Mem Cogn (2016) 44:229–241
DOI 10.3758/s13421-015-0552-0

Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students

Marion Händel[1] · Eva S. Fritzsche[2]

Published online: 5 October 2015

Abstract Two studies were conducted to further examine the *unskilled-and-unaware effect* and to test whether low-performing students are indeed unaware of their (expected) lower metacognitive monitoring abilities. Postdicted judgments of performance and second-order judgments (SOJs) were solicited to test students' metacognitive awareness. Given that global and local judgments tend to differ (the *confidence–frequency effect*), we investigated whether students' (un)awareness pertains to both types of judgments. A first study focusing on global judgments was conducted in a regular exam setting with 196 undergraduate education students. A second study with 115 undergraduate education students examined both global and local judgments. Local judgments were analyzed on an average level and according to different signal detection theory categories (hits, correct rejections, misses, and false alarms). In both studies, students were grouped in four performance quartiles. The results showed that low-performing students highly overestimated their performance (they were *functionally* overconfident). However, their SOJs indicated that they were less confident in their judgments than the other students, and thus seemed to be aware of their low ability to estimate their own performance (they were not *subjectively* overconfident). This was observed for global as well as for averaged local SOJs. Moreover, an analysis of the local judgments revealed that students' SOJs varied depending not only on whether their judgments were accurate but also on whether or not they thought they knew the answer to an item. In sum, SOJs provide valuable information about students' metacognitive awareness.

Keywords Metacognition · Performance judgments · Second-order judgments · Global judgments · Local judgments

The ability to estimate one's own performance accurately—that is, *metacognitive monitoring* ability—is important in the process of self-regulated learning and affects future performance: Only when students are aware that they lack knowledge to achieve high performance can they seek resources to fill these knowledge gaps. As longitudinal data by Rinne and Mazzocco (2014) have shown, students who had better metacognitive monitoring ability also had larger gains in performance. Students' ability to estimate their own performance can be measured via metacognitive judgments. These judgments differ depending on the time of assessment—that is, on whether they are assessed before or after a performance test—and according to their grain size—that is, whether they are related to a whole test or to single items (Hacker, Bol, & Keener, 2008). Judgments prior to testing, which are called *predictions*, are usually solicited to judge future performance on a comprehension test after students have read a specific text (e.g., Lin & Zabrucky, 1998; Maki & McGuire, 2002). One drawback of predictions about test performance, however, is that they do not rely exclusively on the test to be done (which is usually unknown beforehand), but on more general beliefs about one's own performance in a specific domain (cf. Hacker, Bol, Horgan, & Rakow, 2000). In contrast, *postdictions* are judgments made after a test or a specific item has been completed. Students can accordingly base their

✉ Marion Händel
marion.haendel@fau.de

[1] Department of Psychology and Sport Science, University of Erlangen-Nuremberg, Regensburger Straße 160, 90478 Nuremberg, Germany

[2] Department of Education, University of Erlangen-Nuremberg, Nuremberg, Germany

judgments on their knowledge of the test items, and especially on how well they believe they have performed (see Hacker et al., 2000). By referring to the specific test or item, postdictions appear to be a better indicator of metacognitive monitoring ability than are predictions (Schraw, 2009; Stankov & Crawford, 1996). Consequently, examining postdictions especially adds to our understanding of metacognitive monitoring ability—specifically, of whether or not students are able to monitor their learning appropriately. In recent studies, postdictions were generally more accurate than predictions, likely because students then had more detailed information about what to judge (Hacker et al., 2000; Maki, Jonas, & Kallod, 1994; Pierce & Smith, 2001; Zabrucky, Agler, & Moore, 2009). For these reasons, postdictions are the focus of the present studies.

The accuracy of metacognitive judgments depends not only on the point in time at which they are assessed (pre- vs. postdictions). Differences in accuracy have also been shown between local judgments (or microjudgments) that refer to single items and global judgments (or macrojudgments) that refer to an entire test (Nietfeld, Cao, & Osborne, 2005). Previous literature has indicated that local and global judgments must be distinguished from each other because their frames of reference differ (Gigerenzer, Hoffrage, & Kleinbölting, 1991). The studies by Gigerenzer et al. (1991), Mazzoni and Nelson (1995), Nietfeld et al. (2005), Schraw (1994), and Stankov and Crawford (1996) consistently yielded more accurate judgments for global judgments than for local judgments. Whereas the latter type usually results in overconfidence, global judgments tend to be more appropriate and underconfident (Liberman, 2004). This phenomenon was termed the *confidence–frequency effect* by Gigerenzer and colleagues. It denotes a systematic difference between a metacognitive monitoring judgment of confidence in a single event (local judgment) and a metacognitive monitoring judgment of the frequency of correct answers in total (global judgment). Although local judgments appear to be less accurate, they provide important information about whether students are able to discern correctly and incorrectly solved items. To investigate metacognitive monitoring ability for global and local postdictions, two studies were implemented, each focusing on one of these two kinds of judgments.

The accuracy of metacognitive judgments is usually assessed via a difference score that indicates under- or overestimation. However, if some students highly overestimate performance and others highly underestimate, a difference value close to zero would result for the whole sample, suggesting perfect (mean) accuracy (Hacker, Bol, & Bahbahani, 2008). For that reason, the difference score might not be a valid indicator of how accurate students actually are. An absolute difference score provides further information on the accuracy of a judgment and overcomes this limitation of the difference score. To analyze metacognitive judgments at an item level, several approaches can be applied (see Lichtenstein & Fischhoff, 1977, or Schraw, 2009, for discussions of different calibration measures). For example, gamma is a commonly applied measure (Nelson, 1986); however, it has the disadvantage of not being able to measure any bias (Jang, Wallsten, & Huber, 2012).

