Received: 8 June 2018 | Revised: 17 December 2018 | Accepted: 13 March 2019

DOI: 10.1002/acp.3548

RESEARCH ARTICLE

WILEY

Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels

Marloes L. Nederhand[1] | Huib K. Tabbers[1] | Remy M.J.P. Rikers[1,2]

[1] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2] Roosevelt Center for Excellence in Education, UCR, Utrecht University, Utrecht, The Netherlands

Correspondence
Marloes L. Nederhand, Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Woudestein, Mandeville Building, T16.16, Rotterdam 3000 DR, The Netherlands.
Email: m.l.nederhand@essb.eur.nl

Funding information
Research Excellence Initiative

Summary

This experimental study explores whether feedback in the form of standards helps students in giving more accurate performance estimates not only on current tasks but also on new, similar tasks and whether performance level influences the effect of standards. We provided 122 first-year psychology students with seven texts that contained key terms. After reading each text, participants recalled the correct definitions of the key terms and estimated the quality of their recall. Half of the participants subsequently received standards and again estimated their own performance. Results showed that providing standards led to better calibration accuracy, both on current tasks and on new, similar tasks, when standards were not available yet. Furthermore, with or without standards, high performers calibrated better than low performers. However, results showed that especially low performers' calibration accuracy benefitted from receiving standards.

KEYWORDS

calibration accuracy, feedback, monitoring, performance level, self-assessment

1 | INTRODUCTION

To study effectively, students must make adequate decisions about what they already understand and what they need to restudy. This requires accurate calibration: being able to estimate the level of one's own performance (Alexander, 2013; Dunlosky & Thiede, 2013; Lichtenstein, Fischhoff, & Phillips, 1982). Inaccurate calibration is linked to poor academic performance (Bol, Hacker, O'Shea, & Allen, 2005; De Bruin, Kok, Lobbestael, & De Grip, 2017; Dunlosky & Rawson, 2012; Nietfeld, Cao, & Osborne, 2006). When students inaccurately estimate their performance, they may fail to change strategies or prematurely end studying because they wrongly think they already mastered the material (Bol et al., 2005; Dunlosky & Rawson, 2012; Nietfeld et al., 2006; Rawson & Dunlosky, 2007).

Research has shown that calibration accuracy can be improved by providing students with extra cues. For example, feedback in the form of performance standards (i.e., the correct answer) makes students' estimates of their performance more accurately (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Dunlosky & Thiede, 2013; Lipko et al., 2009). Because students regularly use self-testing with feedback as a strategy to monitor their learning progress (Hartwig & Dunlosky, 2012; Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2007), the beneficial effect of standards seems to have a lot of promise for educational practice.

However, it remains yet unclear whether all students benefit equally from receiving standards. Although it has been argued before that performance level may influence the benefit of standards (e.g., Stone, 2000; Zimmerman, 2002), only a few studies investigating the effect of standards on calibration accuracy have included performance

level as a factor. The first aim of our study was therefore to investigate whether the effect of performance standards on calibration accuracy will be different for high and low performers. Furthermore, it has been argued that standards received in the past may also improve performance estimates on future tasks (Koriat, 1997; Zimmerman, 2000). However, empirical evidence for this assumption is scarce. Hence, our second aim was to investigate whether providing performance standards will improve calibration accuracy not only on the current task but also on subsequent, similar tasks when standards are not available anymore.

1.1 | Improving calibration accuracy by providing performance standards

Students experience difficulties in estimating their own performance because they often use unreliable and false cues to estimate, such as the quantity of information they recalled rather than the quality (Baker & Dunlosky, 2006). By comparing their own performance to standards (i.e., does the provided answer match or mismatch with the correct answer?), students generate a much more valid cue of the quality of their performance (Koriat, 1997; Thiede, Griffin, Wiley, & Anderson, 2010), which, in turn, will result in more realistic performance estimates.

