Confidence in performance judgment accuracy: the unskilled and unaware effect revisited

Marion Händel, Markus Dresel

Angaben zur Veröffentlichung / Publication details:

Händel, Marion, and Markus Dresel. 2018. "Confidence in performance judgment accuracy: the unskilled and unaware effect revisited." *Metacognition and Learning* 13 (3): 265–85. https://doi.org/10.1007/s11409-018-9185-6.

Confidence in performance judgment accuracy: the unskilled and unaware effect revisited

Marion Händel[1] · Markus Dresel[2]

Abstract
Since its introduction in the late 1990s, the unskilled and unaware effect motivated several further studies. As it stands, low-performing students are assumed to provide inaccurate and overconfident performance judgments. However, as research with second-order judgments (SOJs) indicates, they apparently have some metacognitive awareness of this. The current study with 266 undergraduate students aimed to provide in-depth insights into both the reasons for (in)accurate performance judgments and the appropriateness of SOJs. We implemented a general linear mixed model (GLMM) approach to study item-specific performance judgments in the domain of mathematics at the person and item level. The analyses replicated the well-known effects. However, the GLMM analyses revealed that low-performing students' lower confidence apparently did not indicate subjective awareness, given that these students made inappropriate SOJs (lower confidence in accurate than in inaccurate judgments). In addition, students' self-generated explanations for their judgements indicated that low-performing students have difficulties recognizing that they possess topic knowledge to solve an item, whereas high-performing students struggle with admitting that they do not know the answer to a question. In sum, our results indicate that students at all performance levels have some metacognitive weaknesses, which, however, occur subject to different judgment accuracy.

Keywords Metacognitive judgments · Performance level · Accuracy · Item-specific judgments · Second-order judgments

Are unskilled students unaware of their low performance or are they—at least slightly—aware of it? Research in the area of metacognitive judgments addresses that question, whose answer can have significant consequences for students' subsequent learning processes. From a self-regulated perspective, performance judgments affect further learning engagement and learning

Marion Händel
marion.haendel@fau.de

[1] Department of Psychology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

[2] Department of Psychology, University of Augsburg, Augsburg, Germany

behavior. For example, a student who believes to already possess elaborate knowledge about a specific topic will not undertake further efforts but instead discontinue studying the topic. In contrast, a student who notices knowledge gaps or difficulties in understanding would ideally invest more study time or would change learning strategies in order to understand and retain the respective contents. Especially low-performing students, who need to invest more effort and engage in learning activities, are less likely to do so if they think they are already well prepared (in other words, if they make overconfident judgments). This, in consequence, can negatively influence future performance and result in poorer grades.

Metacognitive judgments are described in models of metacognition and self-regulated learning (Efklides 2011; Nelson and Narens 1990). For instance, metacognitive judgments are explicitly integrated into the metamemory model by Nelson and Narens (1990), which defines metacognition as the interplay between two levels of information processing (the object-level and the meta-level). The two levels are presumed to interact with each other via monitoring and control processes. Metacognitive judgments are part of the monitoring process, which informs the meta-level about the object level. If deficits are detected, regulation activities are undertaken to modify the object level.

The variety of metacognitive judgments can be distinguished according to the learning/achievement/retention phase they are provided in (Nelson and Narens 1990; Schraw 2009). Whereas predictions are judgments made before the performance phase, postdictions, which are the focus of the current study, are made afterwards. Furthermore, judgments can be made regarding each test item (local or item-specific judgments) or only once regarding a whole test (global judgments). A further distinction of metacognitive judgments that is relevant for our work are second-order judgments (SOJs), established by Dunlosky et al. (2005). SOJs are judgments that assess confidence in previously made performance judgments.

Theoretical assumptions and empirical approaches have been applied to understand whether—and if so, why—low-performing students provide inaccurate and overconfident performance judgments. Two effects studied in the metacognitive judgment literature considerably influenced our research. On the one hand, the unskilled and unware effect, which was postulated by Kruger and Dunning (1999), describes that low-performing students have difficulties in providing accurate performance judgments and overestimate their achievement. In consequence, the authors argued that unskilled students are metacognitively unaware. On the other hand, there is the unskilled but subjectively aware effect (see Miller and Geraci 2011), which questions whether low-performing students are in fact unaware. SOJs revealed that low-performing students are less confident in their (inaccurate) performance judgments than high-performing students. However, the question remains whether low-performing students' lower confidence in performance judgment accuracy actually resembles metacognitive awareness or, for example, whether SOJs are instead provided by default (Fritzsche et al. 2018).

Hence, it seems necessary to generate sophisticated knowledge about the possible mechanisms that lead to inaccurate judgments and respective confidence in judgment accuracy. Based on the recommendation by Dunlosky and Lipko (2007) to use local judgments in order to avoid a mismatch between the grain size of judgments (usually global) and tested information (concerning specific concepts), we will elaborate below how an item-specific analysis can help clarify the meaning of these two different judgments. We will illustrate an empirical procedure to provide insights into which judgments resemble metacognitive awareness or not.

Theoretical background

Unskilled and unaware

The unskilled and unaware effect appeared across several (classroom) studies for different domains and types of judgments: Low-performing students overestimated their performance level, and high-performing students underestimated their performance level with judgments closer to actual performance (Bol and Hacker 2001; Kruger and Dunning 1999; Nietfeld et al. 2005). Overall, although results might differ in dependence of domain and not might be consistent for high-performing students (see Erickson and Heit 2015, who revealed a differntial pattern in mathematics compared to in biology and literature), low-performing students were consistently found overconfident across studies.

Why do students judge their performances inaccurately?

The debut of the unskilled and unaware effect resulted in a controversial debate about its origins and triggering conditions. On the one hand, it is expected that low-performing students are doubly "cursed" (Kruger and Dunning 1999): Low performers not only have no topic knowledge but also do not seem to acknowledge it, indicating low metacognitive awareness. A contrasting explanation of this effect refers to statistical artefacts (Burson et al. 2006; Krueger and Mueller 2002). According to this assumption, students of all performance levels are prone to errors and might be biased upwards, which automatically leads to judgments that are more accurate for high-performing students. However, in a series of studies considering effects such as regression to the mean, low-performing students also provided biased and overconfident judgments (Ehrlinger et al. 2008).

Different empirical approaches have been undertaken to investigate how students generate metacognitive judgments. In general, both personal and environmental variables might influence judgment accuracy (de Bruin et al. 2017; Dinsmore and Parkinson 2013). At the personal level, individual differences regarding test performance, self-concept, or personality are the focus of current research (Buratti et al. 2013; Kröner and Biermann 2007; Schaefer et al. 2004). Other studies take into account task difficulty or test characteristics as environmental factors to explain judgment accuracy (Dutke et al. 2010).

