	30.0	−0.22
	medium (n=36)		45.1	36.4	−0.01
	high (n=21)		55.6	47.1	0.29

[a] = the test involved two easy, three moderate, and four difficult questions

[b] = some teachers skipped giving an overall estimation after they made estimations for each question individually; therefore, the estimated and actual scores and the correlations between them considered only those classes whose teachers made an overall score estimation for their class(es)

* $p<.05$

ficulty into account). The results revealed a significant estimation type main effect; $F(1, 70)=70.311$, $p<.001$, $\eta^2=0.50$. Whereas teachers' item-by-item estimations were below the actual scores of their classes obtained (i.e., underestimation) (M=-6.1%, SEM=1.8), their overall estimations were above their classes' actual scores (M=8.4%, SEM=1.7). The results, however, did not show any class performance main effect and any interaction effect between estimation type and class performance; $F(2, 70)=0.911$, $p=.407$ and $F(2, 70)=2.581$, $p=.083$, respectively.

Correlations between the estimated and the actual scores

Pearson's product-moment correlations (r's) between the teachers' estimated and their students' actual scores in terms of item difficulty and class performance were also calculated (see the far-right column in Table 1). The results showed that the estimations made by the teachers of low-performance classes and their students' actual scores for the difficult questions were significantly correlated ($r(23)=0.46$, $p<.05$) unlike the correlations observed for the easy and the moderate items. As displayed in Fig. 2, teachers of low-performance classes showed a clearer underestimation tendency for the easy and the moderate questions than the remaining groups and so their estimations yielded a significant correlation with their classes' actual scores for the difficult questions only. In other words, they could better

predict their classes' poor performance for the difficult questions, yet they still discredited their students' better performance for the easy and the medium questions. The teachers of medium-performance classes and their students' actual scores were negatively correlated for the easy questions; $r(39)$=-0.35, $p<.05$. On the other hand, the estimations made by the teachers of high-performance classes and their students' actual scores for the difficult questions were positively correlated; $r(20)=0.43$, $p<.05$. That is, the teachers of high-performance classes seemed to estimate their students' relatively better performance for the difficult questions well enough unlike what they estimated for the easy and moderate questions. It should, however, be noted that their judgments were towards overestimation for the difficult questions (see, Fig. 2). Lastly, the correlations between the teachers' overall estimations and their students' actual scores were not correlated in any of the teacher groups.

Discussion and conclusions

The present study investigated a sample of sixth-grade mathematics teachers' JAs by varying the item difficulty, estimation type, and class-performance variables. Previous research has shown that teachers mostly retain an overestimation tendency when estimating their students' academic performance (Bates & Nettelbeck, 2001; Doherty & Conolly, 1985; also see, e.g., Urhahne & Wijnia, 2021 for an extensive review). Though this pattern was observed in this study as well, we found that the direction of estimations varied depending on certain factors. The following paragraphs list these factors together with our related findings.

First, whereas the teachers in this study showed an overestimation tendency for the difficult questions, they retained an underestimation tendency for easy and moderately difficult questions. That is, the students could solve difficult questions not as correctly as their teachers estimated, and they could also solve easy questions more accurately than what their teachers estimated, implying that the teachers did not particularly consider the item difficulty when making estimations.

Second, just like the finding that item difficulty altered the direction of estimation, our results also showed that estimation type changed the teachers' estimation patterns. Whereas item-by-item estimations yielded differential estimation tendencies depending on the item difficulty and class performance, teachers' overall estimations showed an overestimation tendency regardless of class performance. This overestimation tendency, which has been commonly found in previous research, was also observed in this study particularly when teachers made overall estimations (note that item difficulty cannot be a to-be-measured variable when analysing the overall estimations). Interestingly, however, the teachers changed their judgments whilst making overall estimations although the 'average' of what one already estimates for each question individually (item-by-item estimations) should be equal to their overall estimations. Despite that convenience, the teachers seemed to overlook the item difficulty even further when they were asked to make overall estimations and used some alternative cues, presumably previous performance of their class(es) in similar tests (see the second next paragraph for the discussion of this possibility).