In a key study, Rawson and Dunlosky (2007) demonstrated the effect of standards on calibration accuracy. They provided psychology students with six texts that contained four key words with definitions. Students were given time to study each text and to learn the definitions. Afterwards, students were asked to recall the definitions and to estimate how well their recalled definition matched the actual definition. Half of the students received a performance standard (i.e., the correct definition) while estimating their performance, whereas the other half of the students did not. The results showed that students who received performance standards while estimating performance calibrated better than students who did not receive any standards (Rawson & Dunlosky, 2007). This finding has been replicated several times (Dunlosky et al., 2011; Dunlosky & Thiede, 2013; Lipko et al., 2009; Van Loon & Roebers, 2017) and clearly shows that providing a standard improves calibration accuracy.

1.2 | Competence to use standards

Although providing standards improves calibration accuracy, standards do not remedy all miscalibration. Rawson and Dunlosky (2007) also found that students are still limited in their competence to use standards: They often assign more credit to their answers than appropriate (Dunlosky et al., 2011; Lipko et al., 2009; Rawson & Dunlosky, 2007; Thiede et al., 2010). In these cases, students seem to generate incorrect cues from the standard because they overestimate the number of critical elements present in their recalled definition.

Rawson and Dunlosky (2007) did not investigate whether students differ in their competence to use standards. However, in previous studies on calibration accuracy, it was found that performance level plays an important role (Bol et al., 2005; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999). In general, high performers (often defined as those belonging to the upper quartile) are better calibrated than low performers (those belonging to the bottom quartile). It has been argued that low performers use less valid cues to estimate their performance than high performers (Gutierrez de Blume, Wells, Davis, & Parker, 2017).

So how does performance level relate to the effect of standards on calibration accuracy? On the one hand, low performers may benefit more from receiving standards because these standards provide them with more valid cues (Thiede et al., 2010), and low performers have more room for improvement (Bol et al., 2005; Ehrlinger et al., 2008; Kruger & Dunning, 1999). On the other hand, low performers may benefit less from standards than high performers because they are more likely to generate incorrect cues due to their limited competence.

In our study, we thus aim to clarify the role of performance level by investigating whether or not providing performance standards will improve calibration accuracy similarly for both high and low performers.

1.3 | Learning to calibrate accurately

Imagine students reading three definitions they later have to recall. For the first two definitions, the students are asked to estimate the quality of their recalled definitions while receiving standards. On the basis of the previous research (e.g., Rawson & Dunlosky, 2007), we can assume that receiving the standards will improve these students' calibration accuracy. However, what will happen if on the third definition, the students do not receive a standard anymore. Will they still give a more accurate estimate than if they had not received any standards on the previous two definitions? In other words, can providing standards make students learn how to give more accurate estimates on similar tasks?

As previously mentioned, Koriat (1997) argued that the quality of calibration depends on the cues that are used. When students are comparing their own answer to a standard, the standard serves as a cue about the quality of their performance (did it match or mismatch the desired answer?). However, the process of comparing own answer with a standard may also provide students with a cue about the quality of their estimate of performance (did their initial performance estimate match the outcome as scored with the standard present?). In turn, students could use both this performance cue and the calibration accuracy cue to make better judgements on new, subsequent texts. For example, if students recognize that they have overestimated their own performance, they could become more careful and conservative when estimating their performance on new definitions. It could therefore be argued that providing students with standards will not only improve their calibration accuracy on the current task, due to a valid cue about the quality of their performance, but also improve their calibration on a similar subsequent task without a standard present, due to a valid cue about the quality of their estimate.

Empirical findings to support this argument are yet lacking. There are, however, some studies that investigated the issue with other types of feedback. For example, when students had to estimate how well they had performed on an exam, their calibration accuracy improved if they were encouraged to attend to the outcome feedback they had received on previous exams (Hacker, Bol, Horgan, & Rakow, 2000; Labuhn, Zimmerman, & Hasselhorn, 2010; Miller & Geraci, 2011; Nietfeld et al., 2006). So, it seems that reminding students of their previous performance led to better calibration accuracy on subsequent tasks (Lichtenstein & Fischhoff, 1977). Hence, the second aim of our study was to investigate whether the effect of standards on calibration accuracy can also be found on a new task that is similar in structure, but different in content, when standards are not present anymore.

