Received: 5 June 2021 | Revised: 11 January 2022 | Accepted: 17 January 2022

DOI: 10.1002/pits.22665

RESEARCH ARTICLE

WILEY

Beyond counting the correct responses: Metacognitive monitoring and score estimations in mathematics

Tahsin Oğuz Başokçu[1] | Mehmet Akif Güzel[2]

[1]Department of Educational Sciences, Ege University, Izmir, Turkey

[2]Department of Psychology, Abdullah Gül University, Kayseri, Turkey

Correspondence

Tahsin Oğuz Başokçu, Ege University, Gençlik Cad., No: 12, Izmir, Bornova 35040, Turkey.
Email: tahsin.oguz.basokcu@ege.edu.tr

Funding information

The Scientific and Technological Research Council of Turkey (TUBITAK), Grant/Award Number: 115K531

Abstract

This study investigated how well students differentiate their responses' accuracies (metacognitive monitoring) and estimate their test scores beyond counting—and counting on—the number of correct responses alone. Monitoring abilities of 2832 sixth-graders (1410 male and 1422 female native in Turkish) at an 11-item Program for International Student Assessment (PISA)-equivalent mathematics test were measured via response-contingent Type-2 signal detection theory. The students also made score estimations right before and immediately after completing the test (pre- and posttest estimations, respectively). Although high-scoring students underestimated and low-scoring ones overestimated how they would perform in the test, high-scorers were accurate in their posttest estimations unlike the low-scoring group, where the lattaer retained their overestimation tendencies. Having better monitoring performance, the high-scoring group could subsequently calibrate their posttest estimations. Additional assessment methods such as measuring monitoring and score estimations seem to have the potential to reveal how mathematics students behave before, during, and after responding.

KEYWORDS

mathematics, metacognitive monitoring, score estimations

1 | INTRODUCTION

Imagine two students who exhibit very low performance at a given test and you ask them what they expect their scores would be once the test is over. One student accurately predicts a very low score while the other expects a fairly better score than what they indeed obtain, where the latter exhibits an inflated self-appraisal, known as the Dunning–Kruger effect (Kruger & Dunning, 1999). Although the pair can be considered as poor performers based on their test scores and there may be various factors to explain why student estimation accuracies vary (e.g., see, Händel, de Bruin, et al., 2020; Händel, Harder, et al., 2020; Magnus & Peresetsky, 2018; Weisskirch, 2018; also see, Metcalfe, 1998), we can now get a rough yet a further assessment in terms of their overall degree of awareness on their responses. Therefore, the student assessment has been considered as one of the most critical elements in education (see, e.g., Rasooli et al., 2018 for an extensive review) that provide learners and teachers with practical feedback thereby allowing both parties to tailor and adjust their learning and teaching objectives, methods, and contents accordingly (e.g., Newton, 2007; Thiede et al., 2019). However, most of the in-class and particularly large-scale student assessments seem to rely more on structured tests where the number of correct answers is counted for grading due to their relatively better convenience. Even so, mere counting of correct responses does not necessarily reveal how exactly the respondents (e.g., learners) behave when responding, such as how well they are aware of their responses' correctness (O. T. Basokcu & Guzel, 2020; Higham & Arnold, 2007; Koriat & Goldsmith, 1996; also see, e.g., Lai, 2011 for a review).

Investigating the students' higher-order cognitive abilities (i.e., metacognitive performance) at cross-culturally-used standardized tests such as Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) is not something new (e.g., Callan et al., 2017; Gamazo & Martínez-Abad, 2020; Säälik et al., 2015). These large-scale tests can measure the students' higher-order cognitive abilities besides their overall test scores. For instance, one of the feasible options in this arrangement is to develop the questions in a way that they directly assess the students' abilities of reasoning and argumentation, formulating situations mathematically, and so forth (see, e.g., OECD, 2013; Stacey & Turner, 2015). Free-response items have also been considered as a fruitful method to measure students' problem-solving strategies, how they approach the questions, their misconceptions about the problem, and alike (Lie et al., 1996; also see, O'Neil & Brown, 1998 for a comparison between free- and forced-choice tests). Also, a relatively recent approach that was developed to assess students' metacognitive abilities with "scenario-based metacognitive knowledge testing" (Händel et al., 2013) appears as an expanded method that measures the learners' higher-order cognitive abilities, such as performance of learning and problem-solving strategies. Previous works, therefore, have revealed various higher-order cognitive abilities of the students in such tests. For instance, it has been evinced that reading abilities at PISA tests are closely linked with some metacognitive abilities of the students at certain ages (e.g., Artelt et al., 2001; also see, Artz & Armour-Thomas, 1992; White & Frederiksen, 2009 for "developing metacognitive approaches for problem-solving"). As was reported by Artelt and her colleagues, the scores obtained on a test of metacognitive knowledge about reading comprehension were also found highly correlated with reading literacy at PISA 2000 test. The proceeding administrations of the test reported the same relation as well (see, e.g., Artelt & Schneider, 2015 for PISA 2009 results).

Despite the existence of various works on the test development and assessment strategies to measure learners' higher-order abilities, there is still a lack of research line that primarily aims at assessing both metacognitive and cognitive performance of the students in cross-culturally used, standardized tests without demanding to vary the question types and contents and also without administering several self-report scales along with the test, such as a particular questionnaire that assesses the respondents' metacognitive thinking strategies or knowledge (see, e.g., Anthony et al., 2013; Maag Merki et al., 2013; Wirth & Leutner, 2008 that use several related questionnaires annexed to the given tests). The study of Higham (2007), however, emerges as an exception to this. Using a specific type of signal detection theory (SDT; Green & Swets, 1966), namely Type-2 SDT, he investigated the metacognitive performance of the participants on the scholastic aptitude test (SAT). The response-contingent Type-2 SDT essentially assesses how well the respondent judges (i.e., monitors) their responses' correctness (e.g., Higham, 2002) whereas Type-1 SDT, which is also known as the

stimulus-contingent SDT, measures how accurately observer detects whether the signal buried in a noisy environment is present or absent (see the following section for further details about SDT and its types). Higham compared this calculation method with an alternative method of Koriat and Goldsmith (1996), the quantity-accuracy profiles (QAPs) and his results showed the Type-2 SDT, which indexes respondents' awareness of the correctness of their responses (i.e., metacognitive monitoring), was a sound option to the Koriat and Goldsmith's framework in the assessment of SAT scores and it allowed to measure regulation of accuracy well enough while its alternative, QAPs, could not.

Therefore, we tested the mathematical abilities of the students at the PISA mathematics test context in this study and specifically aimed at investigating how the students indeed behave before, during, and after they complete the test. Unlike the traditional gradings (e.g., scoring the test performance based on the responses' dichotomous feature: correct or incorrect), our investigation was implemented in a single testing procedure and without manipulating the questions' contents and types or without administering an attached self-report scale. Pre- and posttest judgments were indexed with score estimations and how the students behaved during responding were measured with their metacognitive monitoring scores that were calculated with the Type-2 SDT approach. Previous research shows that learners may tend to overestimate their actual performance (e.g., Jacobson, 1990; Prohaska, 1994; Svanum & Bigatti, 2006; Weisskirch, 2018) yet the accuracy of score estimation varies with some factors, such as the perception of the assessment's fairness (Cherry et al., 2003), previous academic achievement (Magnus & Peresetsky, 2018), even grit and self-esteem (e.g., Weisskirch, 2018; also see, Händel, de Bruin, et al., 2020; Händel, Harder, et al., 2020), and so forth. As a unique feature of our study, however, we investigated both score estimations and *metacognitive monitoring* performance "together," which herein refers to one's ability to discriminate their correct and incorrect "responses" (e.g., Higham, 2002). We primarily aimed to reveal the relationship between the metacognitive monitoring ability and score estimations and how these parameters are affected by the students' academic performance (e.g., being a high or a low scorer) and had an objective to later suggest researchers, test developers, and particularly the teachers a feasible method to assess their students' metacognitive abilities besides grading the test scores via counting the number of correct responses only.