Third, teachers of low-performance classes were found to be less accurate in predicting their students' actual scores than teachers of high-performance classes. This finding is critical since teachers of high-performance classes could have easily yielded a clearer

underestimation than those of low-performance classes should both teacher groups have retained comparable estimations. That is, high-performance students had already obtained better scores than the low-performance students, which thereby was conducive to allowing more space for their teachers to diverge further away from the actual scores towards *underestimation*. On the contrary, they did not. This allowance was also valid for the teachers of low-performance students in the way that they could have a clearer *overestimation* tendency if they had held similar estimations to the teachers of high-performance classes. Yet again, these teachers did not show clearer overestimation than the teachers of high-performance classes. As shown in Fig. 2, teachers' underestimation gradually diminished from the teachers of low-performance classes to those of high-performance classes. The overestimation tendency, on the other hand, showed the opposite pattern. In short, the teachers in the present study seemed to overvalue their classes' actual performance (as were indexed with and so implied by the performance clusters) by discrediting the to-be-obtained achievements of the low-performance classes or overvaluing this achievement in the high-performance classes further.

The above-mentioned findings imply that the teachers must have used primarily the existing or recent class performance as a judgment cue when making their score estimations rather than the item difficulty.[7] Why exactly such an implication was inferred is as follows. Amongst some possible others, one (in this case, teachers) may use at least two reasonable cues when they judge whether their students will correctly solve any given question: the difficulty of the question item and how well the students are considered in solving such questions (e.g., algebraic calculus, volume/area calculations, etc.). Therefore, teachers should have needed to use both cues equally well to yield an accurate estimation. The results revealed that the teachers of low-performance classes showed clearer underestimations for their classes' test performance than those of high-performance classes and their underestimations were even the clearest for the easiest questions. In other words, the number of students who solved easy questions correctly in low-performance classes was higher than what their teachers predicted. However, should the teachers have assessed these items as easy ones for the class, they would have made higher (and so accurate) estimations for them. Likewise, the students in the high-performance classes did not solve particularly the difficult questions unlike what their teachers predicted. Again, they would have made lower (and so accurate) estimations for their classes if those teachers had considered such items were in fact difficult ones for their classes.

Some earlier works showed that teachers accurately estimate the task difficulty (Conejo et al., 2014; Impara & Plake, 1998; also see, e.g., van de Watering & van der Rijt, 2006 for a review) whereas some more recent evidence has shown that students are better at estimating the task difficulty than teachers (e.g., Wauters et al., 2012). The latter finding seems in line with the 'curse of knowledge' phenomenon (Birch, 2005), referring to the condition that people mainly use what they know when predicting what others know (Fussell & Krauss, 1992; Nickerson et al., 1987). To state it briefly, our findings showed that item difficulty was secondary to another possible cue when teachers estimate their students' test scores, which we infer must be how teachers assess their classes' existing performance in solving similar questions.

[7] Note that the test was not prepared by the teachers themselves and neither the students nor the teachers in the present study were instructed that the questions varied in terms of difficulty.

Previous research has tested some explanations of what cue(s) people use when guessing or predicting what others know. For instance, Tullis (2018) conducted a series of experiments where he tested his cue-utilisation approach and compared his proposition with other possible suggestions to explain how we predict others' knowledge, such as using static knowledge cues and the anchoring-and-adjustment approach (e.g., first using what we know regarding the answers for the questions asked and then adjusting our estimations for others' responses via releasing from this egocentric thinking; see also e.g., Nickerson, 1999; Nickerson et al., 1987). Tullis also showed that we weigh available and salient cues in a dynamic way and do not retain the same way of thinking in all conditions whilst estimating one's own post-test scores, whereas some other works have shown that one is affected mostly by what they initially hold and maintain that as a reference, i.e., as an anchor (e.g., Bard & Weinstein, 2017; Jackson & Greene, 2014; Weinstein & Roediger, 2010). It seems that the teachers in this study were somehow affected by their anchoring and had difficulty adjusting their reference assessments when estimating their students' to-be-obtained scores.