1.4 | Present study

The present study aimed to answer two research questions:

1. Do students from different performance levels benefit equally from receiving performance standards to improve their calibration accuracy?
2. Does providing performance standards also improve calibration accuracy on subsequent, similar tasks, when standards are not present anymore?

Additional to our main research questions, we also investigated whether we could replicate the basic finding that providing standards while estimating performance will benefit calibration accuracy.

We investigated our research questions by using the method and materials from the key study by Rawson and Dunlosky (2007) with some minor adaptations. We hypothesized that we would replicate the positive effect of standards on calibration accuracy, found by Rawson and Dunlosky (2007) and explored whether low performers and high performers benefitted equally from receiving standards. Finally, we explored whether students receiving performance standards indeed improved their calibration accuracy on subsequent tasks when standards were not yet available. Based on theory (Koriat, 1997), we expected that providing standards would indeed improve calibration on subsequent tasks. Because low and high performing students may not benefit equally, we also included performance level in this analysis.

2 | METHOD

2.1 | Participants and design

The participants in this study consisted of 126 first-year psychology students from a Dutch university. Four students experienced technical difficulties while participating in the experiment, and we therefore excluded their answers from our data file, resulting in 122 participants. The participants had a mean age of 19.82 (*SD* = 3.50), with 84.4% females and 15.6% males. Students received course credit for their participation and provided informed consent for their participation. Furthermore, our Institutional Research Committee of the Institute of Psychology provided approval for this experiment.

The experiment conformed to a 2 Standards (Yes vs. No) × 3 Performance level (Low vs. Medium vs. High) design. Students were randomly assigned to the conditions, with 62 students in the standards group and 60 students in the no-standards group. Within each experimental group, we defined three performance level groups based on students' overall performance (i.e., how many definitions were correctly recalled by each student). In both the standard and no-standard group, we defined students as low performing when they scored below the 33th percentile, medium performing when they scored between the 33th and 66th percentile, and high performing when they scored above the 66th percentile.[1] Table 1 displays the performance accuracy of the percentile groups.

2.2 | Materials

Computers presented all materials and recorded the responses by the students, using the online software Qualtrics.

2.3 | Texts

Students had to read the same texts as those used by Rawson and Dunlosky (2007). The texts used in our experiment had been translated into Dutch by De Bruin et al. (2017), and the translated texts ranged between 273 and 303 words. The subjects of the texts were taken from textbooks of undergraduate courses, such as communication and family studies. Each of the six critical texts that were presented to our students contained subjects that had not been part of their curriculum yet. Each text contained four key terms in capital letters that were followed by a definition students needed to learn and recall (e.g., "EMBLEMS are gestures that represent words or ideas"). See Appendix A for a sample text.

2.4 | Recall test

The recall test required students to write down the definitions of the key terms from the text they had just learned. Because each text contained four key terms, students had to recall four corresponding definitions. Students were presented with one key term at a time

[1]The choice of splitting performance level into different groups was inspired by prior research on calibration accuracy, in which performance-level differences are typically operationalized by divided students into different groups, mostly by median split (e.g., Bol et al., 2005; Hacker et al., 2000) or by using quartiles (Kruger & Dunning, 1999). In our case, we decided against a median split because there would have been too much overlap between high and low performers—the two groups that were of most interest to use. Quartiles, however, would have required us to test a substantial larger number of participants (note that we already tested twice as many participants as the original study by Rawson & Dunlosky, 2007) although we were not interested in specific differences between the second and third quartile. Hence, to prevent too much information loss due to overlap between the performance-level groups, while focusing on our main question, we decided to use three performance level groups.