According to the cue utilization approach by Koriat (1997), judgment accuracy depends on the availability of diagnostic cues that either relate to knowledge the student has about her/his abilities and knowledge (information-based cue) or to test-specific characteristics such as perceived item difficulty (experienced-based cue) (Koriat et al. 2008). One means of clarifying the reasons underlying students' performance judgments is to ask students directly about why they provide the respective performance judgment. Within the last couple of years, a number of studies were published that implemented such procedures via closed-ended questionnaires (Bol et al. 2005; Hacker et al. 2008) or open-ended questions (Bol and Hacker 2012; Bol et al. 2010; Dinsmore and Parkinson 2013; Hacker et al. 2008; Thiede et al. 2010). The latter studies with open-ended questions coded students' answers with regard to personal or environmental factors that were derived according to the explanations they most frequently offered. For example, students in a research methods class explained their judgments as a consequence of prior knowledge, guessing, or test characteristics (Dinsmore and Parkinson 2013). In the studies by Bol, Hacker, and colleagues, students reported, inter alia, study time, past experiences, or confidence as possible factors influencing their judgment accuracy (Bol and Hacker

2012; Bol et al. 2010; Hacker et al. 2008). In addition, the work by Thiede et al. (2010) studied explanations regarding a text comprehension task with reading ability as a performance criterion. On a descriptive level, average readers more often provided explanations about prior knowledge, memory, or comprehension than at-risk readers did.

Taken together, previous research studying people's reasons for their judgments has revealed substantial information about how they arrive at metacognitive judgments. That research, however, restricted itself to the predominant use of global judgments, where students make only one single judgment about the number or percentage of correctly solved items. We argue that item-specific judgments should be investigated in order to gain a more detailed picture of reasons for metacognitive judgments. First, the mechanisms for judging personal performance might differ between item-specific and global judgments. When asked about their reasons for a global judgment, students need to make a tradeoff between the reasons or average them across the performance test. For example, if students explain their global judgment on the basis of prior knowledge, this might not apply to all items. On the contrary, prior knowledge might be high for some items but low for other items. Second, global judgments might be driven more by personality factors than local judgments might be (Gigerenzer et al. 1991). Consideration of item-specific judgments would enable the reasons to be examined with regard to the fit between judgment and item performance. That is, a correctly solved item can be accurately judged as correct (hit) or mistakenly judged as incorrect (miss). Similarly, an incorrectly solved item can be recognized as such (correct rejection) or can mistakenly be assumed to have been correctly solved (false alarm). These four categories[1] known from signal detection theory (Green and Swets 1966) allow for investigating how students explain their judgments in dependence of actual and judged item correctness. For example, following the cue-utilization approach, students are expected to explain their judgment via prior knowledge or high topic knowledge if they rightly think that their answer to an item is correct because of information-based cues. Guessing, in contrast, might instead be based on an experience-based cue. We expect that students who are metacognitively aware differ in their explanations for their judgments, depending on whether or not they think they knew the correct answer and if this belief is actually true. Transferring previous results (Thiede et al. 2010) to item-specific judgments, we expect, for instance, that judgments concerning hits are presumably explained via high topic knowledge, whereas false alarms might instead be explained via guessing.

Unskilled but subjectively aware

Current research using SOJs challenges the unskilled and unaware effect. Miller and Geraci (2011) argued that if low-performing students are blissfully incompetent, they should not only provide overly high performance judgments but should also be overconfident about the accuracy of their ratings. Their research, however, indicates that low-performing students "only" provide overly high performance judgments and make moderate SOJs, which are lower than those of students with higher performance levels (Miller and Geraci 2011). The authors hence conclude that low-performing students—despite being unaware of their lacking knowledge—seem aware of their metacognitive deficit. Current research tapping into this methodology of asking students for performance judgments and for their confidence in judgment accuracy replicated the results and transferred them from predictions before exam completion to postdictions after exam

[1] Another classification of these four types of judgments is true positive (hit), true negative (correct rejection), false positive (false alarm), and false negative (miss); see Egan (1975); Schraw et al. (2013).

completion as well as from global judgments to aggregated local judgments (Al-Harthy et al. 2015; Händel and Fritzsche 2016; Shake and Shulley 2014).

In sum, these research results indicate that low-performing students might have some metacognitive awareness. Nevertheless, results (aggregated) at exam level cannot fully explain why low-performing students' SOJs are lower. For one, this pattern of findings may have resulted because these students thought their performance judgments were too high, because they thought these were too low (see Miller and Geraci 2011), because they made SOJs by default (Fritzsche et al. 2018), or for completely other reasons. Serra and DeMarree (2016) discussed that wishful thinking, or more specifically the discrepancy between desired grades and postdictions that results from motivated inferences to protect or enhance self-worth (see Fiske and Taylor 2013), might influence SOJs. Such an approach, however, seems to be more applicable to global judgments because students might have an idea about how well they want to perform in a specific test but not on specific test items. To gain insight into the circumstances under which low-performing students make lower SOJs, one methodological approach would be to assess performance judgments and SOJs at item level and to investigate SOJs by considering the accuracy of the performance judgment (see Fritzsche et al. 2018). Students who are subjectively aware in the sense of Miller and Geraci (2011) should state higher confidence after accurate judgments than after inaccurate judgments. For example, a false alarm, which indicates an inaccurate judgment, should be followed by lower confidence than in the event of a hit. However, accurate judgments of low-performing students might instead result from incorrect items (correct rejections) because low-performing students produce—by definition—more incorrect than correct items. In contrast, high-performing students' accurate judgments might result from correct items (hits). Although both hits and correct rejections are accurate judgments, they stem from either correctly solved or incorrectly solved items, each detected as such. Hence, not only the accuracy of the judgment but also the type of accurate/inaccurate judgment (cf. signal detection theory categories) might have an influence on the respective confidence in it.

Aims of the study

In the study, we pursued three goals. First, we aimed to transfer the unskilled and unaware effect or, respectively, the unskilled but subjectively aware effect to the domain of mathematics, given that no knowledge about SOJs previously existed in this domain, in which all students were shown to be overconfident (Erickson and Heit 2015). Second, we aimed to study reasons for student's alleged unawareness by eliciting their explanations for their judgments. We were interested in which explanations students in the respective performance quartiles provide and, additionally, whether these explanations differ with regard to the adequacy of their judgments and the item solutions. Third, we aimed to investigate the subjective awareness, measured via SOJs, in more detail. Our goal was to understand how SOJs relate to judgment accuracy and whether this relation differs between performance levels.

Hypotheses

We postulated the three following hypotheses.

> H1: Low-performing students in a mathematical knowledge test make more overconfident and inaccurate performance judgments but lower SOJs than do high-performing students.

H2: Students at different performance levels provide different explanations for their performance judgments, which vary in dependence of the fit between performance and performance judgment.
H3: Students' SOJs depend not only on their performance level but also on the accuracy of performance judgment (measured either via a dichotomous accuracy variable or via hits, false alarms, misses, and correct rejections) as well as on the interaction between performance level and judgment accuracy.

Method

Procedure

Students from two German universities took part in the study. Students were tested within group settings. First, students filled in a questionnaire on gender, study term, mathematics as study subject (major, minor, none), and high school grade in mathematics. Second, students took a mathematics test and provided item-specific metacognitive judgments. They had 45 min to complete the test.

Sample

In total, 275 teacher education students participated in the study. Nine students with missing values in more than one third of the items (due to dropout or omission) were excluded from the further analyses, resulting in a final data set of 266 students. The majority of students in our sample were female (78.2%), a gender distribution typical for German teacher education programs. Students were in their first to fourth year of studies ($M = 1.34$, $SD = 0.66$). Eighteen percent of the students studied mathematics as a major, 56.4% as a minor (mathematics education), and 24.8% did not study mathematics at all.

Instruments

Domain-specific performance test As a performance measure, we implemented 18 items from a subtest on mathematical functions of the MTAS (Lienert and Hofer 1972), which we selected in a previous pilot study. The MTAS is a mathematical test for high school graduates and college freshmen. It has several functions, including course guidance (whether to enroll in mathematics/science or in humanistic studies). The test is not mandatory but rather has a counselling function. Students in our sample were not required to pass the test as an admission criterion and were not familiar with the test. For each item, students have to choose one out of three possible answers.