Additionally, as was proposed by Koriat (1997) in his cue-utilization approach for JOL, item difficulty can be considered as a possible intrinsic cue and class performance as a mnemonic cue when estimating others' test scores. Therefore, our findings suggest that the teachers relied more on the mnemonic cues than intrinsic cues when estimating the students' prospective performance. When the proposed multi-layered working model is considered as an extension of JOL, referring briefly to the judgments one makes on their learning (i.e., acquired knowledge), to another likely judgment that can hereinafter be called *judgment of others' learning* (JOOL), this type of judgment now comes about not regarding the knowledge about one's own but of others' knowledge. Therefore, if JOOL shall be measured with teachers' judgment accuracies, we expect that counting on what one already knows may be a more convenient cue to use when predicting others' knowledge levels than relying on another available cue, which is now concerning the items themselves, e.g., item difficulty (also see, e.g., Oudman et al., 2018). In other words, a teacher may not perfectly assess how difficult an item is for the students and so they may need some other possibly available cues, such as what they know about their classes' past performance and/or in-class performance they have observed so far. Previous research also seems to support this speculation. For instance, the past performance of classes was shown to be a highly diagnostic cue for their teachers to make their judgments, particularly when the domains of the to-be-judged test and the previous scores match (Hecht & Greenfield, 2002; Martínez et al., 2009).

Future work may further investigate the effects of setting a clear metacognitive goal to regulate learners' knowledge (i.e., meta-level) along with metacognitive 'control' processes in this regulation as well since this study primarily investigated the 'monitoring' process in the regulation of learners' knowledge via calculating teachers' judgment accuracies. How the link between accurate monitoring and better control (e.g., accurate teacher judgment that is linked to better-adjusted instruction methods) is set towards the final aim that is expectedly to yield higher academic achievement for the students also needs further research (see e.g., Thiede et al., 2019 for further discussion). Besides investigating the teachers' monitoring process only as depicted in the proposed multi-layered model, the following should also be noted as the limitations of the present study. For instance, the sample involved the sixth-grade mathematics teachers only, the total number of the questions in the test was not high and the number of items was not even amongst the clusters of item difficulty, and the items were categorised into various difficulty levels 'after' the test scores were obtained (i.e., by

calculating the number of students who solved the question correctly out of the total number of students taking the test). Also, the teaching domain and the content of the test measured matched in this study; therefore, this variable was not manipulated. Owing to their specific teaching objectives, some courses may also assess students' performance mostly in a similar fashion, whereas others may use several heterogeneous performance assessments, which thereby can affect the accuracy of teachers' estimations.

Despite the above-mentioned limitations, we believe that the current multi-layered metacognitive regulation model in teaching has the potential to drive and answer various research questions. Based on the basic premises and the conceptualisations offered in this revisited model, future research may consider varying the item difficulty 'before' the test administration and/or inform the teachers explicitly that the items vary in terms of difficulty to reveal whether having this information alleviates teachers' under- and over-estimations. They may additionally assess teachers' existing evaluation of their classes, which we did not directly obtain yet inferred from the teachers' score estimations. Besides these suggestions, using an alternative measurement method of teacher JAs rather than the conventional method that uses prepared test(s) or a given task can reveal these JAs more realistically -in other words, closer to the practical application in the education settings-, such as by asking teachers to create their own question items in a way that they predict a certain number of students who would solve these items correctly and compare them against the learners' actual scores.

Lastly, the findings obtained in this study have some pedagogical implications as well. For instance, once the classes' expected success is mainly affected by the teachers' assessments of their classes' existing performance and if this perception is somehow overvalued, the students either in a low-performance or a high-performance class are disadvantaged by their teachers' earlier assessments (see e.g., Zhu et al., 2018). Teachers of high-performance classes may be inclined to maintain a higher criterion for their classes to achieve and so they may miss the needs of those students who lack behind the class. The disadvantage of the low-performance classes, however, can at least be two-fold. Along with possibly being less motivated to teach in such classes, teachers of low-performance classes may not only perpetuate their available assessments and perceptions in their teachings, such as maintaining lower learning objectives for their classes, but they may also be oblivious to those students who would need a higher-level instruction beyond what they are already better at. As what has been shown in earlier works, the effectiveness of teaching largely depends on teachers' accurate monitoring of their class(es)' real performance -no matter what it may be-, which eventually should allow them to adjust their instructional activities and educational strategies fittingly well to their students' true knowledge levels (Brendefur et al., 2016; Thiede et al., 2003, 2019). In short, though metacognitive regulation of others' knowledge would entail the monitoring and control processes that are linked through a meta-level, even 'accurate monitoring' of students' performance or class performance alone seems to appear as a critical element in regulating others' knowledge, i.e., in teaching.