Metacognitive monitoring ability and performance

Alongside general tendencies toward over- or underestimation in local and global judgments, students' performance has been a strong predictor of judgment accuracy: Higher-performing students are more accurate than lower-performing students in their metacognitive monitoring judgments (Bol & Hacker, 2001; Hacker, Bol, & Keener, 2008; Nietfeld et al., 2005). Kruger and Dunning (1999) conducted several studies within different domains to investigate how accurately students at different performance levels estimated their performance in a previously performed test (global postdictions). Kruger and Dunning investigated whether students were able to judge their own performance in comparison to others' performance and with regard to their own raw performance score. After grouping students post-hoc in four performance quartiles, they found across all domains that students in the bottom performance quartile significantly overestimated their performance relative to students in the top performance quartile. On the basis of this finding, Kruger and Dunning argued that "incompetence . . . not only causes poor performance but also the inability to recognize that one's performance is poor" (p. 1130). They concluded that the lowest-performing students were unaware of being unskilled due to a lack of metacognitive skills. This prominent study sparked further studies investigating the *unskilled-and-unaware effect* in different settings and domains (e.g., Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). However, not all previous studies could replicate the proposed effect; heterogeneous results occurred especially with test items in (educational) psychology (cf. Hacker et al., 2000; Hartwig & Dunlosky, 2014). These differences in study outcomes may be due to differences in group performance levels, because students are classified as low-performing in relation to the other study participants. For instance, in the studies by Hacker et al. (2000), only a small group of the students categorized as very-low-performing were unaware with respect to pre- and postdictions. Additionally, alternate (methodological) explanations were provided for the low accuracy of low-performing students, such as the better-than-average effect or regression to the mean (Burson, Larrick, & Klayman, 2006; Krajc & Ortmann, 2008; Krueger & Mueller, 2002). Others suggested that different assessment and analysis approaches to judging performance might lead to different interpretations of the results (Ackerman, Beier, & Bowen, 2002). Nevertheless, "perhaps the leading interpretation is that low performers are overconfident because they have a general deficit of metacognitive insight" (Miller & Geraci, 2011, p. 502).

Awareness of metacognitive monitoring ability

Miller and Geraci (2011) further investigated the supposed unawareness of low-performing students by soliciting their predictions in an exam situation and assessing metacognitive monitoring ability on the basis of these performance judgments. In addition, the authors asked for second-order judgments (SOJs) to assess the students' awareness of their metacognitive monitoring ability. SOJs are confidence judgments for previously given performance judgments (Dunlosky, Serra, Matvey, & Rawson, 2005) and can be regarded as meta-monitoring (Dunlosky et al., 2005) or meta-metacognitive judgments (Buratti & Allwood, 2012). Confirming the results by Kruger and Dunning (1999), Miller and Geraci found that low-performing students' performance judgments were too high in comparison with their actual performance. Analysis of the SOJs revealed, however, that the students seemed aware of their low metacognitive monitoring ability. Hence, although low-performing students provided exaggerated performance estimates, their confidence in their performance judgments was significantly lower than that of higher-performing students, who provided more accurate estimates. The authors differentiated between functional overconfidence—that is, people estimate their performance higher than it actually is—and subjective overconfidence—that is, people are overly certain of their estimates. Accordingly, low-performing students are functionally but not subjectively overconfident.

In their study, Miller and Geraci (2011) extended Kruger and Dunning's (1999) research with predictions in a regular exam situation and by using SOJs. Whether the results concerning functional and subjective overconfidence also apply to SOJs of postdictions, however, has not yet been investigated. Postdictions are based on the privileged knowledge of the type of test and its items (Hacker et al., 2000), do not so heavily rely on general beliefs such as do self-efficacy and self-concept, and are therefore assumed to be better indicators of metacognitive monitoring. They are also of particular importance for individual self-assessment in the process of self-regulated learning (e.g., when solving mock exams). Hence, the question is whether, with postdictions, low-performing students are still functionally unaware. To that end, in the present two studies we investigated postdictions. The first study presented in this article focused on global judgments. In the second study, we additionally investigated local judgments, whose assessment offers several theoretical and methodological advantages. The specific strengths of local judgments are that they provide a fine-grained picture of students' metacognitive monitoring ability and that they are a more specific measure of metacognitive monitoring because students have to refer to specific items (Nietfeld et al., 2005). Moreover, the assessment of local judgments offers the advantage of calculating internal consistency for the judgments, which would not be possible with a single global judgment. Local judgments refer to specific items and might also be less strongly influenced by a personality trait such as self-concept (Gigerenzer et al., 1991). Hence, rules of thumb cannot be applied, and students have to make a new decision for each item. Finally, local judgments are necessary as a source of information when students try to assess their knowledge (i.e., to monitor their knowledge in the process of self-regulated learning) and want to know explicitly which contents they need to study further (and not only to judge the proportion of knowledge that they already possess).

The use of local judgments enables differences in SOJs to be investigated with regard to the accuracy of performance judgments at an item level. Two questions regarding SOJs are of interest: whether the respective item was solved correctly, and whether the person indicated having solved the item correctly (performance judgment). Such a classification of actual performance and judged performance on single items corresponds to the classification of items according to signal detection theory (Green & Swets, 1966). The four possible combinations are *hits* (correctly solved and expected to be correct), *misses* (correctly solved but expected to be wrong), *correct rejections* (wrong solution and expected to be wrong), and *false alarms* (wrong solution but expected to be correct) (cf. Schraw, Kuch, & Gutierrez, 2013, or Winne & Muis, 2011, for overviews of the statistical tools to assess metacognitive judgment accuracy according to signal detection theory). The successful analysis of metacognitive judgments with these measures (see, e.g., Barrett, Dienes, & Seth, 2013; Jang et al., 2012; Maniscalco & Lau, 2012; Masson & Rotello, 2009) was the point of origin for our analyses of SOJs. We investigated how confident students are about items that they accurately judged as being correct/incorrect (hits/correct rejections), as compared to items for which they overestimated (false alarm) or underestimated (misses) their performance. This specific analysis at an item level goes beyond earlier studies based on SOJs that focused on judgments of averaged test performance.

Aims of the studies

The present two studies aimed to broaden previous findings on metacognitive monitoring ability and the awareness of this ability, especially in low-performing students. First, to investigate whether low-performing students are indeed unaware of their expected low metacognitive monitoring ability, in Study 1 we examined SOJs in addition to performance judgments. In accordance with Miller and Geraci (2011), lower-performing students were assumed to be overconfident in their performance judgments (functional overconfidence). However, the low-performing students' confidence in their performance judgments should be lower than that of the high-performing students (no subjective overconfidence). Second, Study 2 tested whether global judgments and their respective SOJs would be more accurate than the averaged local ones (confidence–frequency effect). How this might affect the judgment accuracy of students at different performance levels was a key research question. To investigate judgment accuracy, an absolute difference score, in addition to a difference score, was calculated for global as well as for local judgments. Because the difference score averages the levels of the true difference, the absolute difference score is assumed to result in a stronger discrepancy than the difference score. Third, the SOJs of local judgments were analyzed at an item level to investigate whether students discriminate in their SOJs according to the four possible combinations of performance and performance judgments outlined by signal detection theory. Assuming that students are aware of their metacognitive monitoring ability, the SOJs of correctly estimated items (hits or correct rejections) were expected to be higher than the SOJs for wrongly estimated items (misses or false alarms).