A sample item is "log a + log b =" with the following three answer options: i) log(ab), ii) log (a + b), and iii) $\log(ab^{-1})$. The correct answer is i). Each item was coded as correct (1) if the student chose the correct answer and as incorrect (0) if a student missed the correct answer (ticked none of the answers or chose a wrong one).[2] To prevent students from copying their

[2] This coding is actually only relevant for the calculation of a reliability score, given that all further metacognitive scores were coded as missing if an answer was missing in the performance test (see also our considerations in the section on the assessment of performance judgments and SOJs).

peers' work, we implemented two test forms containing the same items but at different positions. Students' performance did not significantly differ between the two versions ($p = .57$). The test was internally consistent (Cronbach's $\alpha = .71$). The test score significantly correlated with the final high school grade in mathematics, indicating convergent validity (Pearson's $r = .46$, $p < .001$). The test score significantly differed with regard to study subject, $F(2, 261) = 39.47$, $p < .001$, partial $\eta^2 = .23$. Tukey post hoc tests indicate that the students with mathematics as a major had higher performance scores ($M = 0.65$, $SD = .022$) than did the other groups of students (students with mathematics as a minor [$M = 0.43$, $SD = 0.16$], students without mathematics as a subject [$M = 0.38$, $SD = 0.15$], $ps < .001$). Students with mathematics as a minor and students without mathematics as a subject did not significantly differ in their test scores.

After completing each item in the performance test, students were asked to provide several judgments, which are explained in more detail below. Figure 1 illustrates the general procedure of providing item-specific judgments in our study.

Performance judgment After each test item, students were asked to provide a judgment about the correctness of their answer („Do you think you picked the correct answer?"). Students had to tick one of two boxes labeled with "yes" or "no," recoded into 1 for yes and 0 for no. If a student did not provide any answer to the test item itself but nevertheless made a performance judgment, the judgment was recoded as missing. This coding was implemented because no metacognitive ability is required to detect that an item is not correctly solved if the student did not tick any answer at all.

Explanation for the performance judgment We asked students to self-generate answers about the reasons for their performance judgment. After each judgment, students were requested to answer the following question "Why do you think that your provided answer is correct/wrong?"

SOJ We assessed students' confidence in the accuracy of their performance judgment via the following question: "How sure are you that your judgment is correct?" Students were asked to answer this question on a 6-point Likert scale labelled "absolutely unsure, unsure, fairly unsure, fairly sure, sure, absolutely sure," coded from 1 to 6. As with the performance judgments, the SOJ was recoded as missing if the item itself or the performance judgment was missing.

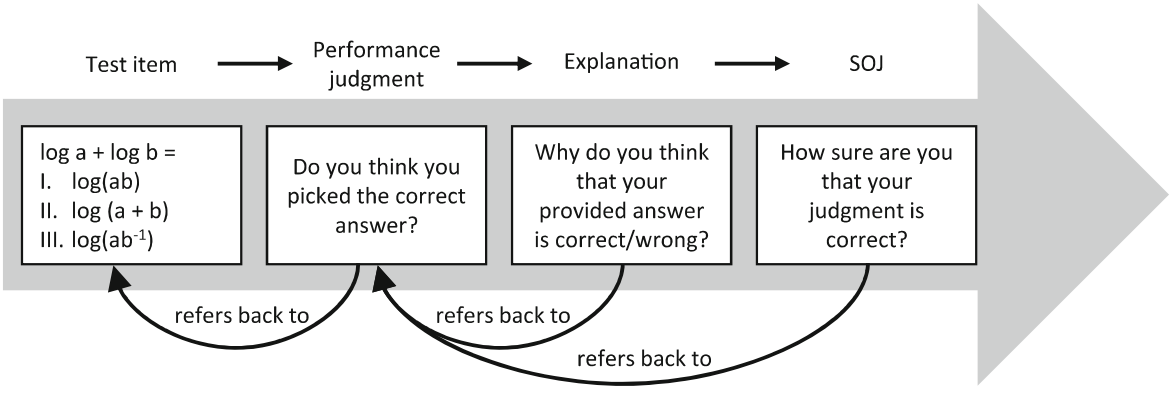

Fig. 1 Procedure of providing metacognitive judgments and SOJs, illustrated via an example test item and the respective item wording

Data preparation

For each item, we calculated a bias and an accuracy score. Bias could result in the values −1 (a student knew the correct answer to the item but thought it was incorrect), 0 (a student accurately judged a correctly/incorrectly solved item as such), or 1 (a student gave an incorrect answer to the test item but thought it was correct). Accuracy values could be 0 (item performance and performance judgment do not correspond to each other) or 1 (the test item is accurately judged as correct/incorrect). In addition, we coded accuracy based on signal detection categories. We calculated whether a judgment is an adequate judgment after a correctly solved item (hit) or after an incorrectly solved item (correct rejection) and, in the same vein, whether an inadequate judgment is a false alarm (item not correctly solved but judged as correctly solved), or miss (item correctly solved but judged as not known).

Two trained raters coded students' open responses that served as explanations for their item-specific judgments into four final categories, which were constructed after a review of the answers and thorough discussions by the authors and raters. The resulting categories were: 1) no/low topic knowledge, 2) guess, 3) gut feeling, and 4) high topic knowledge (see Table 1 for sample explanations). Although one might assume that a guess results from not knowing, we decided to code guessing and not knowing into two separate categories. Especially for a multiple choice test, as used in our study, we presume that students might have some knowledge relevant to the item that leads them to exclude one answer possibility. Nevertheless, they might be confused by one or more other answer options and thus need to guess the correct answer. In sum, 3085 answers out of 4788 explanations (18 items by 266 persons) fit in one of the four categories. 1520 were coded missing when students provided either no explanation or one that was not reasonable. For content considerations, explanations were coded as invalid and were not further taken into consideration if an item judged as incorrect was explained via high topic knowledge (59 explanations)[3] or an item judged as correct was explained via no topic knowledge (124 explanations), resulting in 183 further "missing" codings. Cohen's $\kappa = .90$ (calculated based on 20 double-coded tests) indicates a satisfying interrater-reliability.

Data analyses

To analyze item-specific judgments, we applied a general linear mixed model (GLMM) approach (see Fritzsche et al. 2018), conducted with the package lme4 of the program R (Bates et al. 2015; R Development Core Team 2012). We split students into performance quartiles according to their test score in the mathematical performance test. We then investigated quartile differences according to test performance for the variables performance, performance judgment, bias, accuracy, and SOJ through separate GLMMs under consideration of person and item effects (H1).

Next, we calculated frequencies of different types of self-generated explanations and studied differences between performance quartiles in the relative number of explanations for each signal detection category via Chi-square tests (H2).

[3] High topic knowledge might be a valid source for judging items as not solvable if students' high domain knowledge leads them to the decision that they lack knowledge about the specific requirements of the task. However, this seemed not to be the case in our sample, where such explanations were mostly provided by the lower performing students. We therefore decided to recode these explanations into missing values.