Before overviewing the present study by laying out its specific aims and hypotheses, the following sections first review what is measured in the assessment of metacognitive monitoring ability along with how it is calculated with the Type-2 SDT method.

1.1 | Metacognitive monitoring and Type-2 SDT

Numerous research has extensively investigated the self-regulatory processes that are presumed to involve, for instance, monitoring and control processes in various learning contexts, such as a laboratory or a classroom setting (see, e.g., Higham & Leboe, 2011; Zimmerman, 2001 for extensive reviews). Therefore, it is no surprise to witness the emergence of various methods that have been used to assess the metacognitive abilities of the respondents, such as the QAPs (Koriat & Goldsmith, 1996) and the specific calculation methods that are based on the basic premises of SDT (Green & Swets, 1966). Overall, these measurements quantify the respondents' several metacognitive abilities including how well one regulates their memory accuracies (e.g., lowering the number of their responses [i.e., quantity] for the sake of increasing the number of accurate responses among those reported [i.e., accuracy]; see, e.g., Koriat, 1997), how accurately one differentiates their correct and incorrect responses (i.e., metacognitive monitoring), at which level they set their response criterion when they are tested under free-report option (e.g., stringently or leniently), and so on (see, e.g., Barrett et al., 2013; Galvin et al., 2003; Guzel & Higham, 2013; Higham & Arnold, 2007; Maniscalco & Lau, 2014).

Following the Type-2 SDT, for instance, Higham (2002) suggested the following calculation method to measure participants' monitoring abilities and response biases (control). Before detailing this method, however, let us briefly clarify the Type-1 SDT first which we believe would better reveal why his method is based on a particular variation of this theory; also see, for example, Figure 1 that compares these types of SDT. In conventional Type-1 SDT observations,

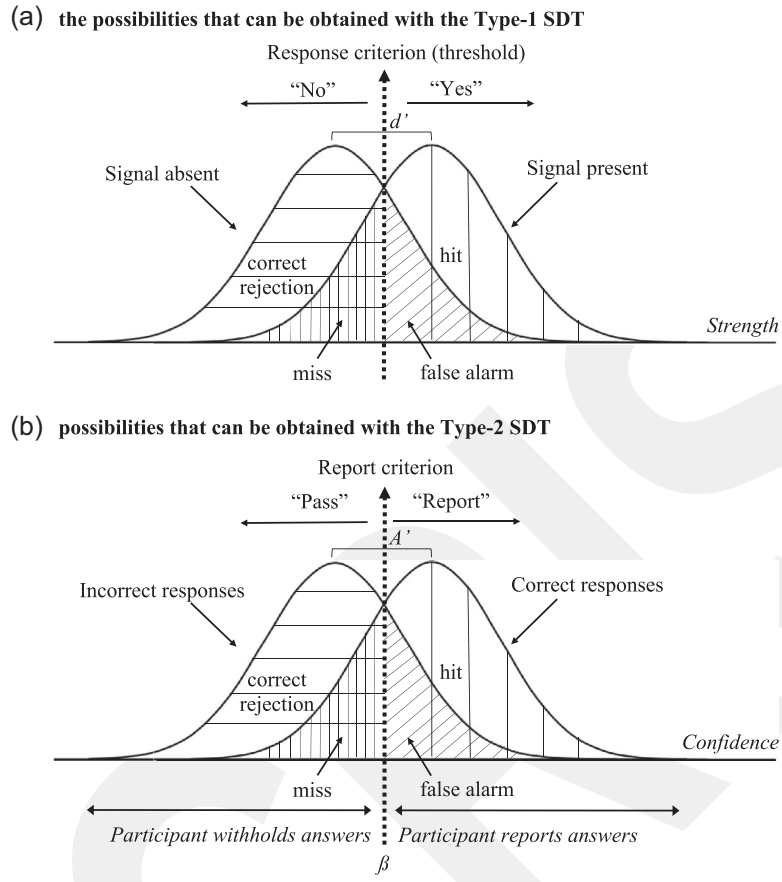

FIGURE 1 Probabilities that can be obtained according to Type-1 and Type-2 signal detection theory (SDT). (a). The possibilities that can be obtained with the Type-1 SDT. (b) Possibilities that can be obtained with the Type-2 SDT. *Source*: Higham and Arnold (2007)

the respondent simply makes a "yes/no" decision whether a signal that is embedded in a noisy environment is present or absent. The stimuli—noise and signal—are assumed to distribute normally in terms of their strengths and the detection of the signal from among the noise is gradually easier as the distributions are further apart, which is also affected by where the threshold is set (Green & Swets, 1966). Therefore, a 2 × 2 contingency table (i.e., 2[response: yes-no] × 2[signal: present-absent]) can be drawn to calculate the parameters of Type-1 SDT. These parameters refer to the following possibilities: *hit* (saying "yes" when signal present), *false alarm* (saying "yes" when signal absent), *miss* (saying "no" when signal present), and *correct rejection* (saying "no" when signal absent) (see, e.g., Abdi, 2010). Type-2 SDT examinations are also derived from the similar rationale of Type-1; however, Type-2 SDT is contingent on the responses of the "observer" on the stimuli unlike Type-1 SDT, which is contingent on the stimuli itself (see, e.g., Galvin et al., 2003 for a detailed comparison). Therefore, whether the respondents accurately differentiate the correct and incorrect responses of their own or given can be measured in Type-2 SDT observations (Galvin et al., 2003). Uniquely, however, participants are free to report or withhold (pass) their answers, although they are still asked to guess the correct answer even if they prefer to pass the question. Lastly, participants rate how confident they are their responses are correct. As displayed in Figure 1, this variant of SDT assumes that the respondents generate correct and incorrect candidates before reporting (see, e.g., Watkins & Gardiner, 1979 for the generate-recognize model of recall), these candidates distribute normally over a continuum of confidence, and that the participants set a response criterion to decide whether the generated candidate shall be reported or withheld depending on how sure they are to report this candidate (Galvin et al., 2003;