General method

Two studies were conducted to extend previous work on students' ability to judge their own performance, and thus their metacognitive awareness. In both studies, undergraduate education students were asked to postdict their own performance as well as to provide a respective SOJ. In Study 1, a single global judgment was investigated—that is, a judgment about performance on the entire test. In Study 2, local judgments were investigated, as well. The data from Study 1 were collected within an ecologically valid setting during a regular exam situation, and Study 2 was conducted as a laboratory study. The contents of the implemented tests were similar in both studies.

Study 1

Overview

The study was implemented in the final exam at the end of a study term. All data included in this study were assessed by all participating students at the same point in time. After students completed their exam, they were asked to voluntarily provide two judgments (a performance judgment and an SOJ) on a separate paper sheet. Altogether, students were given 45 min to complete the exam and to make the two judgments.

Method

Participants

The participating students attended an educational psychology course for undergraduate education students. Of the 351 students who took the corresponding exam, 196 voluntarily provided performance judgments and SOJs. Further analyses relied on this subsample only. Most were first-year students (89.2 %) and female (72.8 %), which is typical for an introductory course in education.

Instruments

Performance The final exam served as an indicator of students' performance in educational psychology. The students had to answer 32 multiple-choice questions (four options, single select, Cronbach's α = .81). A sample item was "Which method is suitable to assess learning processes? (A) Portfolio, (B) Oral exam, (C) Written exam, or (D) Interview" [the correct answer is "(A) Portfolio"].

Postdicted performance judgment After students completed their exam, they were asked to judge the raw score of items that they answered correctly (global performance judgment). The implemented question was, "What do you think: How many of the 32 test items did you solve correctly? _____ items."

SOJ Finally, students were asked to provide an SOJ. They were requested to judge their confidence in their performance judgment on a 5-point-Likert scale. The implemented item was "How confident are you that your performance judgment is correct?" Students were asked to select a confidence rating on a 5-point smiley scale, displayed in Fig. 1 (see Händel & Fritzsche, 2015; Jäger, 2004). The frowning face on the left represented low confidence(1), and the smiling face on the right represented high confidence(5).

Data analysis

To interpret the resulting scores more easily, the performance score and the global performance judgment score were recoded into percentage scores; for example, 16 of 32 items solved correctly was recoded to 50. A difference score was calculated as "estimated performance– actual performance"

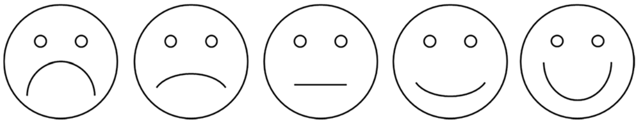

Fig. 1 Implemented 5-point rating smiley scale of the second-order judgments

in order to investigate the degree of over- or underestimation of the global performance judgment. Negative values (max = –100) refer to an underestimation, and positive values (max = 100) to an overestimation of performance. The absolute value of the difference score was calculated in order to investigate the accuracy of the performance judgments. The absolute difference score was |estimated performance– actual performance|. Values close to 0 indicated high accuracy, and values close to 100 indicated low accuracy.

To investigate the influence of performance on the accuracy of the respective judgments, we followed the procedure implemented by Kruger and Dunning (1999) and Miller and Geraci (2011) by grouping students into four performance quartiles (Q1 = lowest-performing students, to Q4 = highest-performing students). A multivariate analysis of variance (MANOVA) was performed with the performance quartile as the independent variable and judgment, difference, absolute difference, and SOJ as dependent variables.

Results

Overall, students had high performance scores and slightly underestimated their performance.[1] Table 1 shows the descriptive statistics for each performance quartile in which students were grouped. Figure 2 additionally provides a quick overview of the descriptive study results, summarized for both studies.

A MANOVA resulted in the following statistical differences between the performance quartiles. First, the difference scores differed significantly between the performance quartiles [$F(3, 196) = 23.34$, $p < .001$, $\eta_p^2 = .267$]: Students in the bottom quartile overestimated their performance, whereas students in the other three quartiles underestimated their performance (Tukey post-hoc comparisons between all quartiles except Q2–Q3 and Q3–Q4 were statistically significant, *p*s < .001). The quartile differences in absolute differences did not reach statistical significance [$F(3, 196) = 2.34$, $p = .075$, $\eta_p^2 = .035$]. However, students in Q1 showed the highest absolute difference on a descriptive level. Finally, a significant difference emerged in the SOJs: $F(3, 185) = 3.64$, $p = .014$, $\eta_p^2 = .054$. Tukey post-hoc tests indicated that students in Q1 and Q2 were less confident in their ratings than the students in the top quartile (*p*s = .023 and .034, respectively).

Discussion

The aim of the first study was to investigate whether low-performing students are unaware of their (expected) low metacognitive monitoring ability. To create a combined investigation of metacognitive monitoring ability and of the awareness of this ability, Kruger and Dunning's study procedure (1999) was extended to include SOJs (cf. Miller & Geraci, 2011) and was conducted in a regular exam setting (see also the first study by Ehrlinger et al., 2008). Before considering low-performing students' degrees of awareness of their expected low metacognitive monitoring ability, we first discuss whether their metacognitive monitoring ability was in fact low. Our results replicated the unskilled-and-unaware effect in its initial consideration (in other words, Q1 students were functionally overconfident) with respect to educational psychology questions. This result is noteworthy because it contradicts an earlier study also conducted with introductory psychology items in which the unskilled-and-unaware effect was not shown (Hartwig & Dunlosky, 2014). As expected, students in the lowest performance quartile overestimated their performance according to the difference score (functional overconfidence). In contrast, students in higher performance quartiles underestimated their own performance, which might be explained by a statistical artifact: The students in Q4 had a performance score of nearly 90. Hence, they had little opportunity to overestimate, but much space to underestimate, their own performance. In addition, the effect discussed by Liberman (2004) might have played a role: Insofar as students have to provide global judgments, they do not sufficiently take into account the guessing rate in their judgment.

Furthermore, in the present study we investigated the accuracy of the postdictions via the absolute difference score as a measure for metacognitive monitoring ability. The absolute difference scores did not differ significantly between the four performance quartiles. That is, students in the four performance quartiles seemed equally able to judge their global performance accurately. However, the descriptive statistics and the medium (albeit nonsignificant) effect size indicate that students in the lowest performance quartile tended to be the least accurate.

Regarding the issue of main interest—awareness of metacognitive monitoring ability assessed via SOJs—the students in the lowest performance quartile (who were functionally overconfident) were least confident in their performance estimations. That is, although low-performing students seemed to overestimate their own performance and appeared—on a descriptive level—least accurate, they seemed to acknowledge this in their confidence about their performance judgments. In other words, they did not seem to be *subjectively* overconfident. By contrast, the underestimation of students in the high performance quartile was related with the highest SOJs.