Table 1 Sample explanations for each of the four categories describing reasons for the provided judgments

Category	Sample explanations
No/low knowledge	"I do not know how to calculate this"; "I've never heard about such a task before"
Guessing	"I just guessed"; "guessed"
Gut feeling	"Sounds plausible"; "I think I remember having heard that it works this way"
High topic knowledge	"I just worked it out in my mind"; "I learned this"

Finally, we investigated H3 via two separate GLMM analyses, each with SOJs as dependent variable and considering person and item effects. In a first step, we examined whether confidence in the performance judgment depended on the accuracy of the performance judgment (dichotomous variable: accurate versus inaccurate) across quartiles. In a second step, we studied how differences in students' confidence are contingent on signal detection theory categories across performance quartiles. Signal detection categories were categorically coded as 1 = correct rejection, 2 = false alarm, 3 = miss, and 4 = hit.

Results

Descriptive results for the overall sample show a mean value of item correctness of $M = 0.46$ ($SD = 0.50$). Students' performance judgment was $M = 0.62$ ($SD = 0.49$), and their SOJ $M = 3.91$ ($SD = 1.64$). In other words, item difficulty was about 50%, estimated item correctness was higher, and students were relatively confident in their performance judgment (the SOJ was higher than the average scale value). Students' mean bias score per item was $M = 0.15$ ($SD = 0.63$), indicating overconfidence, and their mean accuracy score per item was $M = 0.59$ ($SD = 0.49$). Students' judgments resulted in 732 correct rejections, 850 false alarms, 369 misses, and 1134 hits.

Unskilled but subjectively aware

The GLMM analyses used to test H1, that is, to replicate the unskilled but subjectively aware effect in mathematics, resulted in significant performance quartile differences for all variables under investigation ($p < .001$ for performance, performance judgment, accuracy, and SOJ and $p < .01$ for bias). The significant random effects indicate that it is necessary to consider item and person level as random variables. Effects were as expected (see Table 2). Figure 2 illustrates the performance quartile differences for each of the variables under investigation. For readability and because effects are most apparent here, we focus exclusively on comparing the effects of Quartile 4 and Quartile 1.

Compared to low-performing students, high-performing students provided significantly higher performance judgments; that is, high-performing students judged items as solved correctly more often than low-performing students did. Students in Quartile 4 were significantly less biased and more accurate in their judgments than students in Quartile 1. Hence, low-performing students had difficulties detecting whether they had solved an item correctly or not and overestimated their item-specific performance. However—in line with the unskilled but subjectively aware effect—low-performing students were less confident than high-performing students in the accuracy of their performance judgments, shown by a significant increase in the SOJ for Quartile 4 compared to Quartile 1.

Table 2 Fixed effects estimates (top) for performance quartiles and variance-covariance estimates (bottom) for the variables performance, performance judgment, bias, accuracy, and SOJ

Parameter	Performance		Performance judgment		Bias		Accuracy		SOJ	
	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)	Estimate	(*SE*)
Fixed effects										
Intercept	0.24***	(0.33)	0.47***	(0.05)	0.22***	(0.05)	0.52***	(0.02)	3.72***	(0.17)
Quartile 2	0.13***	(0.19)	0.04	(0.04)	−0.08*	(0.04)	0.01	(0.02)	−0.26	(0.17)
Quartile 3	0.25***	(0.18)	0.17***	(0.04)	−0.09*	(0.04)	0.03	(0.02)	0.21	(0.16)
Quartile 4	0.51***	(0.19)	0.38***	(0.04)	−0.13***	(0.04)	0.23***	(0.02)	0.85***	(0.17)
Random effects										
Person	0.00	(0.00)	0.03***	(0.19)	0.02***	(0.15)	0.00**	(0.06)	0.83***	(0.91)
Item	0.02***	(0.13)	0.04***	(0.19)	0.04***	(0.20)	0.01***	(0.08)	0.24***	(0.49)
Residual	0.20	(0.45)	0.15	(0.38)	0.33	(0.57)	0.23	(0.47)	1.46	(1.21)

The intercept is equivalent to Quartile 1

* $p < .05$, ** $p < .01$, *** $p < .001$

Students' explanations for their performance judgments

To further explore the effects, we investigated why low-performing students provided inaccurate performance judgments (H2). To that end, we analyzed performance quartile differences in students' self-generated explanations for their judgments in general as well as with respect to correct rejections, false alarms, misses, and hits (cf., Table 3).

Overall, the categories no/low knowledge, guessing, and gut feeling were found similarly often—around 500 times per category. In comparison, high topic knowledge was referred to three times more often. Students with different performance levels provided a different number

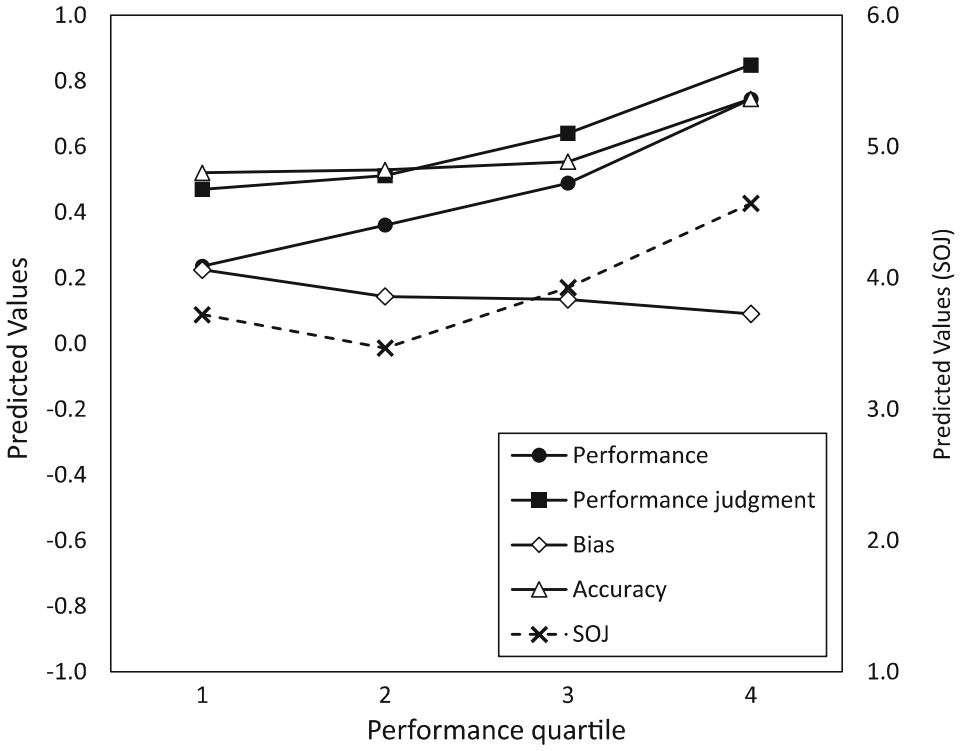

Fig. 2 Predicted values for performance, performance judgment, bias, accuracy, and SOJ for the four performance quartiles (SOJs are displayed on the separate axis on the right and via the dashed line)

Table 3 Frequencies of different explanations for the performance judgments, separately reported for correct rejections, false alarms, misses, and hits as well as for performance quartiles (Relative frequencies within each answer category and each quartile in brackets)