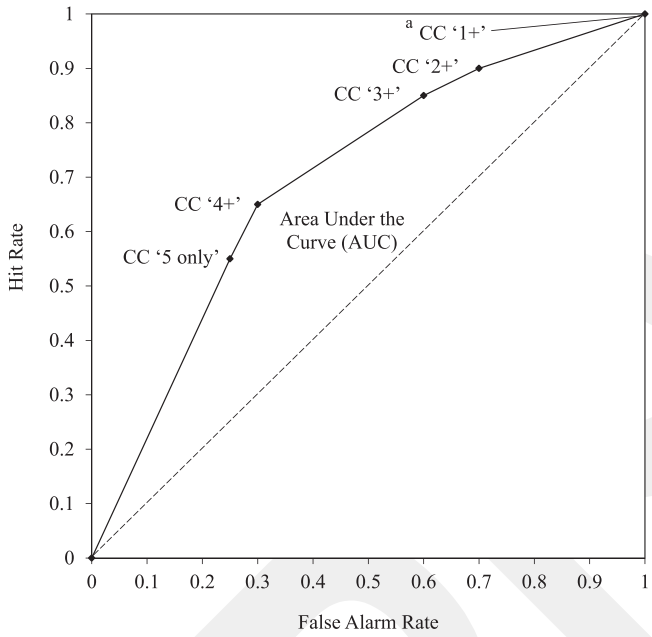

FIGURE 2 The receiver operating characteristics (ROC) curve of a hypothetical case. Monitoring performance (i.e., area under the curve [AUC]) is measured as the area appearing in between the broken (solid) line and the diagonal (dashed) line, where the former is the intersection points of the hit and false alarm rates calculated at each cumulative confidence (CC) level. [a] The intersection at the top right end, for instance, refers to the CC "1+" of hit rate and false alarm rate of this case, which is "1.00" for both rates. Also note that the confidence levels of the students on their responses' correctness range between "1" (not at all confident correct) and "5" (completely confident correct)

Higham, 2002, 2007; also see, e.g., Higham & Higham, 2019 for a detailed summary). Hence, the better-monitoring respondents tend to report mostly correct ones and rate these responses with higher confidence levels, yet they withhold incorrect candidates and rate them with lower confidence levels. The contingency table in Type-2 SDT, therefore, appears as 2(response: report-withhold) × 2(candidate answer: correct-incorrect; Higham & Tam, 2005, 2006). The calculations of hit and false alarm rates eventually allow us to measure *monitoring ability* (e.g., A'), referring to the degree to which participants tend to report correct candidate responses and withhold incorrect ones, and *report bias* (β), the participant's tendency to report their candidate responses regardless of the generated candidates' accuracies (see, e.g., Higham, 2002; Higham & Tam, 2005 for the mathematical formulae).

Figure 2 displays a hypothetical case where the participant's monitoring performance is calculated individually via their hit and false alarm rates measured at each cumulative confidence (CC) level, that is, "1+," "2+," "3+," "4+," and "5" (1 = not at all confidence correct; 5 = completely confident correct). The CCs refer to the responses rated with a particular confidence level and above. The intersection points of all hit rate (HR) and false alarm rates (FARs) at a particular CC are pinned on a scatterplot and they are connected. The trapezoidal area appearing in between the lines connecting each intersection and the diagonal line, where the latter directly connects the intersections of "0" and "1.00" rates and assumes a pure guess (i.e., every HR and FAR are 0.50 at all confidence levels), yield an area under the curve (AUC) score for the given case. In other words, the further away is this trapezoidal curve from this diagonal line (pure guess line) towards higher hit rates and lower false alarm rates, the better the monitoring ability of the participant. Thus, it implies that these respondents tend to give overwhelmingly more correct answers with the "report" option and rate them with higher confidence levels and prefer not to give a response (e.g., "withhold" their potential answers) if they accurately assess their candidate answers are potentially false and now rate these with lower confidence levels (Higham, 2002; Higham & Higham, 2019).

1.2 | Overview of the study

The current study assessed sixth-graders on an 11-item mathematics test that was developed to be equivalent to PISA's mathematics section (see the "materials" section for details). For each answer, the students were asked to decide whether they think they could solve the problem or not (referring to "report" or "pass", respectively), then to make their best guesses even if they wished to pass the question, and lastly to rate their confidence levels on the correctness of their responses (i.e., "How confident are you that your response is correct?"); see Appendix A for an exemplary question item and the instructions involved in the original test booklet. The "report" and "pass" options conventionally refer to the conditions where participants think they wish to provide an answer if they believe they know the correct answer. However, we needed to replace the standardly-used "report" or "pass" options with the following ones in the current study: "I think I can solve the question" and "I don't think I can solve the question," respectively. These phrases were considered to delineate the same options as what it is meant by the "report/pass" option, yet were found much appropriate for the participants (i.e., the sixth-graders whose first language is Turkish) due to the followings. The present testing did not use a list-learning paradigm as used in the conventional procedures such as a recall or a recognition test, and the nature of the participants' native language and their vocabulary level needed us to modify these options with more suitable ones in this language, particularly for what is exactly meant by the "pass" option. Eventually, this whole procedure allowed us to measure the individual AUC scores of the participants. For the score estimations, the participants predicted their prospective test scores just before solving the test (i.e., after briefly looking over the whole questions only yet not starting to solve them already) and immediately after completing the test fully, referring to pre- and posttest estimations, respectively.

We expected that pretest estimations would not be as accurate as posttest estimations simply because the questions are not already solved when the students provide their pretest estimations. However, we expected that the monitoring ability and particularly the posttest score estimations would be well related. That is, those who would be better at the test (i.e., high-scorers) would also be better at monitoring their responses' correctness during solving the test and this better awareness would subsequently yield more accurate immediate posttest estimations than those who would obtain lower scores from the test. Just like what the Dunning–Kruger effect would expect (1999), the low-scoring students would not be able to recognize their responses' correctness as successfully as the high-scoring students during responding so that their posttest estimations consequently be inaccurate as well.

2 | METHOD

2.1 | Participants

The population involved 448 state primary schools in 30 districts of Izmir, Turkey, which covered 1822 classes and 45069 sixth-grade students in total (note that the sixth-grade in the Turkish Education system strictly covers only those aging between 11 and 12). The minimum sample size that meets 99% of confidence level and 2% margin error for this population was calculated as 3809 (Thompson, 2012). The schools varied in terms of their students' socio-economic status (SES; e.g., A, B, and C, where A refers to the highest SES) and students' national exam results obtained in the previous year. A stratified random sampling method that considered these two characteristics proportionately was used to determine which schools were to be reached (Lodico et al., 2006). Hence, the necessary number of schools were selected randomly from the same clusters of SES and exam results. The administrators and the parents were written to get participation consent for the students after the schools were determined. Eventually, 2832 sixth-grade students (1410 male and 1422 female) from 15 elementary schools volunteered to participate in the study. The average student number of the selected schools was 189 (range = 90–239).

2.2 | Materials

An 11-item test was developed to assess the sixth-graders' performance in mathematics. Seven questions were multiple-choice, one question was a true-false, and three questions were short-answer questions in the test.[1] The test measured algebraic statements, area calculations, geometrical objects, and volume calculations that were constructed to be appropriate for the level of sixth-grade (see Appendix A for a question that involved in the test booklet). Item-discrimination difficulties of the test were calculated via classical test and item-response theories (IRTs). The test's mean item difficulty was 0.35 (range = 0.06–0.76), mean discrimination index was 0.46 (range = 0.21–0.65), mean point-biserial correlation was 0.52 (range = 0.33–0.73), and Kuder-Richardson-20 was 0.71. The calculations showed the developed test had high reliability (Gronlund, 1982; Gronlund & Linn, 1990). IRT (two-parameters logistic model) analyses showed the mean of threshold *b* parameter, indexing item difficulties, was found 0.70 (range = −1.71–2.86) and the mean of slope (*a* parameter), which calculated item discrimination, was 1.25 (range = 0.44–3.26). The analyses showed the test had high item discriminations, had high reliability, and the test items were fairly difficult (see T. O. Basokcu & Canpolat, 2018 for the full psychometric properties of the developed test).[2]

2.3 | Procedure

The test was administered to 103 classes of 15 public elementary schools on the same day and time. Before the participants were tested, a series of meetings were held with the members of school administrations and with the teachers regarding the study's implementation procedure. Invigilators were informed in a detailed way on how to administer the test. They were supplied with the instructions list along with a checklist to ease following the testing procedure. The question booklets and optical answer sheets along with the extra copies of the above-mentioned documents were sent back to the schools in sealed parcels through the Ministry of National Education (MNE) officials. Invigilators informed the students again that the participation was completely voluntary. The students who did not wish to participate in the study were asked to stay in the class and to read while the others were being tested and asked to hand in the test empty after the testing was over. Then, the invigilators 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, and asked them not to use any calculator and not to check their course materials such as course book or notes. After the study was completed, invigilators handed in the signature lists, the checklists, and the test materials back to the MNE officers in the schools who then surrendered the collected materials back to the project office in the university.