These results have implications for self-regulated learning. When students test their knowledge via solving a mock exam for which they do not have the sample solution, their further learning process will be based on their anticipated success. By considering their largely overestimated performance judgments only, low-performing students would presumably not invest enough effort in the further learning process, whereas high-performing students, who largely underestimate their

[1] Table 2 reports the overall values for both studies.

Table 1 Descriptive statistics [*M* (*SD*)] for performance, performance judgments, differences, absolute differences, and second-order judgments (SOJs) in Study 1, reported separately by performance quartiles for global judgments

Q	*N*	Performance	Postdicted Performance Judgment	Difference	Absolute Difference	SOJ
1	44	48.93 (10.26)	57.74 (16.27)	8.81 (17.83)	14.49 (13.51)	3.00 (0.94)
2	52	67.43 (3.73)	65.87 (11.46)	–1.56 (10.90)	9.13 (6.03)	3.04 (0.74)
3	44	78.55 (2.29)	70.24 (12.69)	–8.31 (12.39)	11.58 (9.33)	3.20 (0.67)
4	56	88.67 (3.89)	76.90 (10.23)	–11.77 (10.24)	12.00 (9.97)	3.43 (0.60)

Q = Quartile

performance, would presumably invest more effort than necessary. This learning pattern seems especially detrimental for low-performing students. On closer examination, however, this potential consequence appears to be attenuated by the lower SOJs, indicating that low-performing students are indeed aware of their inaccurate performance judgments.

Study 1 was limited by the fact that each student provided only one performance judgment and only one SOJ. From a methodological perspective, this did not allow us to compute internal consistencies of the performance judgments or SOJs. A global judgment can be used to gain a general idea of whether students are over- or underconfident, or whether they correctly estimate their performance. However, global judgments do not provide any information about whether students are able to discern correctly and wrongly solved items. Imagine two students (Susan and Robert) who both provide a correct judgment of 20 correctly solved items. Asking them which items were correctly solved could result in Susan showing perfect calibration by picking the exact items correctly solved, but Robert showing worse calibration, by being able to pick only ten items correctly solved (and also picking ten wrong items, while not being able to identify the other ten items that were correctly solved). In consequence, global judgments might be a good indicator of general over- or underconfidence, but they are limited regarding any information about students' ability to distinguish between correct and wrong items. It was hence our aim to overcome this limitation in Study 2. To gain a finer-grained picture of students' metacognitive monitoring abilities, performance judgments and their respective SOJs were investigated at an item level in Study 2.

Study 2

Taking into account the confidence–frequency effect, in the second study we investigated local judgments. In particular, the study addressed the question: Do the differences between performance quartiles occur not only with global judgments, but also with averaged local judgments? In addition, we analyzed

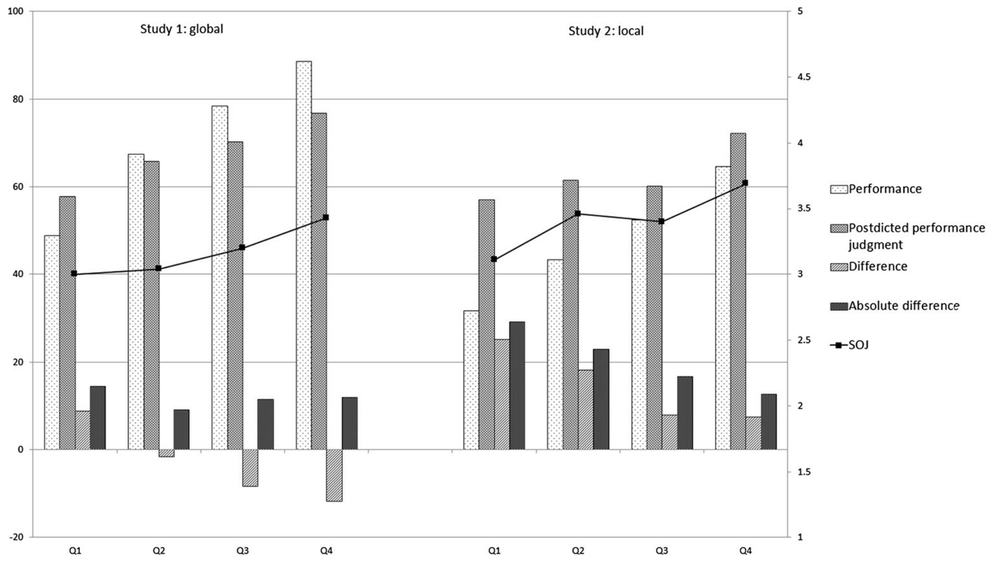

Fig. 2 Summarized results for Studies 1 and 2: Performance, performance judgments, differences, absolute differences, and second-order judgments (SOJs), reported separately for each performance quartile. Note that the performance and difference scores are represented via bar charts measured on the left-hand ordinate; the SOJs are represented via line graphs measured on the right-hand ordinate. Q = Quartile

whether local SOJs differ on the basis of whether an item was correctly or mistakenly judged as being correct or incorrect.

Overview

This study was conducted as a laboratory study in which students were asked for global and local judgments. Students were tested in groups of 10–20 persons. Lasting about 45 min, the session was guided by one of the authors of this article. The participating students were asked to complete a performance test covering educational psychology topics. Local performance judgments as well as their respective SOJs were collected after each test item.[2] Furthermore, after students had completed the whole test, they were asked to provide a global judgment and SOJ, as in Study 1.

Method

Participants

A total of $N = 115$ undergraduate education students voluntarily participated in the study (75.9 % female). The students were recruited from different advanced courses in educational psychology. They were enrolled in different terms, most of them (78.3 %) in the third to the fifth term. Only students who had already passed their exam for the introductory course in educational psychology were recruited for the study. The reason was that participating students should not gain any informational advantage over other students for an exam they had not yet passed.

Instruments

Performance A test with 32 multiple-choice questions (four options, single select) about educational psychology topics served as an indicator for students' performance (Cronbach's $\alpha = .60$). A sample item was, "What kind of learning strategy is the actualization of prior knowledge? (A) An organizational strategy, (B) An elaboration strategy, (C) A self-regulatory strategy, or (D) A resource strategy" [the correct answer is "(B) An elaboration strategy"].