Funding This work was financed by The Scientific and Technological Research Council of Türkiye (TÜBİTAK), Grant No. 115K531.

Declarations

Conflict of interest We have no known conflict of interest to disclose.

References

Bard, G., & Weinstein, Y. (2017). The effect of question order on evaluations of test performance: can the bias dissolve? *Quarterly Journal of Experimental Psychology (2006)*, *70*(10), 2130–2140. https://doi.org/10.1080/17470218.2016.1225108.

Basokcu, T. O., & Canpolat, A. (2018). Ankor Test Deseninde Bilissel Tanı Modeli Ortuk Yetenek Sınıfları ile Test Esitleme Calismasi (A Test Equation Study with Latent Classes following Cognitive Diagnostic Model as an Anchor Test Design). *6th International Congress on Measurement and Evaluation in Education and Psychology*, 384–388.

Basokcu, T. O., & Guzel, M. A. (2022). Beyond counting the correct responses: metacognitive monitoring and score estimations in mathematics. *Psychology in the Schools*, *59*(6), 1105–1121. https://doi.org/10.1002/pits.22665.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology*, *21*(2), 177–187. https://doi.org/10.1080/01443410020043878.

Birch, S. A. J. (2005). When knowledge is a curse: children's and adults' reasoning about mental states. *Current Directions in Psychological Science*, *14*(1), 25–29. https://doi.org/10.1111/j.0963-7214.2005.00328.x.

Brendefur, J. L., Thiede, K., Strother, S., Jesse, D., & Sutton, J. (2016). The effects of professional development on elementary students' mathematics achievement. *Journal of Curriculum and Teaching*, *5*(2), 95. https://doi.org/10.5430/jct.v5n2p95.

Carr, M. (2010). The importance of metacognition for conceptual change and strategy use in mathematics. *Metacognition, strategy use, and instruction* (pp. 176–197). Guilford Press.

Conejo, R., Guzmán, E., Perez-de-la-Cruz, J. L., & Barros, B. (2014). An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications*, *41*(2), 594–606. https://doi.org/10.1016/j.eswa.2013.07.084.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives*, *8*, 1. https://doi.org/10.14507/epaa.v8n1.2000.

Doherty, J., & Conolly, M. (1985). How accurately can primary school teachers predict the scores of their pupils in standardised tests of attainment? A study of some non-cognitive factors that influence specific judgements. *Educational Studies*, *11*(1), 41–60. https://doi.org/10.1080/0305569850110105.

Duffy, G. G., Miller, S., Parsons, S., & Meloth, M. (2009). Teachers as metacognitive professionals. *Handbook of Metacognition in Education*. Routledge.

Dunlosky, J., & Metcalfe, J. (2009a). Judgments of learning. *Metacognition*. Sage Publications, Inc.

Dunlosky, J., & Metcalfe, J. (2009b). *Metacognition*. Sage Publications, Inc.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*(3), 378–391. https://doi.org/10.1037//0022-3514.62.3.378.

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.

Gralewski, J., & Karwowski, M. (2019). Are teachers' ratings of students' creativity related to students' divergent thinking? A meta-analysis. *Thinking Skills and Creativity*, *33*. https://doi.org/10.1016/j.tsc.2019.100583

Gronlund, N. E. (1982). *Constructing achievement tests*. Prentice-Hall.

Gronlund, N. E. (1993). *Measurement and evaluation in teaching* (10th ed.). New Jersey: The Macmillan Company.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items, 3rd ed* (pp. xi, 306). Lawrence Erlbaum Associates Publishers.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009.