2.4 | Data analytic plan

2.4.1 | Coding data and calculating the test scores and monitoring performance

The students' pre- and posttest score estimations, their answers given to the test questions and their responses regarding metacognitive monitoring instructions (i.e., whether they thought they could solve the questions and the

[1]The responses given to the short-answer questions were scored just like scoring the answers given to the forced-choice items. In other words, only the correctness of their final responses was scored so that how the students approached the question was not considered. The true-false question item involved three sub-items yet the students were asked to "report" or "pass" the question item as a whole and they rated their confidence level for this question as a single question item. Therefore, report or pass option selected and the confidence rating given was considered same for each of these three true-false statements.

[2]The current study is one of the studies of a three-year longitudinal research project. The project had also an aim to develop a test that assesses the PISA mathematical abilities of sixth-graders. Therefore, the current study utilized this developed test that was shown to be comparable to the PISA test's mathematics section (see T. O. Basokcu & Canpolat, 2018 for complete psychometric properties of the test).

confidence ratings) were coded onto a Statistical Package for the Social Sciences (SPSS) data file. Since the students were provided with optical forms to code their answers, the data transfer was managed by an optical code reader. The remaining data covering demographic questions (name and surname, gender, student number, school name, and class section) that existed on the cover page of the test booklet and the score estimations made by the students on the cover page (pretest estimations) and at the very last page of the booklet (posttest estimations) were coded manually on to the SPSS file by the research assistants. Test scores, monitoring performance (AUC scores), and the score estimations of the students were calculated via the calculate option on SPSS. Before being transferred to the SPSS data file, the AUC scores of the participants were first calculated on Microsoft Excel due to its better convenience; see the "introduction" section and Figure 2 for the details of how the AUC scores are calculated.

2.4.2 | Clustering performance groups

The descriptive statistics showed the participants solved 3.34 questions correctly (SD = 1.58) out of 11 question items on average (30.39%). The students were, then, clustered into three performance groups in terms of their test scores in a way that the means of groups' test scores differed significantly from each other: low-, medium-, and high-scoring groups; $F(2, 2829) = 7629$, $p < .001$,$\eta_p^2 = 0.84$, [0.84, 0.85]. The students who obtained a total score of "2" and below constituted the low-scoring group ($n = 851$; $M = 1.53$, SD = 0.64; 403 male and 448 female), those who obtained a total score of "5" and above composed the high-scoring group ($n = 627$; $M = 5.58$, SD = 0.82; 312 male and 315 female), and those obtained a total score in between these (i.e., 3 and 4) constituted the medium-scoring group ($n = 1354$; $M = 3.44$, SD = 0.5; 695 male and 659 female). The following analyses on monitoring performance and score estimations were conducted between these performance groups.

2.4.3 | Monitoring performance

A one-way analysis of variance (ANOVA) was run to detect whether the performance groups differed in terms of their monitoring abilities. It was expected that the low-performance group would obtain the lowest AUC scores. Since the comparisons were between the group means, a further correlational analysis (Pearson's product-moment correlation) was also calculated between the test scores and AUC scores of the students where the scores are considered individually.

2.4.4 | Score estimations

The predictions that the students were asked to make before responding to the questions (i.e., just after brief browsing of the test booklet) were expected to be not as accurate as of the postdictions (posttest estimations) because a strong monitoring ability at the test was expected to yield subsequently a better or even accurate estimation. Therefore, first, a 3(group: low-, medium-, high-scoring) × 2(score: actual vs. predicted) mixed factorial ANOVA was conducted with the group variable was a between- and the score was a within-subjects factor. Some students did not provide their pretest estimations so that the mean comparisons considered 2770 students in total for the analyses (*N*s in the low-, medium-, and high-scoring groups are: 823, 1331, and 616, respectively).[3] Secondly, a 3(group: low-, medium-, high-scoring) × 2(score: actual vs. estimated) mixed factorial ANOVA was

[3]Only 62 students out of 2832 students (2.2%) did not provide their pretest score estimations, which we considered was a negligible drop in the response rate. Further $\chi 2$ analysis confirmed that the percentages of the drops in the response rates were comparable across the performance groups; $\chi^2(2) = 0.002$, $p > .05$.

conducted where the group variable was a between-subjects factor and the score was a within-subjects factor. Some students did not provide their posttest estimations. Therefore, this analysis involved a total of 1934 students' data (*N*s for low-, medium-, and high-scoring groups are: 533, 940, and 461, respectively).[4]

3 | RESULTS

3.1 | Monitoring performance

The results showed monitoring performance was significantly different between the performance groups (low-, medium-, and high-scoring groups); $F(2, 548) = 25.672$, $p < .001$, $\eta_p^2 = 0.09$, [0.05, 0.12].[5] As expected, post hoc comparisons revealed the low-scoring group obtained significantly lower monitoring performance ($M = 0.15$, SD = 0.29) than the medium- and high-scoring groups ($M = 0.27$, SD = 0.32; $M = 0.42$, SD = 0.32, respectively), where the latter two also significantly differed from each other. Though the participants together yielded a fairly poor monitoring ability ($M = 0.29$, SD = 0.33) just like their test performance (3.4 correct responses out of 11 questions), the results confirmed the hypothesis that monitoring and test performance are highly related. As displayed in Figure 3, metacognitive monitoring ability gradually increased as the number of correct answers increased. Pearson's product-moment correlation analysis also showed the test scores and monitoring performance were positively correlated, $r(551) = .33$, $p < .001$.

3.2 | Score estimations

3.2.1 | Pretest estimations

The results showed a significant group main effect; $F(2, 2767)=410.817$, $p < .001$, $\eta_p^2 = 0.23$, [0.21, 0.25]. Pair-wise comparisons showed the groups differed significantly from each other: low-scoring group obtained the lowest score (i.e., the average of actual and expected scores) ($M = 2.12$) than medium- ($M = 3.29$) and high-scoring groups ($M = 4.73$). Score main effect was also significant; $F(1, 2767)=19.719$, $p < .001$, $\eta_p^2 = 0.01$, [0.003, 0.013]. The average actual score of the participants was significantly lower ($M = 3.23$) than what they predicted their scores would be ($M = 3.52$). Critically, the results showed an interaction effect between group and score factors as well; $F(2, 2767) = 129.946$, $p < .001$, $\eta_p^2 = 0.09$, [0.07, 0.10]. The pretest estimations of the groups differed significantly from each other, and these estimations increased gradually from low- towards high-scoring groups ($M = 2.69$; $M = 3.12$; $M = 3.88$, respectively) just like their actual scores did. However, the low-scoring group "overestimated" their actual scores whereas the medium- and the high-scoring groups "underestimated" their prospective scores. A greater underestimation was observed in high-scoring students than the medium-scoring group; see Figure 4.