Postdicted performance judgments After each test item, students were asked to indicate whether their answer was correct or not, resulting in 32 local performance judgments (Cronbach's $\alpha = .88$). The implemented question was, "Do you think your answer is correct?" Students had to tick one of two boxes labeled "yes" or "no." In addition, the students made a global performance judgment on a separate paper sheet after completing the whole test. The global performance judgment was the same as in Study 1.

SOJs For each item, students were asked how confident they were about their performance judgment (again operationalized via a 5-point Likert smiley scale, with the frowning face on the left representing low confidence [1] and the smiling face on the right representing high confidence [5]; see Fig. 1). The 32 SOJs revealed high reliability (Cronbach's $\alpha = .90$). In addition, an SOJ was requested for the students' global performance judgment (see Study 1).

Data analysis

For the global judgments, the same scores were calculated as in Study 1 (a difference score and an absolute difference score). To compare the results of the global to the local judgments, a mean percentage score was calculated for the test score and the corresponding local judgments (*averaged score*). This was done by counting the number of "yes" judgments, divided by the number of total items and multiplied by 100 (e.g., a student who judged half of the items to be solved correctly, indicated by a "yes" judgment for half of the items, would obtain a mean score of 50). This recoding facilitated calculating an averaged difference and an averaged absolute difference score, as we described for the global judgments in Study 1. For the 32 SOJs, a mean score was calculated. The levels of the local and global scores (performance, differences, absolute differences, and SOJs) were compared via paired-sample *t*tests.

The local judgments were further investigated in more detail with two different approaches. First, the averaged local judgments and SOJs were investigated subject to performance. That is, students were again grouped in quartiles according to their respective performance. As in Study 1, and in line with Miller and Geraci's (2011) analyses, a MANOVA was performed, with performance quartile as the independent variable and the averaged judgments, differences, absolute differences, and SOJs as dependent variables.

Second, the local SOJs were analyzed at an item level with regard to four categories: whether or not students solved the item correctly, combined with the question of whether or not students thought they had solved the item correctly (hits, misses, correct rejections, and false alarms).[3] This classification

[2] Because item-level judgments are more time-consuming than only one global rating (with the addition of a performance judgment and an SOJ after each test item, students had to answer three times as many items), these additional judgments might have negatively influenced their performance if they had been presented as part of a real exam, as in Study 1. For that reason, the second study was instead conducted in a laboratory setting. In addition, from a research perspective, this should prevent many missing values (due to students' focusing on the exam rather than on their facultative judgments).

[3] Signal detection theory was applied only to classify the performance judgments in relation to actual performance in order to investigate SOJs on the basis of the four categories. Of course, further measures related to signal detection theory can be calculated (see the General Discussion), but this would go beyond the scope of our research.

was used to analyze whether SOJs differed as to whether an item was accurately or inaccurately judged as correct or incorrect. That is, SOJs were analyzed according to performance (item correct or incorrect) and to the performance judgment (which could be accurate or not). To do so, paired-sample *t* tests were computed between the SOJs of each category.

Results

The descriptive statistics of the local and global judgments are shown in Table 2.[4] As expected, students' global judgments and their SOJs after having completed the whole test were lower than the mean value of the local judgments and the corresponding SOJs. A negative difference score was found for the global judgments, as compared to a higher and positive difference score for the local judgments. In addition, the global judgment was closer than the mean value of the local judgments to the actual performance score (lower absolute difference score). All paired-sample *t*tests comparing the respective local and global scores were significant (postdicted performance: $p < .001$, Cohen's $d = 0.88$; difference: $p < .001$, $d = 0.89$; absolute difference: $p = .004$, $d = 0.38$; SOJ: $p = .003$, $d = 0.37$).

Students were grouped into four performance quartiles (see Table 3 for the descriptive statistics for performance, expected performance, and SOJs, grouped in the four performance quartiles). Those with higher performance showed less overestimation than did students with lower performance [$F(3, 116) = 5.22$, $p = .002$, $\eta_p^2 = .123$; Tukey post-hoc tests indicated significant differences between Q1 and Q3/Q4; $p = .006/p = .009$]. Significant differences were also observed for absolute differences: The higher the performance, the lower the absolute difference [$F(3, 116) = 7.19$, $p < .001$, $\eta_p^2 = .161$]. Significant post-hoc differences existed between Q1 and Q3/Q4 ($p = .005/p < .001$) and between Q2 and Q4 ($p = .036$). That is, students in the lower-performing quartiles were less accurate than those in the higher-performing quartiles in assessing their performance judgments. Finally, the SOJs differed between the four performance quartiles [$F(3, 116) = 6.99$, $p < .001$, $\eta_p^2 = .158$; Tukey post-hoc: $p < .001$ for Q1 as compared to Q4, $p = .021$ for Q1 as compared to Q2]. Students in the higher performance quartiles were more confident that their performance judgments were correct, indicated by higher SOJs.

For the investigation of SOJs at an item level, the local performance judgments were additionally coded in accordance with signal detection theory. For more than half of the items, students correctly estimated their performance (hits and correct rejections). Students mistakenly believed that they had solved an item correctly (false alarm) more often than they mistakenly believed that they had incorrectly solved one (misses); see Table 4 for the exact percentages of items.

The respective SOJs differed with regard to the accuracy of the judgments (see Table 4). The highest SOJs were provided for hits, followed, however, not by the SOJs for correct rejections but by those for false alarms. The paired-sample *t*tests of the SOJs for the four types of judgments were significant at the $p < .001$ level, except for the comparison of the items that were judged as incorrect (misses and correct rejections: $p = .120$). The effect sizes were medium to high (Cohen's *d*s: hit vs. correct rejection = 1.21, hit vs. false alarm = 0.60, hit vs. miss = 1.37, correct rejection vs. false alarm = –0.76, false alarm vs. miss = 0.92).

Discussion

The aim of the second study was to investigate the confidence–frequency effect and to analyze SOJs at an item level. First, the local judgments and their respective SOJs were shown to be reliable measures. Second, the global judgment was more accurate than the averaged local one. This is in line with the confidence–frequency effect and with previous research comparing global and local judgments (Gigerenzer et al., 1991; Mazzoni & Nelson, 1995; Nietfeld et al., 2005; Schraw, 1994).

Two relevant findings emerged from the analyses with the averaged local judgments: Students in the lowest performance quartile overestimated their performance to the highest degree and were least accurate. Furthermore, students in the lowest performance quartile experienced lower confidence in their performance ratings, indicated by the lower SOJs. This indicates that low-performing students were aware of their inaccurate judgments (see Miller & Geraci, 2011).