Hecht, S. A., & Greenfield, D. B. (2002). Explaining the predictive accuracy of teacher judgments of their students' reading achievement: the role of gender, classroom behavior, and emergent literacy skills in a longitudinal sample of children exposed to poverty. *Reading and Writing*, *15*(7), 789–809. https://doi.org/10.1023/A:1020985701556.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69–81.

Jackson, A., & Greene, R. L. (2014). Impression formation of tests: retrospective judgments of performance are higher when easier questions come first. *Memory & Cognition*, *42*(8), 1325–1332. https://doi.org/10.3758/s13421-014-0439-5.

Jenkins, L. N., & Demaray, M. K. (2016). Teachers' judgments of the academic achievement of children with and without characteristics of inattention, impulsivity, and hyperactivity. *Contemporary School Psychology*, *20*(2), 183–191. https://doi.org/10.1007/s40688-015-0073-7.

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

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

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.

Machts, N., Kaiser, J., Schmidt, F. T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: a meta-analysis. *Educational Research Review*, *19*, 85–103. https://doi.org/10.1016/j.edurev.2016.06.003.

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: evidence from the ECLS. *Educational Assessment*, *14*(2), 78–102. https://doi.org/10.1080/10627190903039429.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition*, *18*(2), 196–204. https://doi.org/10.3758/BF03197095.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of Learning and Motivation—Advances in Research and Theory* (pp. 125–173). https://doi.org/10.1016/S0079-7421(08)60053-5

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the "delayed-JOL effect". *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x.

Nelson, T. O., Kruglanski, A. W., & Jost, J. T. (1998). Knowing thyself and others: progress in metacognitive social psychology. *Metacognition: cognitive and social dimensions* (pp. 69–89). Sage Publications, Inc. https://doi.org/10.4135/9781446279212.n5.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition?. *Metacognition: knowing about knowing* (pp. 1–25). The MIT Press.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737.

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, *64*(3), 245–259. https://doi.org/10.1016/0001-6918(87)90010-2.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237.

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.

Prohaska, V. (1994). "I know I'll get an A": confident overestimation of final course grades. *Teaching of Psychology*, *21*(3), 141–143. https://doi.org/10.1177/009862839402100303.

Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2016). Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educational Psychology*, *36*(5), 863–878. https://doi.org/10.1080/01443410.2014.998629.

Ready, D. D., Wright, D. L., 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.

Rhodes, M. G. (2016). Judgments of learning: methods, data, and theory. *The Oxford handbook of metamemory* (pp. 65–80). Oxford University Press.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(1), 204–221. https://doi.org/10.1037/0278-7393.26.1.204.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627.

Svanum, S., & Bigatti, S. (2006). Grade expectations: informed or uninformed optimism, or both? *Teaching of Psychology*, *33*(1), 14–18. https://doi.org/10.1207/s15328023top3301_4.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66.

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., Oswalt, S., Snow, J. L., Sutton, J., & 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. E. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (pp. 678–695). Cambridge University Press. https://doi.org/10.1017/9781108235631.027.

Tullis, J. G. (2018). Predicting others' knowledge: knowledge estimation as cue utilization. *Memory & Cognition*, *46*(8), 1360–1375. https://doi.org/10.3758/s13421-018-0842-4.

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

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: a review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, *1*(2), 133–147. https://doi.org/10.1016/j.edurev.2006.05.001.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: an auspicious collaboration between data and judgment. *Computers & Education*, *58*(4), 1183–1193. https://doi.org/10.1016/j.compedu.2011.11.020.

Weinstein, Y., & Roediger, H. L. (2010). Retrospective bias in test performance: providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition*, *38*(3), 366–376. https://doi.org/10.3758/MC.38.3.366.

Zhu, M., Urhahne, D., & Rubie-Davies, C. M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology*, *38*(5), 648–668. https://doi.org/10.1080/01443410.2017.1412399.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: historical background, Methodological Developments, and future prospects. *American Educational Research Journal*, *45*(1), 166–183. https://doi.org/10.3102/0002831207312909.