[4]Out of 2832 sixth-graders, 898 students (31.7%) skipped giving their posttest score estimations. The frequencies of all students who took the test thereby constituted a particular performance group were found to drop comparably to the frequencies of performance groups where the students provided their posttest estimations; $\chi^2(2) = .0176$, $p > .05$.

[5]Numerous students systematically skipped at least one item empty. Since they were excluded from the AUC calculations, participant attrition appearing when comparing AUC scores may be raised by some readers. Therefore, further analysis was conducted with including all the cases who responded to at least one question and its related question items (i.e., report/pass and confidence) as asked ($N = 2152$). Even with the most liberal comparison of AUC values, the results again showed that the groups differed in terms of their monitoring abilities; $F(2, 2149) = 54.841$, $p < .001$, $\eta_p^2 = 0.05$, [0.03, 0.07]: low-, < medium-, < high-scoring groups; $M = 0.13$ (SD = 0.28); $M = 0.23$ (SD = 0.32); $M = 0.33$, (SD = 0.34), respectively. It should also be noted the sizes of the performance groups (*N*s for low-, medium-, and high-performance groups: 666, 1223, and 587) that involved all students who responded at least one student as instructed dropped to the following sizes when the analysis considered strictly only those who completed the test fully (i.e., responded to 11 questions and their related ratings): 108, 273, and 551, respectively. Pearson's $\chi 2$ test revealed the frequencies of the performance groups when their monitoring performance was compared leniently ($N = 2476$) dropped comparably to the frequencies of the groups in which the students provided all the necessary data for all of the questions ($N = 551$); $\chi^2(2) = 2.095$, $p > .05$, suggesting the missing cases were not cumulated in a particular performance group (or groups).

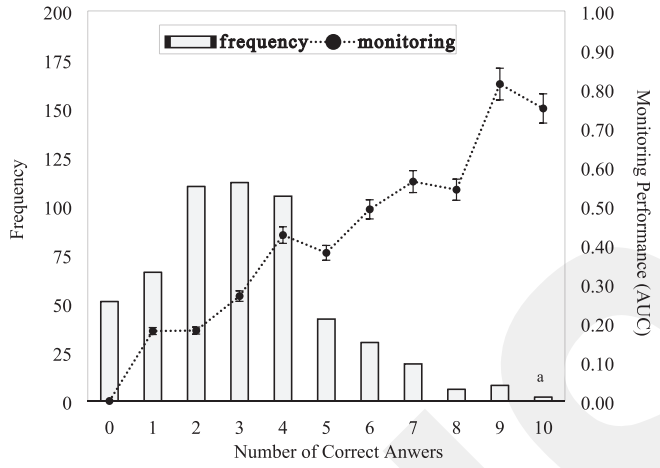

FIGURE 3 Student frequencies in terms of their number of correct answers and their monitoring performance. The frequencies are shown only for those who completed each item fully as instructed in the test booklet ($N = 551$) thereby allowed us to measure their area under the curve (AUC) scores observed for the whole test. [a] The number of correct answers is displayed up to 10 only since "none" of the students who fully completed the test as instructed solved all of the questions correctly (i.e., 11)

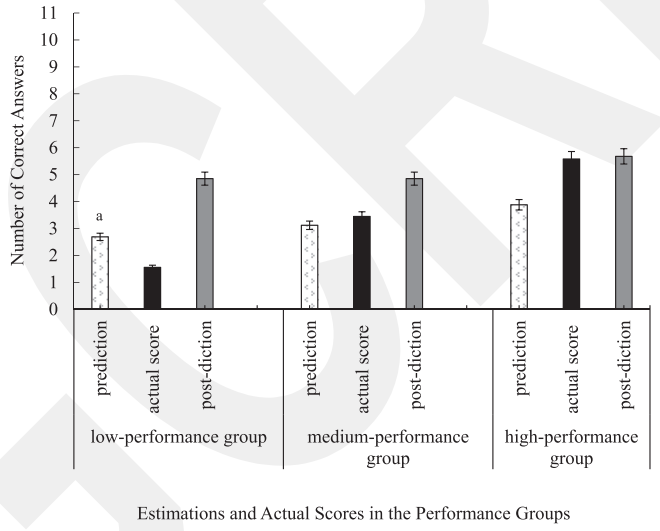

FIGURE 4 Means of pretest estimations, actual scores, and posttest estimations in each of the performance groups. The graph displays the pretest estimations (i.e., predictions), actual scores, and posttest estimations (i.e., postdictions) in each performance clusters which were grouped in terms of the students' test scores: low-, medium-, and high-performance groups. [a] Standard errors are displayed with error bars

3.2.2 | Posttest estimations

The results showed a significant group main effect; $F(2,1931)=413.366$, $p < .001$, $\eta_p^2 = 0.30$, [0.27, 0.33]. Post hoc comparisons revealed the following: As the test performance was gradually better in the group (i.e., low-, < medium-, < high-scoring group), the postdictions (i.e., posttest estimations) also increased significantly ($M = 3.21$; $M = 4.37$; $M = 5.66$, respectively). Score main effect was also found significant; $F(1,1931)=734.111$, $p < .001$, $\eta_p^2 = 0.28$, [0.25, 0.30]. The groups together overestimated what their actual scores would be ($M = 5.27$ and $M = 3.55$,

respectively). Moreover, the results showed an interaction effect between group and score factors; $F(2, 1931) = 185.620$, $p < .001$, $\eta_p^2 = 0.16$, [0.14, 0.19]. Just like low-, medium-, and high-scoring groups differed significantly from each other in terms of their actual scores, their posttest estimations were again significantly different from each other ($M = 4.85$; $M = 5.29$; $M = 5.68$, respectively). Unlike low- and medium-scoring groups who "overestimated" their performance, the means of high-scoring group's actual scores and postdictions were *comparable* (i.e., accurate); see Figure 4. As expected, this pattern confirmed that when the students were better at the test and likewise retained a better monitoring ability during testing, their immediate posttest estimations were accurate as well.

As displayed in Figure 4, the low-scoring group had inflated score estimations about their actual performance no matter they made such estimations before or after the questions were solved. This group's monitoring performance at testing was also the poorest one among all performance groups. Conversely, however, the students who obtained better scores than the remaining groups tended to discredit their prospective performance at pretest estimations yet they (particularly the high-scoring students) seemed to fix their inaccurate pretest estimations with accurate immediate posttest estimations.

4 | DISCUSSION

In the present study, we utilized the measurement strategy of the Type-2 SDT approach (e.g., Higham, 2002) and collected score estimations of a large group of sixth-graders at a mathematics test that was developed to be equivalent to the PISA mathematics section. The findings showed that the monitoring ability and test performance are highly related in the PISA mathematics context in the way that the students who obtained higher test scores concomitantly yielded better monitoring of their responses than those who obtained lower scores from the test.