A further aim of the second study was to investigate whether students can discern between correct and incorrect answers. To that end, local judgments were analyzed with reference to the classification system of signal detection theory. Until now, previous research on local (second-order) judgments had used confidence scales rather than dichotomous items for performance judgments (Buratti & Allwood, 2012; Dunlosky et al., 2005). This made it impossible to categorize performance judgments according to signal detection theory. Our analyses based on the four categories showed that students demonstrated higher confidence in hits than in misses or false alarms. However, they were also more confident of hits than of correct rejections. In addition, the SOJs were higher for false alarms than for misses or correct rejections. Hence, students were more confident if they thought that they had solved the item correctly (regardless of whether this was true or not) than if they actually provided an accurate judgment (i.e., rightly judged that they did not know an item). This result conflicts with our expectation that SOJs would be higher for correct judgments (hits and correct rejections) than for wrong

[4] For a better overview, Table 2 also reports the descriptive results for global judgments in Study 1.

Table 2 Descriptive statistics [*M* (*SD*)] for performance, performance judgments, differences, absolute differences, and SOJs, reported separately for (averaged) local and global judgments

Score	Local (Study 2)	Global (Study 2)	Global (Study 1)
Performance	47.36 (12.24)		71.84 (15.73)
Postdicted performance judgment	62.34 (20.81)	44.62 (19.66)	68.13 (14.37)
Difference	15.03 (20.82)	–2.93 (19.73)	–3.18 (14.99)
Absolute difference	20.74 (15.08)	15.43 (12.56)	11.70 (10.03)
SOJ	3.41 (0.50)	3.14 (0.91)	3.18 (0.75)

judgments (misses and false alarms). It means that expecting to have solved an item correctly also evokes a more positive SOJ than does expecting to have solved an item incorrectly. Taken together, students' SOJs differed not only with regard to the accuracy of their performance judgments but also with regard to their judgments themselves (yes/no).

Four different assumptions might be responsible for the counterintuitive result of higher confidence in false alarms than in correct rejections. First, if students had been absolutely sure that the answer they chose was wrong, they could have selected another one of the four possible responses in the multiple-choice test. Therefore, it seems plausible that their confidence in answers judged to be incorrect would be lower. Second, it is assumed that individuals tend to be more confident about results that are desirable outcomes (like a correctly solved test item). Wishful thinking might play a role, meaning that students truly hoped that their judgment about a correct answer was accurate. If this applies, students' high confidence in false alarms, although unfavorable from a self-regulated learning perspective, can be explained from a motivational perspective. Third, students' performance judgments and their SOJs might not really differ (an assumption discussed by Buratti & Allwood, 2012). That is, if students state that they have solved an item correctly, they confirm the statement with their SOJ. However, a previous validation study by the authors (unpublished) does not support this conjecture. In the interview study, the majority of students (17 out of 18) distinguished the performance judgments and the SOJs in the intended manner (i.e., they described the performance judgment as such and the SOJ as confidence in the previously given performance judgment). Finally, the result can be associated with the trace accessibility model of Koriat (1993, 1995). Although Koriat's studies investigated another type of judgment than we did in our research—namely, feelings of knowing—his suggestions about how judgments emerge might be applicable to SOJs of postdictions at an item level. Judgments then might be higher for false alarms than for correct rejections because the amount of partial information that comes to mind is higher if students think that the answer is correct than if they correctly acknowledge that they do not know the answer.

Higher SOJs for false alarms and hits than for misses might also be explained by the different amount of partial information that comes to mind. The low SOJs for misses might additionally be a result of guessing. That is, if students make a guess because they have absolutely no idea of the correct answer, they might judge the item as not being correctly solved. However, they might assume that the item could nonetheless be correct, on the basis of a 25 % chance of correctly guessing for a multiple-choice item with four options, and might provide a correspondingly low SOJ.

The present study is the first to reveal that SOJs differ with regard to hits, correct rejections, misses, and false alarms. A methodological constraint of this procedure, however, is that the SOJs in each of the four categories were based on unequal numbers of items per person, because the numbers of hits, correct rejections, misses, and false alarms differed between individual students. The implications for future analyses of SOJs in the context of signal detection theory will be outlined in the General Discussion.

General discussion

Are low-performing students aware or unaware of their metacognitive monitoring ability (the unskilled-and-unaware effect proposed by Kruger & Dunning, 1999)? Two studies were conducted with global and local postdictions as indicators for metacognitive monitoring ability, with SOJs as the indicator for metacognitive awareness. In both studies, students were grouped in four performance quartiles, and their

Table 3 Descriptive statistics [*M* (*SD*)] for performance, performance judgments, differences, absolute differences, and SOJs in Study 2, reported separately by performance quartiles for local judgments

Q	*N*	Performance	Postdicted Performance Judgment	Difference	Absolute Difference	SOJ
1	28	31.81 (5.11)	57.03 (23.88)	25.22 (23.12)	29.24 (17.54)	3.11 (0.59)
2	34	43.38 (2.75)	61.58 (20.46)	18.20 (19.95)	22.98 (13.97)	3.46 (0.39)
3	30	52.50 (2.65)	60.21 (17.81)	7.92 (18.22)	16.67 (10.46)	3.40 (0.47)
4	24	64.71 (5.72)	72.27 (18.73)	7.55 (16.86)	12.76 (13.19)	3.69 (0.36)

Q = Quartile

Table 4 Percentages of items and the respective SOJs [*M* (*SD*)] of items that were accurately or inaccurately judged as correct/incorrect

Type of Judgment	Percentage of Items	SOJ
Hits	33.16 (15.30)	3.78 (0.51)
Correct rejections	23.11 (13.59)	3.02 (0.75)
False alarms	29.09 (12.85)	3.48 (0.51)
Misses	14.12 (10.31)	2.91 (0.75)

metacognitive monitoring ability was investigated via two scores: a difference score and an absolute difference score. In addition, SOJs were analyzed at an item level, and we investigated whether students differ in their metacognitive awareness depending on the combination of items' correctness and performance judgments classified in terms of signal detection theory.

Functional versus subjective overconfidence in low-performing students

The two studies revealed that in their global and local judgments, low-performing students overestimated their performance (functional overconfidence). Furthermore, the accuracy of performance judgments was lowest in low-performing students (at a descriptive level for the global judgments in Study 1, and significantly different in Study 2, with local judgments). This implies that low-performing students seem to have few chances to improve their future performance, because their overestimations would lead them to feel comparably capable and would not induce them to invest further resources for learning. However, the low-performing students seemed to be aware of their low metacognitive ability: The students in the four performance quartiles differed significantly in their SOJs, with low-performing students being least confident in their judgments, and high-performing students being most confident. These results suggest that low-performing students (although functionally overconfident) are aware of their comparably low metacognitive ability (no subjective overconfidence). That is, the results of Miller and Geraci (2011) for global predictions were extended by using global postdictions and SOJs in an exam setting (Study 1), along with averaged local postdictions and the respective SOJs in Study 2.