TABLE 1 Test performance scores

	Standards								
	No			Yes			Total		
Performance level	*N*	*M* (*SE*)	95% *CI*	*N*	*M* (*SE*)	95% *CI*	*N*	*M* (*SE*)	95% *CI*
Low	24	.44 (.02)	[.39, .48]	24	.51 (.02)	[.47, .55]	48	.47 (.01)	[.44, .50]
Medium	17	.61 (.01)	[.59, .63]	21	.69 (.01)	[.67, .70]	38	.65 (.01)	[.63, .67]
High	19	.78 (.02)	[.75, .81]	17	.83 (.01)	[.80, .86]	36	.80 (.01)	[.78, .83]
Total	60	.59 (.02)	[.55, .64]	62	.66 (.02)	[.62, .69]	122	.63 (.01)	[.60, .65]

Note. This table displays test performance scores of low, medium, and high performers in both the no-standard group and the standard group. Low performers perform least well in both standard groups. Furthermore, high performers perform best in both standard groups. There are no test performance differences between the no-standard and standard group.

and were asked to type in the definition they thought corresponded to this key term. The definitions recalled by the students were scored by the first author with a scoring grid used in previous studies (e.g., Dunlosky, Rawson, & Middleton, 2005; Rawson & Dunlosky, 2007). Definitions were awarded with full (1 point), partial (0.5 point), or no credit (0 point). A second rater independently scored a random selection (9.84%) of the entire data set. A sufficient degree of agreement was found between the two raters, with an intraclass correlation for single measures of .83, with a 95% confidence interval from .79 to .87. Consequently, the scoring of the first rater was used as measure of actual obtained credit per definition.

2.5 | Performance standards

The standard group received a performance standard in the form of a correct definition of each key term (cf. Rawson & Dunlosky, 2007). Such a standard was presented together with the definition provided by the student, so students could compare their own definition to the correct definition.

2.6 | Performance estimates

2.6.1 | Global prediction

Only because we aimed to follow the procedure of Rawson and Dunlosky (2007) as closely as possible, we included a global prediction measure in our study. Right after reading a text, students were presented with the following question: "How well will you be able to complete a test over this material?" Students rated their answer on a scale from 0 (*definitely will not be able*) to 10 (*definitely will be able*).

2.6.2 | Postdiction without standard present

For each recalled definition, all students estimated the credit they would thought they would obtain on a three-point scale, ranging from no credit (0 point), partial credit (0.5 point), to full credit (1 point). For each text, the average of the four estimates was taken as a measure of postdiction without standard present.

2.6.3 | Postdiction with standard present

Students in the standard group also had to provide a second estimate but this time in the presence of a performance standard. Students used the same three-point rating scale, and for each text, the average of the four estimates was taken as a measure of postdiction with standard present.

2.7 | Calibration accuracy

To investigate their hypotheses on the effect of standards on calibration accuracy, Rawson and Dunlosky (2007) made a qualitative distinction between different recall responses. They divided the students' responses into five categories: omission error (no response); commission error (students provided a completely incorrect response); partially correct (a response that can be rewarded with some, but not all, credit); partial plus commission (although a student provided some correct information, he or she also reported incorrect information); and correct (fully correct response). Subsequently, Rawson and Dunlosky compared the standard and no-standard condition on their average performance estimate within each response category. However, in our study, we wanted to use a more general estimate of calibration accuracy (cf. Labuhn et al., 2010; Nietfeld et al., 2006).[2] Therefore, we defined calibration accuracy as the quantitative difference between performance estimate and actual obtained credit. Calibration accuracy is optimal when performance estimates are similar to actual obtained credit. So, the closer the calibration accuracy score is to zero, the better. Operationalizing calibration accuracy this way enabled us to compare our conditions not only on accuracy but also on direction of miscalibration (bias), to explore whether students overestimated or underestimated themselves. The different calibration accuracy scores are explained below.

[2]For archival purposes, we also performed the response category analysis. The graphical depiction of the results is added to Appendix B, showing an identical pattern as in Rawson and Dunlosky (2007).

2.7.1 | Global prediction accuracy

Although the quality of predictions was not of central interest in our study, we explored whether students' predictions improved after receiving standards. Global prediction accuracy was calculated as the absolute difference between the global prediction of each text and the average obtained credit for each text (i.e., mean obtained credit of the four recalled definitions, multiplied by 10 to get the same 10-point scale). As a measure of direction, we also calculated a bias score, as the non-absolute difference between global predictions and average obtained credit.