Expecting better monitoring ability among high-scoring students rather than low-scoring ones is based on the Type-2 SDT's predictions. For instance, the high-scoring students are likely to be better learners of the subject and so the presumed distribution of their correctly generated candidate answers should be further apart from the distribution of their incorrectly generated ones compared to the divergence that the low-scoring students may yield (see, e.g., Figure 1). This advantage can consequently be conducive to a better recognition ability for the high-scoring students when they detect their correct candidates from among the incorrect ones mainly because the distributions are presumably further apart. However, a possible advantage can exist providing that, for instance, the groups retain comparable report biases (i.e., where they set their response criteria). Further results showed that each group was fairly liberal in their tendencies to report, which was indexed by the β parameter (report bias; e.g., Higham, 2007). However, the high-scoring group held a more liberal criterion to report their answer (i.e., withheld fewer answers; $M = -0.48$, $SD = .74$) than both low- and medium-scoring groups ($M = -0.22$, $SD = 0.90$ and $M = -0.38$, $SD = 0.76$, respectively); $F(2, 548)=3.645$, $p > .027$, $\eta_p^2 = 0.013$, [0.001, 0.031]. Although the high-scoring group was more liberal in responding which thereby might have yielded higher false alarms (i.e., falsely reported incorrect answers), they were still better at monitoring their responses' correctness than the other groups. Therefore, this result renders our finding even more robust in the way that being a high-scorer is closely linked with being better at monitoring when responding.

The close linkage between academic performance and monitoring ability shown in the current study converges with the previous findings. For instance, it has been reported that enhancing the metacognitive abilities of the students has the potential to yield better academic performance (e.g., Leal, 1987; Mayer, 2001; Swanson, 1990). The current evidence, however, may still be retained with caution. For instance, one may give correct responses yet this does not always guarantee or necessitate a good monitoring ability because the cognitive and metacognitive processes are, in fact, different processes even though they can presumably be well related (Flavell, 1979; Nelson & Narens, 1994). For instance, as was reported by Guzel and Higham (2013), generation of correct answers and recognition of them at a later stage, referring to early- and late-selection processes respectively, may result in a dissociative pattern at retrieval depending on the features of the study material (e.g., if the study lists involve highly-related to-be-remembered items). In other words, several factors may adversely affect the discrimination ability of correct and incorrect answers (metacognition) such as how probable the incorrect answers, lures, or distractors are

(Benjamin & Bawa, 2004) while they facilitate cognitive performance (cognition) such as categorical relatedness between target items (e.g., Guynn & McDaniel, 1999; Guynn et al., 2014; Hunt et al., 2016).

Additionally, the existence of making a pretest estimation was not manipulated as a variable since all students made pretest estimations in the current study. Providing that making pretest estimations somehow affect the students' subsequent test performance by such as changing their response bias, further research is needed to reveal how the groups vary in terms of whether they make pretest estimations or not. In other words, whether a preset mental readiness shifts the report criterion. We expect that an "overestimation" tendency that may be caused by expecting to get better scores from the test may lead the respondents to allocate a relatively liberal report criterion. This loosened criterion would thereby end up with a higher false alarm rate (i.e., reporting a higher number of incorrect responses). Although previous research has shown that participants may shift their response criteria even on an item-by-item basis rather than retaining a stable one throughout the testing session and revealed various factors that can shift the response criterion, such as the external influences (e.g., task difficulty, immediate feedback given on the response correctness, type of stimulus in the to-be-remembered lists, etc.; see, e.g., Hockley, 2011 for an extensive review), further research is needed to better clarify whether making pretest estimations affect the response criterion towards a lenient one and increase the confidence ratings.

The inflated score estimations observed among the low-scorers in this study remind us that the Dunning–Kruger effect must be in harness (Kruger & Dunning, 1999). Although it is beyond the scope of the present study to evaluate this effect, we believe that the low-performers should *not* be considered as those who have no awareness at all on how they behave particularly when responding due to two main reasons. First, it is statistically a higher probability to observe a clearer overestimation when one's actual score is low than one having a higher test score providing that low- and high-scorers diverge upwards from their actual scores with a comparable amount. Second, the further analysis on monitoring abilities showed that the low- and medium-scoring groups had comparable confidence ratings ($M = 2.92$, SD = 1.07; $M = 3.15$, SD = 0.94, respectively) and these groups rated their responses with significantly *lower* confidence ratings than the high-scoring group ($M = 3.46$, SD = 0.85); $F(2, 548) = 11.613$, $p < .001$, $\eta_p^2 = 0.04$, [0.02, 0.07]. In other words, the low-scorers solved the lowest number of questions in the test and they overestimated their actual scores no matter it was before or immediately after solving the questions, yet they were still *not* as much confident about the correctness of their responses as the high-scoring students.

As expected, the present results on the posttest estimations rather than pretest estimations were found to be better related to the monitoring abilities. For instance, the high-scoring group seemed to calibrate their inaccurate pretest estimations with accurate postdictions most likely due to a better monitoring performance that they retained during testing. In other words, once the students have well-functioning metacognitive awareness of whether their answers are correct or incorrect, then it is reasonable to expect that they would accurately estimate what their overall actual scores will be. We believe that the accurate postdictions observed among the high scorers were also affected by the detailed instructions that were available when responding, such as choosing to report or pass the items, giving an answer even if it was passed, and later giving a confidence rating for each of their responses' accuracies (i.e., a deeper evaluation of the question item by the students). This availability, however, seemed to assist primarily the high-scoring students although the same instructions were already available for all of the students. Therefore, future studies on students' score estimations may consider varying the existence of Type-2 SDT testing procedure to reveal whether the participants who estimate their scores without this testing procedure (i.e., when Type-2 instructions are excluded during responding) yield comparable results as those who estimate their scores along with these instructions. Such experimental design can better clarify whether the high-scorers' posttest estimations are accurate since they have already better metacognitive monitoring abilities than the low-scorers no matter the students are tested with Type-2 monitoring instructions or not. There is, however, recent evidence that metacognitive instructions have a facilitative effect on monitoring and academic achievement (e.g., Händel, de Bruin, et al., 2020; Händel, Harder, et al., 2020).

Despite the above-mentioned suggestions for future research, the current study still appears as a unique one by revealing that the metacognitive monitoring ability and score estimation (i.e., posttest estimations) are highly related. That is, a better monitoring ability expectedly ends up with an accurate posttest estimation, unlike a poor monitoring ability. As

was reported previously (e.g., Boekaerts & Rozendaal, 2010; Erickson & Heit, 2015; Hines et al., 2009; Jacobson, 1990; Prohaska, 1994), posttest estimations have been considered as a judgment that necessitates a certain degree of awareness whether one has an accurate assessment on their overall test performance. Therefore, we propose that should one prefers a rough yet instant assessment of the respondents' monitoring abilities, even immediate posttest estimations alone can be preferred as a fairly informative judgment to measure a learner's higher-order ability (see, e.g., Jacobson, 1990 for a discussion on the subject). Even so, score estimation would appear as a limited parameter as compared to the Type-2 SDT calculations simply because the latter allows us to measure various metacognitive abilities accurately via its specific testing procedure, such as the regulation of memory accuracy, metacognitive control, and monitoring, as well as report bias (see, e.g., Higham, 2007). Despite this limitation, we suggest that assessing students' higher-order abilities beside—or even beyond—counting the responses' correctness is a critical assessment of the learners no matter a test entails a detailed and accurate measurement of various metacognitive abilities and behaviors or a rough yet relatively more instant measurement of monitoring ability.