Signal detection category	Explanation	$\chi^2(3)$	Whole sample	Quartile 1	Quartile 2	Quartile 3	Quartile 4
All	No/low knowledge	144.99***	467 (15.1)	207 (28.5)	119 (17.5)	104 (12.5)	37 (4.4)
	Guessing	59.55***	517 (16.8)	153 (21.1)	164 (24.1)	143 (17.2)	57 (6.7)
	Gut feeling	20.40***	546 (17.7)	148 (20.4)	151 (22.2)	157 (18.9)	90 (10.6)
	High topic knowledge	512.42***	1555 (50.4)	219 (30.1)	246 (36.2)	426 (51.3)	664 (78.3)
Correct rejection	No/low knowledge	29.09***	334 (45.6)	170 (56.1)	82 (39.1)	63 (38.7)	19 (33.9)
	Guessing	11.94**	279 (38.1)	87 (28.7)	102 (48.6)	65 (39.9)	25 (44.6)
	Gut feeling	8.36*	119 (16.3)	46 (15.2)	26 (12.4)	35 (21.5)	12 (21.4)
	High topic knowledge						
False alarm	No/low knowledge						
	Guessing	6.80	43 (5.1)	20 (7.9)	8 (3.6)	12 (5.0)	3 (2.2)
	Gut feeling	24.88***	210 (24.7)	69 (27.3)	79 (35.6)	43 (18.0)	19 (14.0)
	High topic knowledge	22.82***	597 (70.2)	164 (64.8)	135 (60.8)	184 (77.0)	114 (83.8)
Miss	No/low knowledge	2.13	133 (36.0)	37 (38.5)	37 (39.0)	41 (33.9)	18 (31.6)
	Guessing	0.93	172 (46.6)	41 (42.7)	47 (49.5)	57 (47.1)	27 (47.4)
	Gut feeling	5.90	64 (17.3)	18 (18.8)	11 (11.6)	23 (19.0)	12 (21.1)
	High topic knowledge						
Hit	No/low knowledge						
	Guessing	17.26***	23 (2.0)	5 (6.7)	7 (4.56)	9 (2.9)	2 (0.3)
	Gut feeling	21.86***	153 (13.5)	15 (20.0)	35 (22.9)	56 (18.2)	47 (7.9)
	High topic knowledge	96.13***	958 (84.5)	55 (73.3)	111 (72.6)	242 (78.8)	550 (91.8)

For each signal detection category, chi-square tests analyzed differences in the relative amount of the respective explanation between performance quartiles

* $p < .05$, ** $p < .01$, *** $p < .001$

of explanations for their performance judgments. Chi-square statistics indicate significant differences between performance quartiles for all categories. Low-performing students more often referred to no/low topic knowledge, guessing, and gut feeling. High-performing students more often referred to topic knowledge.

Finally, students' explanations were investigated for each category of signal detection theory separately. For correct rejections, low-performing students referred more often to no/low knowledge but less often to guessing or gut feeling than high-performing students did. For false alarms, differences were found for gut feeling (low-performing students provided this explanation more often) and for high topic knowledge (provided more often by high-performing students). For misses, the frequencies of each explanation did not significantly differ between performance quartiles. Finally, in the case of hits, high-performing students referred more often to high topic knowledge and less often to guessing or gut feeling compared to low-performing students. These results indicate that students at different performance levels provided different explanations for their performance judgments and that these explanations differed with regard to the fit between item correctness and judgment (that is, by the signal detection categories).

SOJs and judgment accuracy

A GLMM analysis with SOJ as dependent variable and accuracy, performance quartile, and their interaction as independent variables under consideration of person and item level shows significant influences of performance quartile ($p < .001$) and accuracy by performance quartile ($p < .001$) on the confidence in performance judgment accuracy (H3). At the personal level 37.4% of variance could be explained and at the item level 10.5% of variance could be explained (each $p < .001$). For the resulting fixed and random effects, please consult Table 4; comparisons between and within quartiles are presented in Table 5.

Differences in estimated SOJs are illustrated in Fig. 3. Low-performing students reported significantly higher confidence in inaccurate than in accurate performance judgments, indicating a low metacognitive awareness. High-performing students generally reported significantly

Table 4 Fixed effects estimates (top) for performance quartile and accuracy and variance-covariance estimates (bottom) for the SOJ

Parameter	SOJ	(*SE*)
Fixed effects		
Intercept	3.83***	0.17
Accuracy	−0.21**	0.08
Quartile 2	−0.32	0.18
Quartile 3	0.11	0.17
Quartile 4	0.37*	0.19
Accuracy by Quartile 2	0.11	0.11
Accuracy by Quartile 3	0.17	0.10
Accuracy by Quartile 4	0.72***	0.12
Random effects		
Person	0.82***	0.91
Item	0.23***	0.48
Residual	1.41	1.19

The intercept is equivalent to an inaccurate judgment in Quartile 1

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 5 Differences within and between performance quartiles for accurate and inaccurate judgments

Contrast	ΔSOJ	(*SE*)
Quartile 1 inaccurate vs. Quartile 1 accurate	0.21**	0.08
Quartile 1 inaccurate vs. Quartile 4 inaccurate	−0.37*	0.19
Quartile 1 accurate vs. Quartile 4 accurate	−1.10***	0.17
Quartile 4 inaccurate vs. Quartile 4 accurate	−0.52***	0.09

* $p < .05$, ** $p < .01$, *** $p < .001$

higher confidence than low-performing students did—for both inaccurate and accurate judgments. Finally, Quartile 4 students were significantly more confident in accurate performance judgments than in inaccurate ones.

The second GLMM analysis with SOJ as dependent variable and signal detection categories, performance quartile, and their interaction as independent variables under consideration of person and item level shows significant influences of performance quartile ($p = .028$), signal detection categories ($p < .001$), and their interaction ($p < .001$) on the SOJ (H3). Again, a greater amount of variance was explained at the person level (33.8%) than at the item level (4.7%), both significant at $p < .001$. Fixed and random effects are provided in detail in Table 6.

Effects within and between the two extreme quartiles are presented in Fig. 4; for the respective contrasts, see Table 7. First, we depict differences between signal detection categories within each quartile. For Quartiles 1 and 4, a false alarm and a hit led to significantly higher confidence than did a correct rejection or a miss. That is, as long as students thought they had correctly solved an item (independent of whether this was actually the case or not), they reported higher confidence than for items they judged as solved incorrectly. In addition, in both extreme quartiles, students did not show higher confidence in correct rejections as accurate judgments than in misses as incorrect judgments. In other words, if students thought