2.7.2 | Calibration accuracy without standards present

For each text, calibration accuracy without standards present was calculated as the absolute difference between postdictions without standards present and actual obtained credit, averaged over the four definitions. We also calculated bias scores, by calculating the (non-absolute) difference between post-dictions with standards present and actual obtained credit (cf. Dunlosky & Thiede, 2013; Schraw, 2009).

2.7.3 | Calibration accuracy with standards present

Calibration accuracy with standards present could only be calculated for the standard group. We did so by calculating the absolute difference between postdictions with standards present and actual obtained credit, averaged over the four definitions. Again, bias scores were calculated by taking the (non-absolute difference) between postdictions without standards present and actual obtained credit (cf. Dunlosky & Thiede, 2013; Schraw, 2009).

2.8 | Procedure

With the exception of receiving standards or not, the procedure for the two experimental groups was the same and is depicted in Figure 1. All students sat behind a computer and were tested individually. They were informed that they had to read several texts (one practice text, six critical text) and had to memorize the key definitions in each text. The critical texts were presented in random order. First, students were instructed to read the practice text (about different measurement scales: nominal, ordinal, interval, and ratio) and made a practice test (i.e., recalling the definitions and providing performance estimates) to get comfortable with the materials and procedure. When students thought they were ready, they could continue with the critical texts. After each text, students could click "continue" when they thought they were done studying. Immediately after doing so, they were asked to make a global prediction and then continued with the recall test. The four key terms were presented one-by-one in a random order, and students were asked to recall their definition. After recalling a definition, students had to provide a postdiction without standard present before they could continue to the next key term.