Depending on the evidence gathered in this study, we propose that the mathematics teachers may readily consider particularly the posttest score estimation method as a further assessment that is annexed to their tests if they consider AUC calculation is relatively more cumbersome. Again, if AUC calculations are considered as a less practical one as compared to score estimations, even asking students whether they believe they can solve the question items or not just before answering the items alone (i.e., providing a report/pass option and asking to guess even if it is passed) may also inform the teachers in mathematics education of their students' responding styles, such as how they set, maintain, or alter their report criteria (e.g., strictly or leniently) in terms of, for instance, item difficulty. We believe that even this application alone (i.e., contemplating on the accuracies of own responses) can function as a workable educational activity besides being an in-class assessment for the teachers by, for instance, allowing the students to better realize how they approach the questions and evaluate their answers during responding. Whether one prefers using score estimations and/or AUC calculations, various instructional adjustments can also be generated for the students who have good or poor monitoring abilities accordingly. In this vein, the findings we gathered in this study and the suggestions emerging from these findings seem to be in line with the education literature. For instance, it has been reported that the students who have poor mathematical performance are also more inaccurate in their postdictions (i.e., posttest estimations) than those who have better mathematics abilities (Desoete et al., 2019; Donker et al., 2014). Better monitoring and self-regulating abilities of the students have also been shown to facilitate solving these questions correctly (Garofalo & Lester, 1985; Lester, 1982; Silver, 1982). Additionally, Garofalo and Lester (1985), for instance, have also underlined that a good metacognitive monitoring ability is necessary to learn particularly more complex mathematics topics (also see, Schoenfeld, 1983). In short, success in mathematics seems to be positively affected by how well the students are aware of their knowledge and regulate this knowledge. Therefore, teachers may wish to consider rearranging their instructions for those who have better monitoring abilities yet low scores towards enhancing these students' knowledge about the topic more (i.e., their cognitions). On the other hand, for those students who are good learners yet somehow have low performance on monitoring their responses' accuracies, teachers may consider providing them with, for instance, additional feedbacks about their awareness of their responses. It should, however, be kept in mind that having better monitoring ability and being a good learner (e.g., better scorer) at the same time are well-linked and so is more likely to observe. Therefore, we believe that teachers should suspect that such observation may readily emerge due to the possibility that the assessment method such as the test might have involved a proportion of questions that are solved correctly even by pure guesses; in other words, with not much of a clear awareness on the responses' correctness at all. Hence, this detection may give a reason to teachers -or the test developers- that the assessment tool needs to be readdressed and so modified in a way that the questions can mainly be answered correctly without the help of any guesses.

Lastly, the current study was restricted to the mathematical abilities of the sixth-graders only. Therefore, future studies may consider assessing the monitoring performance of learners among various other grades and in other domains of the PISA test (i.e., science and reading), in any other assessment types, or even between cultures. Should the prospective studies, teachers, or even the future PISA administrations consider the assessment procedure of Type-2 SDT or the posttest estimations are feasible and fruitful ones, they may also investigate whether monitoring ability and estimation

accuracies change between different educational formats, such as between online, face-to-face, and hybrid contexts as well.

ACKNOWLEDGMENT

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

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

Abdi, H. (2010). Signal detection theory. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (pp. 407–410). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01364-6

Anthony, J. S., Clayton, K. E., & Zusho, A. (2013). An investigation of students' self-regulated learning strategies: Students' qualitative and quantitative accounts of their learning strategies. *Journal of Cognitive Education and Psychology*, *12*(3), 359–373. https://doi.org/10.1891/1945-8959.12.3.359

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education*, *16*(3), 363–383. https://doi.org/10.1007/BF03173188

Artelt, C., & Schneider, W. (2015). Cross-country generalizability of the role of metacognitive knowledge in students' strategy use and reading competence. *Teachers College Record*, *117*(1), 1–32.

Artz, A. F., & Armour-Thomas, E. (1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. *Cognition and Instruction*, *9*(2), 137–175. https://doi.org/10.1207/s1532690xci0902_3

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/a0033268

Basokcu, O. T., & Guzel, M. A. (2020). Metacognitive monitoring and mathematical abilities: Cognitive diagnostic model and signal detection theory approach. *TED Eğitim ve Bilim (Education & Science)*, *46*(205), 221–238. https://doi.org/10.15390/EB.2020.7991

Basokcu, T. O., & Canpolat, A. (2018). *Ankor test deseninde bilissel tanı modeli ortuk yetenek siniflari ile test esitleme calismasi (A study on test equation with latent classes following cognitive diagnostic model as an anchor test desing)*. 6th International Congress on Measurement and Evaluation in Education and Psychology, 384–388.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, *51*(2), 159–172. https://doi.org/10.1016/j.jml.2004.04.001

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction*, *20*(5), 372–382. https://doi.org/10.1016/j.learninstruc.2009.03.002

Cherry, B., Ordóñez, L. D., & Gilliland, S. W. (2003). Grade expectations: The effects of expectations on fairness and satisfaction perceptions. *Journal of Behavioral Decision Making*, *16*(5), 375–395. https://doi.org/10.1002/bdm.452

Callan, G. L., Marchant, G. J., Finch, W. H., & Flegge, L. (2017). Student and school SES, gender, strategy use, and achievement. *Psychology in the Schools*, *54*(9), 1106–1122.

Donker, A. S., de Boer, H., Kostons, D., Dignath van Ewijk, C. C., & van der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, *11*, 1–26. https://doi.org/10.1016/j.edurev.2013.11.002

Erickson, S., & Heit, E. (2015). Metacognition and confidence: Comparing math to other academic subjects. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00742

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

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. https://doi.org/10.3758/BF03196546

Gamazo, A., & Martínez-Abad, F. (2020). An exploration of factors linked to academic performance in PISA 2018 through data mining techniques. *Frontiers in Psychology*, *11*, 575167. https://doi.org/10.3389/fpsyg.2020.575167

Garofalo, J., & Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education*, *16*(3), 163–176. https://doi.org/10.2307/748391

Green, D. G., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley. https://psycnet.apa.org/record/1967-02286-000