The question remains why low-performing students—although overestimating their performance—seem to be aware of their low judgment accuracy. Students who possess the requested test knowledge (i.e., high-performing students) seem to find it easier to judge the number of items correct, and therefore to be confident in their judgments. By contrast, low-performing students likely know that they made some guesses in the multiple-choice test, and consequently are less confident in their performance judgments. Another potential explanation is that low-performing students might have given up on improving their learning or knowledge and are resigned to receiving low grades (cf. Hacker et al. 2000). In that case, at least some of the low-performing students might not be heavily engaged in the monitoring process and might not be motivated to provide accurate judgments that could allow them to take responsibility for future learning. These students might consequently provide lower SOJs.

When we compared global and local performance judgments across all performance quartiles, the global judgments were more accurate than the averaged local ones, which is in line with the confidence–frequency effect. Consistent with the results discussed by Liberman (2004), the local judgments were higher than the actual performance score (overconfidence), and the global judgments were lower than actual performance (underconfidence). Of interest was how this influenced quartile differences; when we compared students at different performance levels, the patterns of results were quite similar for local and global judgments. That is, the type of judgment had no influence on the differences in accuracy between the quartiles.

Item-level analyses

Finally, the local SOJs were categorized in line with signal detection theory, and differences in the SOJs were analyzed according to whether the performance judgments were hits, misses, correct rejections, or false alarms. As expected, SOJs were highest for hits. Contrary to our expectations, however, SOJs were higher for items classified as false alarms than for items classified as correct rejections. Students were more conservative in their confidence ratings if they thought that their answer to an item was incorrect. For the SOJs, it thus seems relevant not only whether the person made a correct performance judgment, but whether or not the person believed that an item was correctly solved. Assuming that students make these considerations while solving a (mock) exam and that SOJs affect future learning, students might overly study content already known (but not identified as such), while neglecting to study content that they (incorrectly) assume they know already.

Limitations of the studies

Four methodological aspects might limit the validity of the studies. First, the settings between the studies differed (real exam setting and laboratory study), and the tests were of different difficulties and reliabilities. Nevertheless, the patterns of results were comparable. The lower test scores and low internal consistency of the test in Study 2 might have resulted from the different setting, in which students did not explicitly prepare for the test. Students had already learned the subject matter of the test, but some students seemed to have forgotten some of it, and other students seemed to have forgotten another part of it. Nonetheless, the low reliability of the test limits the significance of the results presented in Study 2.

Second, assessing local and global judgments within the same sample and the same test, as was done in Study 2, has a drawback that might also have affected earlier studies: Given the fact that the local judgments had already been provided beforehand, the global judgments might have been influenced by them. Recent research by Hartwig and Dunlosky (2014), however, has refuted this concern, inasmuch as the authors did not find any dependence of local on global judgments. A comparison between two different tests, one using local and the other using global judgments, would, in addition, be less economical and would hamper comparison of the judgments—for instance, through possibly discrepancies in test difficulty.

Third, the analyses of differences and absolute differences according to performance quartiles lacked statistical independence, since both scores shared some variance with the performance score.

Finally, our data do not provide information about whether low-performing students are more conservative in their confidence ratings because they think their performance judgments are too high or because they think their performance judgments are too low (see also Miller & Geraci, 2011). It also might well be that low-performing students generally provide lower performance and confidence judgments than high-performing students. Hence, although the low-performing students were admittedly the most overconfident, they nevertheless provided the lowest performance judgments relative to the other quartiles. However, a review of data from another study (Jacob, Händel, Markus, & Eberle, 2014) that used a subsample of the Study 1 participants indicated that students in the lowest performance quartile did not generally score lower on questionnaire items that asked, for example, about self-efficacy, learning goals, or the use of learning strategies. Further measures, such as the fit of the absolute difference scores and the SOJs (cf. Händel & Fritzsche, 2015), could provide more evidence on whether low-performing students' SOJs can in fact be regarded as appropriate.

Be that as it may, in our study, the analysis of SOJs at an item level provided information about the types of items for which students in general made lower SOJs (i.e., correct rejections and misses).

Implications

First, we discuss the implications of the assessment procedure we used. Second, we discuss potential work for future studies. Our studies applied Miller and Geraci's (2011) procedure of asking for performance judgments and the respective SOJs through both global and local postdictions. In our view, postdictions of raw performance scores are a valid measurement of metacognitive monitoring. In particular, judgments of raw performance are assumed to be an easier task than percentile rank judgments (Hartwig & Dunlosky, 2014), because judging raw performance requires only judging the self, whereas judging performance relative to others (percentile ranks) requires both judging the self and estimating the performance of others. Although methodological effects (such as regression to the mean or the better-than-average effect) and motivational biases (such as the self-enhancement bias or wishful thinking) might be an issue for the accuracy of performance judgments, they are assumed to be more evident in research that asks for individuals' estimations of performance relative to others than for judgments of raw performance scores (cf. Kwan, John, Kenny, Bond, & Robins, 2004, on the distinction of self-enhancement as social comparison and self-enhancement as self-insight). This argument also applies to SOJs, which are regarded as a valid measurement of metacognitive awareness. In addition, the results for the SOJs in our studies differed from those for the performance judgments, revealing that SOJs are not functionally equivalent to performance judgments. Specifically, asking for SOJs is not the same as asking twice for a performance judgment (as was discussed by Allwood, Granhag, & Johansson, 2003). Notwithstanding, performance judgments and SOJs are probably both influenced by the particular content knowledge. That is, the more information or content knowledge regarding a specific item that students have at their disposal, the more strongly they believe that they provided the correct answer to the item, and consequently they are confident that their postdiction was correct. Nevertheless, because SOJs are not directly linked to performance in a domain but to metacognitive monitoring ability, they are regarded as a valid indicator for assessing metacognitive awareness. When asked for SOJs, students likely feel compelled to take a critical stance and to reflect on behavioral consequences, such as engaging in learning activities to fill knowledge gaps (Buratti & Allwood, 2012; Metcalfe, 1998). Indeed, making an SOJ might be similar to the second stage of the two-stage process discussed for delayed judgments of learning (cf. Son & Metcalfe, 2005; see also the two-process hypotheses discussed by Dunlosky et al., 2005). Our results indicate that low-performing students are not completely unaware, which should encourage researchers and practitioners to support students in providing accurate judgments. Indeed, study results by Ryvkin, Krajč, and Ortmann (2012) indicated that the unskilled are not doomed to remain unaware.