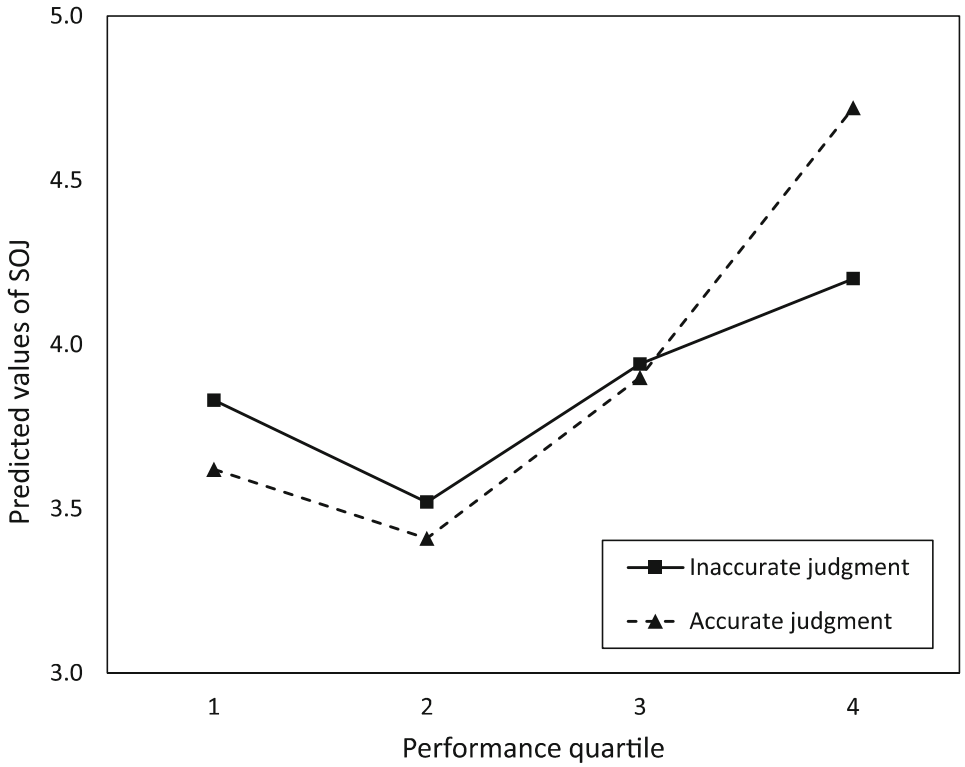

Fig. 3 Interaction effect of performance quartile and accuracy

Table 6 Fixed effects estimates (top) for performance quartile and the four categories of performance judgments and variance-covariance estimates (bottom) for the SOJ

Parameter	SOJ	(*SE*)
Fixed effects		
Intercept	3.45***	0.14
False alarm	0.58***	0.09
Miss	−0.03	0.12
Hit	0.56***	0.12
Quartile 2	−0.52**	0.17
Quartile 3	−0.24	0.17
Quartile 4	−0.31	0.21
False alarm by Quartile 2	0.36**	0.13
Miss by Quartile 2	0.10	0.15
Hit by Quartile 2	0.61***	0.16
False alarm by Quartile 3	0.52***	0.13
Miss by Quartile 3	0.05	0.16
Hit by Quartile 3	0.58***	0.15
False alarm by Quartile 4	0.82***	0.18
Miss by Quartile 4	0.14	0.22
Hit by Quartile 4	1.21***	0.19
Random effects		
Person	0.71***	0.84
Item	0.10***	0.32
Residual	1.29	1.14

The intercept is equivalent to a correct rejection in Quartile 1

** $p < .01$, *** $p < .001$

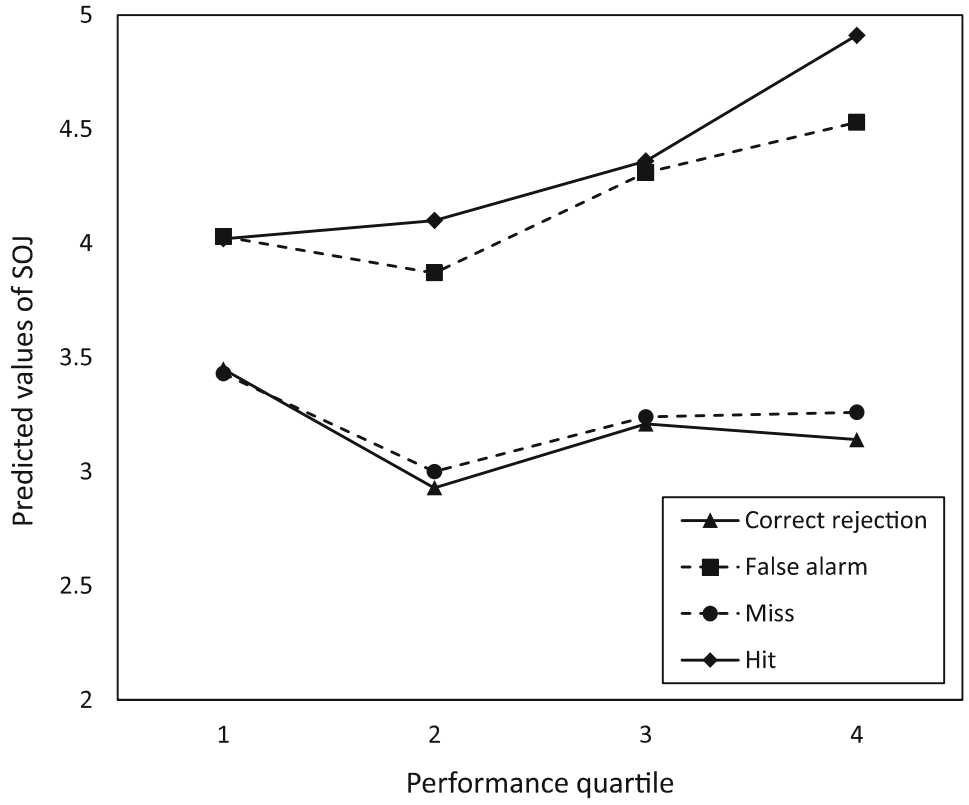

Fig. 4 Interaction effect of performance quartile and the adequacy of judgment in terms of correct rejection, false alarm, miss, and hit

Table 7 Differences within and between performance quartiles for the signal detection categories

Contrast	ΔSOJ	(*SE*)
Quartile 1 correct rejection vs. Quartile 1 false alarm	−0.58***	0.09
Quartile 1 correct rejection vs. Quartile 1 miss	0.02	0.12
Quartile 1 correct rejection vs. Quartile 1 hit	−0.57***	0.12
Quartile 1 false alarm vs. Quartile 1 miss	0.60***	0.12
Quartile 1 false alarm vs. Quartile 1 hit	0.01	0.12
Quartile 1 miss vs. Quartile 1 hit	−0.59***	0.15
Quartile 4 correct rejection vs. Quartile 4 false alarm	−1.39***	0.16
Quartile 4 correct rejection vs. Quartile 4 miss	−0.11	0.18
Quartile 4 correct rejection vs. Quartile 4 hit	−1.77***	0.15
Quartile 4 false alarm vs. Quartile 4 miss	1.27***	0.16
Quartile 4 false alarm vs. Quartile 4 hit	−0.38***	0.10
Quartile 4 miss vs. Quartile 4 hit	−1.65***	0.14
Quartile 1 correct rejection vs. Quartile 4 correct rejection	0.31	0.21
Quartile 1 false alarm vs. Quartile 4 false alarm	−0.50**	0.18
Quartile 1 miss vs. Quartile 4 miss	0.17	0.22
Quartile 1 hit vs. Quartile 4 hit	−0.89***	0.19

** $p < .01$, *** $p < .001$

they had not solved the item correctly, their confidence did not change according to whether they actually solved the item correctly (miss) or not (correct rejection). Comparing hits and false alarms shows significant differences in confidence for high-performing students only. Quartile 4 students were more confident for hits than for false alarms. In contrast, Quartile 1 students did not differ in their confidence for hits or false alarms. Finally, comparing the confidence for specific signal detection categories across performance quartiles indicates that high-performing students reported significantly higher confidence in hits and in false alarms than low-performing students did. No significant differences between performance quartiles were found for correct rejections and misses.

Discussion

Our study conducted an in-depth investigation of the unskilled but subjectively aware effect and implemented an approach with item-specific performance judgments and second-order judgments, accompanied by self-generated explanations. Using GLMM analyses revealed that the unskilled but subjectively aware effect needs to be reinterpreted. Contrary to the assumption that unskilled students' inaccurate performance judgments lead to more moderate SOJs because these students are subjectively aware (Miller and Geraci 2011), our findings indicate that low-performing students fairly irrationally reported lower confidence.