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

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching*. Macmillan.
Guynn, M. J., & McDaniel, M. A. (1999). Generate—Sometimes recognize, sometimes not. *Journal of Memory and Language*, *41*(3), 398–415. https://doi.org/10.1006/jmla.1999.2652
Guynn, M. J., McDaniel, M. A., Strosser, G. L., Ramirez, J. M., Castleberry, E. H., & Arnett, K. H. (2014). Relational and item-specific influences on generate–recognize processes in recall. *Memory & Cognition*, *42*(2), 198–211. https://doi.org/10.3758/s13421-013-0341-6
Guzel, M. A., & Higham, P. A. (2013). Dissociating early- and late-selection processes in recall: The mixed blessing of categorized study lists. *Memory & Cognition*, *41*(5), 683–697. https://doi.org/10.3758/s13421-012-0292-3
Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal for Educational Research Online/Journal Für Bildungsforschung Online*, *5*, 162–168.
Händel, M., de Bruin, A. B. H., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning*, *15*(1), 51–75. https://doi.org/10.1007/s11409-020-09220-0
Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction*, *65*, 101245. https://doi.org/10.1016/j.learninstruc.2019.101245
Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, *30*(1), 67–80. https://doi.org/10.3758/BF03195266
Higham, P. A. (2007). No special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, *136*(1), 1–22. https://doi.org/10.1037/0096-3445.136.1.1
Higham, P. A., & Arnold, M. M. (2007). How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. *European Journal of Cognitive Psychology*, *19*(4–5), 718–742. https://doi.org/10.1080/09541440701326121
Higham, P. A., & Higham, D. P. (2019). New improved gamma: Enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves. *Behavior Research Methods*, *51*(1), 108–125. https://doi.org/10.3758/s13428-018-1125-5
Higham, P. A., & Leboe, J. P. (2011). Constructions of remembering and metacognition. In P. A. Higham, & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honour of Bruce Whittlesea*. Palgrave Macmillan. https://doi.org/10.1057/9780230305281
Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory and Language*, *52*(4), 595–617. https://doi.org/10.1016/j.jml.2005.01.015
Higham, P. A., & Tam, H. (2006). Release from generation failure: The role of study list structure. *Memory & Cognition*, *34*(1), 148–157. https://doi.org/10.3758/BF03193394
Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. *Psychology and Aging*, *24*(2), 462–475. https://doi.org/10.1037/a0014417
Hockley, W. E. (2011). Criterion changes: How flexible are recognition decision processes? In P. A. Higham, & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honour of Bruce Whittlesea* (pp. 155–166). Palgrave Macmillan. https://doi.org/10.1057/9780230305281_12
Hunt, R. R., Smith, R. E., & Toth, J. P. (2016). Category cued recall evokes a generate-recognize retrieval process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(3), 339–350. https://doi.org/10.1037/xlm0000136
Jacobson, J. M. (1990). Congruence of pretest predictions and posttest estimations with grades on short answer and essay tests. *Educational Research Quarterly*, *14*(2), 41–47.
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
Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517. https://doi.org/10.1037/0033-295X.103.3.490
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121
Lai, E. R. (2011). Metacognition: A literature review research report. Pearson.
Leal, L. (1987). Investigation of the relation between metamemory and university students' examination performance. *Journal of Educational Psychology*, *79*(1), 35–40. https://doi.org/10.1037/0022-0663.79.1.35
Lester, F. K. (1982). Building bridges between psychological and mathematics education research on problem solving. In F. K. Lester, & J. Garofalo (Eds.), *Mathematical problem solving* (pp. 55–85). Franklin Institute Press.
Lie, S., Taylor, A., & Harmon, M. (1996). Scoring techniques and criteria. In M. O. Martin, & D. K. Kelly (Eds.), *Third international athematics and science study (TIMSS) technical report, volume I: Design and development*. Chesnut Hill.
Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2006). *Methods in educational research: From theory to practice*. John Wiley.
Maag Merki, K., Ramseier, E., & Karlen, Y. (2013). Reliability and validity analyses of a newly developed test to assess learning strategy knowledge. *Journal of Cognitive Education and Psychology*, *12*(3), 391–408. https://doi.org/10.1891/1945-8959.12.3.391

Magnus, J. R., & Peresetsky, A. A. (2018). Grade expectations: Rationality and overconfidence. *Frontiers in Psychology*, *8*, 2346. https://doi.org/10.3389/fpsyg.2017.02346

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance SDT model. In S. M. Fleming, & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Springer-Verlag. https://doi.org/10.1007/978-3-642-45190-4_3

Mayer, R. E. (2001). Cognitive, metacognitive, and motivational aspects of problem solving. In H. J. Hartman (Ed.), *Metacognition in learning and instruction* (vol 19, pp. 87–101). Neuropsychology and cognition. Springer. https://doi.org/10.1007/978-94-017-2243-8_5

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review*, *2*(2), 100–110. https://doi.org/10.1207/s15327957pspr0202_3

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing*. (pp. 1–25). The MIT Press.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

OECD. (2013). *PISA 2012 assessment and analytical framework*. OECD. https://doi.org/10.1787/9789264190511-en

O'Neil, Jr., H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education*, *11*(4), 331–351. https://doi.org/10.1207/s15324818ame1104_3

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

Rasooli, A., Zandi, H., & DeLuca, C. (2018). Re-conceptualizing classroom assessment fairness: A systematic meta-ethnography of assessment literature and beyond. *Studies in Educational Evaluation*, *56*, 164–181. https://doi.org/10.1016/j.stueduc.2017.12.008

Säälik, Ü., Nissinen, K., & Malin, A. (2015). Learning strategies explaining differences in reading proficiency. Findings of Nordic and Baltic countries in PISA 2009. *Learning and Individual Differences*, *42*, 36–43. https://doi.org/10.1016/j.lindif.2015.08.025

Schoenfeld, A. H. A. (1983). Episodes and executive decisions in mathematical problem solving. In R. A. Lesh, & M. Landau (Eds.), *Acquisition of mathematics concepts and processes*. Academic Press.

Silver, E. A. (1982). Knowledge organization and mathematical problem solving. In F. K. Lester, & G. J. (Eds.), *Mathematical problem solving* (pp. 15–25). Franklin Institute Press.

Stacey, K., & Turner, R. (2015). Assessing mathematical literacy. In K. Stacey, & R. Turner (Eds.), *Assessing mathematical literacy: The PISA experience*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10121-7

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

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of Educational Psychology*, *82*(2), 306–314. https://doi.org/10.1037/0022-0663.82.2.306

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

Thompson, S. K. (2012). *Sampling*. Wiley.

Watkins, M. J., & Gardiner, J. M. (1979). An appreciation of generate-recognize theory of recall. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 687–704. https://doi.org/10.1016/S0022-5371(79)90397-9

Weisskirch, R. S. (2018). Grit, self-esteem, learning strategies and attitudes and estimated and achieved course grades among college students. *Current Psychology*, *37*(1), 21–27. https://doi.org/10.1007/s12144-016-9485-4

White, B. Y., & Frederiksen, J. R. (2009). Inquiry, modelling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, *16*(1), 3–118. https://doi.org/10.1207/s1532690xci1601_2

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence. *Zeitschrift Für Psychologie/Journal of Psychology*, *216*(2), 102–110. https://doi.org/10.1027/0044-3409.216.2.102

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman, & D. H. Schuck (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 1–37). Lawrence Erlbaum Associates Publishers.

How to cite this article: Başokçu, T. O., & Güzel, M. A. (2022). Beyond counting the correct responses: Metacognitive monitoring and score estimations in mathematics. *Psychology in the Schools*, 59, 1105–1121. https://doi.org/10.1002/pits.22665

APPENDIX A

See Figure A1.

Above is a rectangular prism taken from the Minecraft game that is composed of same-size and color cubes inserted one another.[6]

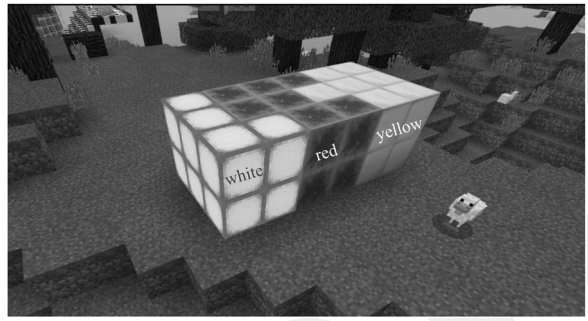

FIGURE A1 An exemplary question item involved in the test booklet

Question 7.	*Please code the answer you think is correct on the optical form as well.*	
I think	'I can solve the question below'	○
	'I cannot solve the question below'	○→Even though you indicate this, please choose the best alternative below that you think would be the correct answer.

What is the ratio of the white cubes' volume to the yellow cubes' one?

How confident are you that your response is correct? Please indicate by choosing a number given below.				
Not at all confident correct	Not confident correct	A bit confident correct	Fairly confident correct	Completely confident correct
1	2	3	4	5
○	○	○	○	○

[6]The test booklet was printed in color. The color names tagged on the cubes are for the display purpose only.