To strengthen the outcomes on postdictions combined with SOJs, the present studies need to be replicated using, for instance, other tests, domains, and samples. In particular, a more reliable test would be needed when investigating local performance judgments and their respective SOJs. Because the focus of our studies was on SOJs rather than on the metacognitive monitoring judgments themselves, we decided to calculate a difference score and to categorize the performance judgments according to signal detection theory.

Because our results are aggregated across individuals (who of course differed in terms of their numbers of hits, correct rejections, false alarms, and misses), multilevel analyses would be needed to gain further insight into SOJs for local judgments. The analysis of SOJs from a multilevel perspective looks very promising for further studies on the nature of SOJs (Murayama, Sakaki, Yan, & Smith, 2014). In addition, further measures to assess students' awareness of their metacognitive monitoring abilities need to be implemented (see above for recent developments of measures based on signal detection theory). For instance, measures of resolution can provide information about whether students are able to discern between correct and incorrect answers. More specifically, future studies could include measures based on signal detection theory, such as d_a, which is the distance between the means of the distributions of metacognitive judgments of hits versus false alarms (cf. Masson & Rotello, 2009); other measures to assess students' calibration are discussed in Schraw et al. (2013).

Our results indicate that the difference score was averaged due to under- and overestimation, which would lead to an overvaluation of students' accuracy if it were used as a single measure of metacognitive judgments, as had been done in earlier studies. For that reason, we recommend calculating an accuracy score, such as an absolute difference score, in addition to a difference score. Finally, future research will need to investigate the impact of our findings on students' learning processes.

Conclusion

According to the results of the present studies and the work by Miller and Geraci (2011), there is strong evidence that, although low-performing students are inaccurate in their performance judgments and tend to overestimate their own performance, they do not appear to be overconfident in the accuracy of these judgments. An analysis of their SOJs suggests that even if students are functionally unaware, they are subjectively aware. That is, although low-performing students overestimated their performance, they seemed to know that, or at least were less confident in their performance judgments than high-performing students. For self-regulated learning, functional overconfidence seems less problematic, provided that students are at least subjectively aware. If they implicitly know that they might be not as good as they think they are, they will presumably take responsibility for their learning.

Author note E.S.F. is now at the Department of Teacher Education, School of Education, Technische Universität München, Germany. This research was supported by a grant from the "Sonderfonds für wissenschaftliches Arbeiten an der Universität Erlangen-Nürnberg." We thank Albert Ziegler, Stephan Kröner, Bettina Harder, Kathrin Lockl, and Luc P. Beaudoin for in-depth discussions of our study results.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33,* 587–605. doi:10.1016/S0191-8869(01)00174-X

Allwood, C. M., Granhag, P. A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgements: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17,* 545–561. doi:10.1002/acp.888

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

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. doi:10.1080/00220970109600653

Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive Processing, 13,* 243–253. doi:10.1007/s10339-012-0440-5

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. doi:10.1037/0022-3514.90.1.60

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *Journal of General Psychology, 132,* 335–346. doi:10.3200/GENP.132.4.335-346

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. doi:10.1016/j.obhdp.2007.05.002

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528. doi:10.1037/0033-295X.98.4.506

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

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92,* 160–170. doi:10.1037/0022-0663.92.1.160

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

Hacker, D. J., Bol, L., & Keener, M. C. (2008b). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). New York: Routledge.

Händel, M., & Fritzsche, E. S. (2015). Students' confidence in their performance judgements: A comparison of different response scales. *Educational Psychology, 35,* 377–395. doi:10.1080/01443410.2014.895295

Hartwig, M., & Dunlosky, J. (2014). The contribution of judgment scale to the unskilled-and-unaware phenomenon: How evaluating others can exaggerate over- (and under-) confidence. *Memory & Cognition, 42,* 164–173. doi:10.3758/s13421-013-0351-4

Jacob, B., Händel, M., Markus, S., & Eberle, T. (2014). *Der 3 × 2 Achievement Goal Questionnaire und das Problem der semantischen Ähnlichkeit* [Semantic overlap with the 3 × 2 Achievement Goal Questionnaire]. Paper presented at the 79. Meeting of the Arbeitsgruppe für empirische pädagogische Forschung (AEPF) in Hamburg, Germany.

Jäger, R. (2004). Konstruktion einer Ratingskala mit Smilies als symbolische Marken [Development of a rating scale with smileys as anchors]. *Diagnostica, 50,* 31–38. doi:10.1026/0012-1924.50.1.31

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review, 119,* 186–200. doi:10.1037/a0025960

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639. doi:10.1037/0033-295X.100.4.609

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124,* 311–333. doi:10.1037/0096-3445.124.3.311

Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology, 29,* 724–738. doi:10.1016/j.joep.2007.12.006

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. doi: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. doi:10.1037/0022-3514.77.6.1121

Kwan, V. S., John, O. P., Kenny, D. A., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review, 111,* 94–110. doi:10.1037/0033-295X.111.1.94

Liberman, V. (2004). Local and global judgments of confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 729–732. doi:10.1037/0278-7393.30.3.729

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23,* 345–391.

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review, 1,* 126–129. doi:10.3758/BF03200769

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 39–67). Cambridge: Cambridge University Press.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21,* 4222–4430. doi:10.1016/j.concog.2011.09.021

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509–527. doi:10.1037/a0014876

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1263–1274. doi:10.1037/0278-7393.21.5.1263

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

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. doi:10.1037/a0021802

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1287–1306. doi:10.1037/a0036914

Nelson, T. O. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100,* 128–132. doi:10.1037/0033-2909.100.1.128

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.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29,* 62–67. doi:10.3758/BF03195741

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS ONE, 9*(e98663), 1–11. doi:10.1371/journal.pone.0098663

Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? *Journal of Economic Psychology, 33,* 1012–1031. doi:10.1016/j.joep.2012.06.003

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19,* 143–154.

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

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition, 33,* 1116–1129. doi:10.3758/BF03193217

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personal and Individual Differences, 21,* 971–986. doi:10.1016/S0191-8869(96)00130-4

Winne, P. H., & Muis, K. R. (2011). Statistical estimates of learners' judgments about knowledge in calibration of achievement. *Metacognition Learning, 6,* 179–193. doi:10.1007/s11409-011-9074-8

Zabrucky, K. M., Agler, L.-M. L., & Moore, D. (2009). Metacognition in Taiwan: Students' calibration of comprehension and performance. *International Journal of Psychology, 44,* 305–312.