Unskilled but subjectively aware

In the first instance, we replicated the unskilled but subjectively aware effect (Miller and Geraci 2011) in mathematics as another domain (H1). High-performing students provided more accurate judgments than low-performing students did. However, all students, not only the low-performing ones, were overconfident. This pattern of results is consistent with global judgments in the domain of mathematics (Erickson and Heit 2015) and also with results that stem from item-specific judgments in the domain of psychology (Händel and Fritzsche 2016).

Hence, it is not clear whether the general overconfidence is due to being tested in the domain of mathematics or rather to our implementation of item-specific judgments, which are generally known to lead to more overconfident judgments (Gigerenzer et al. 1991). Admittedly, the mathematic-specific argumentation by Erickson and Heit (2015) is limited because their implemented test was comparably difficult. Hence, even the high-performing students in their study might have had difficulties in solving the test. This would be in line with the hard-easy effect (Gigerenzer et al. 1991; Merkle 2009), which describes how people tend to overestimate their performance in hard tasks compared to in easy ones.

Taking item difficulty and the research on the hard-easy effect into account, we believe that the unskilled and unaware effect (Kruger and Dunning 1999) is stable and replicable in the domain of mathematics as well.

As for the subjective awareness postulated by Miller and Geraci (2011), our findings preliminarily indicated that low-performing students provide lower SOJs and therefore might be aware of their low judgment accuracy. To further investigate the unskilled but subjectively aware effect, we performed two additional analysis strategies that gave a different picture.

What were performance judgments comprised of?

Our first procedure to disentangle the whens and whys of inaccurate judgments was based on previous work investigating students' self-generated explanations for their performance judgments (Bol et al. 2010; Dinsmore and Parkinson 2013). On the basis of students' answers, we developed new coding categories and found that students mostly referred to high topic knowledge, followed by three equally frequent explanations: gut feeling, guessing and no/low topic knowledge. Students at different performance levels varied in their frequency of explanations per coded category.

Our analyses of the self-generated explanations for item-specific performance judgments provide further insights into when unskilled students seem unaware of their personal performance (H2). Low-performing students did not appear to acknowledge a hit as such because they less often explained items they thought they had solved correctly by knowledge; that is, they seemed unaware of the knowledge they had. High-performing students, in contrast, explained hits mostly via high topic knowledge, that is, they seemed aware that they knew the answer. In the case of correct rejections, the explanation low-performing students provided for their judgment was no/low topic knowledge about the item. Hence, for incorrect items judged as such, they seemed to be more aware of the knowledge they did not possess. That is, although low-performing students overestimated their performance and had less accurate performance judgments, they nevertheless seemed to be aware of when they really were unskilled. Conversely, high-performing students likely did not want to confess not having knowledge about an item—they explained these judgments instead as guessing rather than not knowing. Hence, we think that in the cases where high-performing students thought they missed the correct answer, they too were likely unaware. As we indicated in the "Method" section, it is not obvious whether guessing and not knowing are distinct categories. One might thus argue that a guess results from not knowing the answer. However, besides theoretically driven assumptions about why there might be a difference—especially in a multiple choice test—we checked our data and found that students intra-individually provided explanations for each of the categories and in doing so might have wanted to express distinct thoughts with them.

Students' mathematical self-concept might have influenced their explanations (Marsh 1986; Skaalvik 1994). Domain-specific self-concept is known to be correlated with respective test performance (Marsh and Craven 2006) and with expectancy of success (Dickhäuser and Reinhard 2006). Consequently, even if high-performing students think they missed the answer, they ascribe this to bad luck rather than to no or low knowledge because their high mathematical self-concept might lead to the assumption that they are generally able to solve such questions. Conversely, the presumably low self-concept of low-performing students might result in explanations about lacking knowledge in the case of correct rejections.

Do SOJs reflect metacognitive awareness?

The two GLMM analyses of SOJs facilitated investigating the adequacy of SOJs under consideration of performance level and absolute accuracy or signal detection categories, respectively. The first GLMM analysis indicated that although confidence was generally lower for low-performing students than for high-performing students, low-performing students' confidence seemed inappropriate: It was lower for accurate judgments than for inaccurate judgments. For high-performing students, the reverse held. The second GLMM analysis considered that inaccurate/accurate judgments might predominantly result from different item correctness (e.g., low-performing students' accurate judgments presumably stem from correct rejections whereas those of high-performing students stem mostly from hits) and yielded further interesting insights into the unskilled but subjectively aware effect. High-performing students seemed better able to discriminate in their confidence in the case of items judged as correct. They seemed able to distinguish whether their judgment that the item had been solved correctly was true or not (with higher confidence for hits than false alarms). Low-performing students, in contrast, did not seem to be aware of the same. Across performance quartiles, students did not seem to be metacognitively aware when they thought they had provided an incorrect item solution. Here, the two extreme performance quartiles did not differ in their confidence regarding misses or correct rejections.

Hence, results of the two GLMMs shed light on the cases in which students' SOJs are appropriate or not. The findings of our approach contradict Miller and Geraci's (2011) assumption that low-performing students are subjectively aware. Both GLMM analyses revealed incongruent SOJs of low-performing students. In addition, the analyses showed under which circumstances even high-performing students lack metacognitive awareness. Students' confidence may well be influenced by motivational factors such as self-concept or biased information processing to protect and enhance self-worth. For example, in a study with fifth-graders in the domain of mathematics, students at equal performance levels who overestimated performance prior to testing were more satisfied with their performance than students who underestimated their performance (Narciss et al. 2011). In our study, high confidence in false alarms might have a similar self-protective function. Furthermore, if students think they do not know the correct answer to a test item, they might nevertheless hope that it might still be correct. Wishful thinking thus might have influenced students' confidence in performance judgment accuracy (see Saenz et al. 2017; Serra and DeMarree 2016). Presumably, the influence of wishful thinking or motivational influences pertains especially to low-performing students, who by definition more often face situations in which they miss correct answers. Moreover, low-performing students might not have recourse to item-specific cues that help them to judge the adequacy of their judgments. In consequence, it seems that SOJs provide further information about

metacognitive awareness solely for high-performing students in the case of items judged as correct (hits versus false alarms).

Limitations and future directions for research

Our study replicated the unskilled but subjectively aware effect in the domain of mathematics and provided relevant insights into its functioning. However, results might be limited because of the chosen test format. The implemented test in our study provided only three answer options per item. That is, students had a guessing probability of 33.3%. Although test difficulty was appropriate for the sample under investigation, this might have influenced students' metacognitive judgments and respective explanations. While the open responses yielded interesting insights into students' reasons for performance judgments, the results are limited by the number of missing values that might have resulted from the considerable amount of effort needed to explain 18 judgments in own words. In addition, another possible limitation of the analyses of SOJs might be that these were assessed within the same students who provided the explanations for their performance judgments. Hence, the SOJs in our study might have been influenced by explanations for the performance judgment, seeing as students did not provide them spontaneously after the performance judgment but had already reflected about their judgments. With regard to the question of the domain-specificity or domain-generality of metacognition (Schraw et al. 1995; Schraw and Nietfeld 1998) and in view of the results for the unskilled and unaware effect in different domains (Erickson and Heit 2015), research in other domains is needed. In particular, we suggest further investigating the open responses in more ecologically valid settings in which students prepare for performance tests.

Conclusion

At first glance, our results indicate that low-performing students reported lower confidence, which would be in accordance with the unskilled but subjectively aware effect. However, the assumption that their lower confidence indicates a (subjective) metacognitive awareness was unsustainable: Their confidence levels did not rationally connect to the adequacy of their performance judgments, neither for the accuracy score nor for the signal detection categories. Hence, contrary to the assumption that unskilled students are subjectively aware (Miller and Geraci 2011), it seems that low-performing students provided lower confidence rather irrationally, likely by default (see Fritzsche et al. 2018). Their confidence differed solely between items judged as correct/incorrect, not with regard to the accuracy of those judgments. Students' self-generated explanations support this result by showing that low-performing students had difficulties in accrediting hits to high topic knowledge and instead referred to guessing or gut feelings. Hence, the two approaches implemented in our study indicate that the unawareness of low-performing students compared to high-performing students might be ascribed not only to the unawareness of lacking knowledge (that is, of being unskilled) but additionally to the unawareness of existing knowledge. Taken together, our results indicate that (low-performing) students indeed have difficulties in accurately estimating their performance. To further clarify the underlying mechanisms, the study of individual differences in self-concept and other motivational variables might be a promising path (Kröner and Biermann 2007). Based on our results, training studies also seem a worthwhile enterprise (see de Bruin et al. 2017; Foster

et al. 2016; Hacker et al. 2008; Johnson et al. 2012; Roelle et al. 2017) in order to help students generate accurate judgments that allow them to pursue self-regulated learning activities accordingly.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Al-Harthy, I. S., Was, C. A., & Hassan, A. S. (2015). Poor performers are poor predictors of performance and they know it: can they improve their prediction accuracy? *Journal of Global Research in Education and Social Science, 4*, 93–100.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*, 133–151. https://doi.org/10.1080/00220970109600653.

Bol, L., & Hacker, D. J. (2012). Calibration research: where do we go from here? *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00229.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*, 269–290.

Bol, L., Riggs, R., Hacker, D. J., Dickerson, D., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education, 21*, 81–96.

Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First- and second-order metacognitive judgments of semantic memory reports: the influence of personality traits and cognitive styles. *Metacognition and Learning, 8*, 79–102. https://doi.org/10.1007/s11409-013-9096-5.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*, 60–77. https://doi.org/10.1037/0022-3514.90.1.60.

de Bruin, A. B. H., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: overconfidence, learning strategy, and personality. *Metacognition and Learning, 12*, 21–43. https://doi.org/10.1007/s11409-016-9159-5.

Dickhäuser, O., & Reinhard, M.-A. (2006). Daumenregel oder Kopfzerbrechen? [rule of thumb or causing headache?]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 38*, 62–68. https://doi.org/10.1026/0049-8637.38.2.62.

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension. *Current Directions in Psychological Science, 16*, 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x.

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology, 132*, 335–346.

Dutke, S., Barenberg, J., & Leopold, C. (2010). Learning from text: knowing the test format enhanced metacognitive monitoring. *Metacognition and Learning, 5*, 195–206. https://doi.org/10.1007/s11409-010-9057-1.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: the MASRL model. *Educational Psychologist, 46*, 6–25. https://doi.org/10.1080/00461520.2011.538645.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002.

Erickson, S., & Heit, E. (2015). Metacognition and confidence: comparing math to other academic subjects. *Frontiers in Psychology, 6*, 1–10. https://doi.org/10.3389/fpsyg.2015.00742.

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition*. Los Angeles: Sage.

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2016). Even after thirteen class exams, students are still overconfident: the role of memory for past exam performance in student predictions. *Metacognition and Learning, 12*, 1–19. https://doi.org/10.1007/s11409-016-9158-6.

Fritzsche, E. S., Händel, M., & Kröner, S. (2018). What do second-order judgments tell us about low-performing students' metacognitive awareness? *Metacognition and Learning., 13*, 159–177. https://doi.org/10.1007/s11409-018-9182-9.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*, 101–121. https://doi.org/10.1007/s11409-008-9021-5.

Händel, M., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: metacognitive monitoring ability and respective awareness in low-performing students. *Memory & Cognition, 44*, 229–241. https://doi.org/10.3758/s13421-015-0552-0.

Johnson, A., Smyers, J., & Purvis, R. (2012). Improving exam performance by metacognitive strategies. *Psychology Learning and Teaching, 11*, 180–185. https://doi.org/10.2304/plat.2012.11.2.180.

Koriat, A. (1997). Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In I. J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 117–135). New York: Psychology Press.

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept — Towards a model of response confidence. *Intelligence, 35*, 580–590. https://doi.org/10.1016/j.intell.2006.09.009.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180–188. https://doi.org/10.1037//0022-3514.82.2.180.

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*, 1121–1134.

Lienert, G. A., & Hofer, M. (1972). *MTAS. Mathematiktest für Abiturienten und Studienanfänger [mathematics test for high-school graduates and freshmen]*. Göttingen: Hogrefe.

Marsh, H. W. (1986). Self-serving effect (bias?) in academic attributions: its relation to academic achievement and self-concept. *Journal of Educational Psychology, 78*, 190–200.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: beysond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*, 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x.

Merkle, E. C. (2009). The disutility of the hard-easy effect in choice confidence. *Psychonomic Bulletin & Review, 16*, 204–213. https://doi.org/10.3758/PBR.16.1.204.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 502–506. https://doi.org/10.1037/a0021802.

Narciss, S., Koerndle, H., & Dresel, M. (2011). Self-evaluation accuracy and satisfaction with performance: are there affective costs or benefits of positive self-evaluation bias? *International Journal of Educational Research, 50*, 230–240. https://doi.org/10.1016/j.ijer.2011.08.004.

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *The Psychology of Learning and Motivation, 26*, 125–141.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*, 7–28.

R Development Core Team (2012). R: A language and environment for statistical computing. Retrieved from Vienna, Austria: http://www.R-project.org/.

Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology, 109*, 99–117. https://doi.org/10.1037/edu0000132.

Saenz, G. D., Geraci, L., Miller, T. M., & Tirso, R. (2017). Metacognition in the classroom: the association between students' exam predictions and their desired grades. *Consciousness and Cognition, 51*, 125–139. https://doi.org/10.1016/j.concog.2017.03.002.

Schaefer, P. S., Williams, C. C., Goodie, A. S., & Campbell, W. K. (2004). Overconfidence and the big five. *Journal of Research in Personality, 38*, 473–480. https://doi.org/10.1016/j.jrp.2003.09.010.

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415–429). New York: Routledge.

Schraw, G., & Nietfeld, J. L. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology, 90*, 236–248.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & DeBacker Roedel, T. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87*, 433–444.

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007.

Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unaware in the classroom: college students' desired grades predict their biased grade predictions. *Memory & Cognition, 44*, 1127–1137. https://doi.org/10.3758/s13421-016-0624-9.

Shake, M. C., & Shulley, L. J. (2014). Differences between functional and subjective overconfidence in postdiction judgments of test performance. *Electronic Journal of Research in Educational Psychology, 12*, 263–282. https://doi.org/10.14204/ejrep.33.14005.

Skaalvik, E. M. (1994). Attribution of perceived achievement in school in general and in maths and verbal areas: relations with academic self-concept and self-esteem. *British Journal of Educational Psychology, 64*, 133–143.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*, 331–362. https://doi.org/10.1080/01638530902959927.