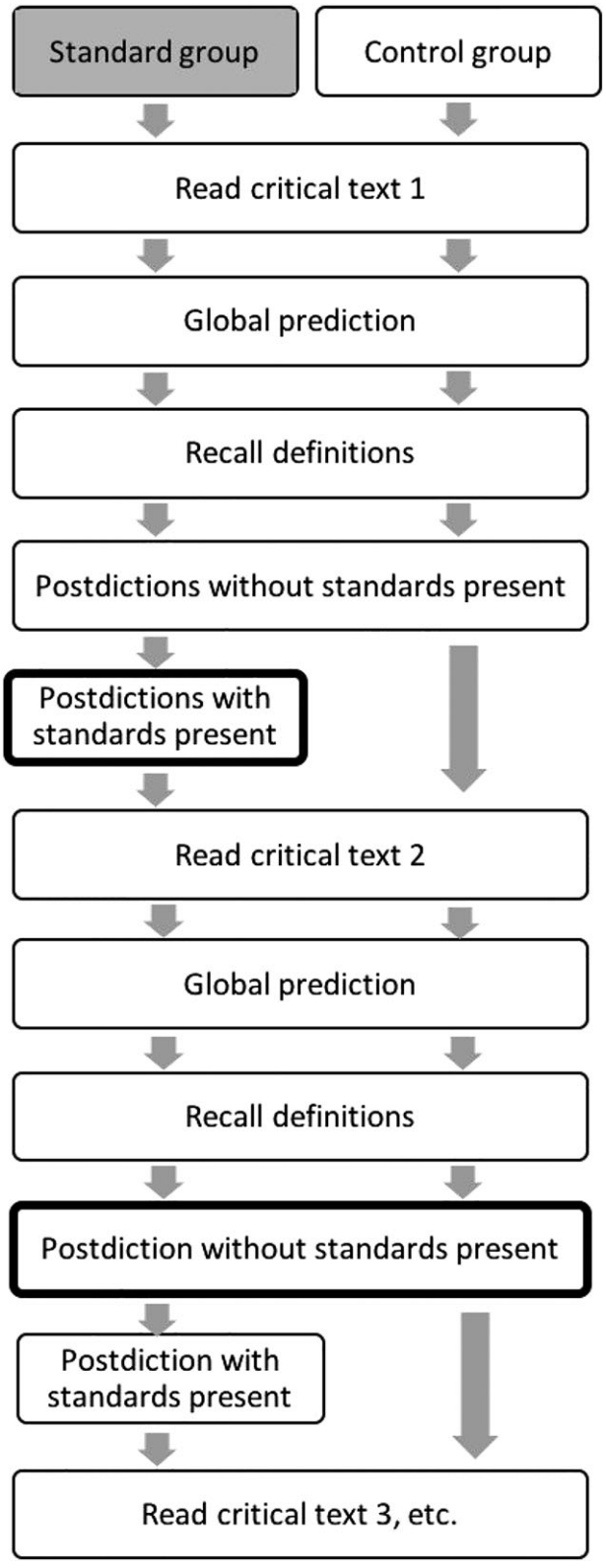

FIGURE 1 A graphical display of the experimental procedure

When students in the no-standard group had recalled the four definitions and provided their estimates, they continued with reading the next text. Students in the standard group, however, first received performance standards of the four key terms, to compare with their

recalled definitions, and provided a postdiction with standard present for each definition. Students in the standard group then also continued with the next text. After following this procedure for all six texts, students finished the experiment. On average, the experiment took about an hour.

Our procedure differs in two ways from that of Rawson and Dunlosky (2007). First, students in our standard group also provided postdictions when standards were not available yet. Note that in the study of Rawson and Dunlosky, the aim was to investigate whether providing standards while estimating performance would improve calibration accuracy. Therefore, Rawson and Dunlosky compared postdictions without standards present of the no-standard group with the postdictions with standards present of the standard-group. In our study, we also aimed to investigate the effect of standards on calibration accuracy on subsequent, similar tasks. Therefore, we included the postdictions without standards present in the standard group. A second difference between our procedure and that of Rawson and Dunlosky is that in their study, students had to complete a final test, in which the definitions students had learned and recalled during the experiment again had to be recalled. To answer our research questions however, there was no need for such an extra test because we focused on the possible learning effect of how well students were able to estimate their performance instead of direct improvements of (final) test performance.

3 | RESULTS

In all our analyses, a significance level of .05 was used. It is important to note that ideally, scores on calibration accuracy are zero—there should be no mismatch between estimated performance and actual performance. So, the lower the scores on calibration, the better the calibration accuracy is.

3.1 | Calibration accuracy with versus without standards present

We first examined whether we could replicate the positive impact of providing standards on calibration accuracy while estimating performance (cf. Rawson & Dunlosky, 2007) and whether students' performance level influenced this effect. To do so, we compared the mean calibration accuracy with standards of the standards group to the calibration accuracy without standards of the no-standard group over all six critical texts (see also Figure 1). We ran a two-way analysis of variance (ANOVA), with standards (Yes vs. No) and performance level (Low vs. Medium vs. High) as independent variables, and calibration accuracy on the six critical texts as the dependent variable. Our analysis showed that students who received standards while estimating their performance were better calibrated (M = .19, SD = .08) than students who did not receive standards while estimating their performance (M = .28, SD = .09), $F(116) = 44.96$, $p < .001$, $\eta^2 = .221$, replicating the findings of Rawson and Dunlosky (2007).

Second, we explored whether low and high performers would benefit equally from receiving standards. We found a nonsignificant interaction effect between standards and performance level, $F(116) = 1.13$, $p = .325$, $\eta^2 = .011$, indicating that low, medium, and high performers benefitted equally from receiving standards. Results did show a main effect of performance level however. Calibration accuracy of high, medium, and low performers differed significantly, $F(116) = 19.73$, $p < .001$, $\eta^2 = .195$. Follow-up pairwise comparisons showed that medium performers (M = .23, SD = .08) calibrated better than low performers (M = .28, SD = .10), $p = .003$ and that high performers (M = .18, SD = .07) calibrated better than both low and medium performers, $p < .001$ and $p = .002$,a respectively. So, no matter whether students received standards or not, the calibration accuracy of high performers was the highest, followed by the medium performers, and the calibration accuracy of low performers was the worst.

When analyzing bias scores, results showed a main effect of standard group $F(116) = 10.67$, $p = .001$, $\eta^2 = .084$. Students in the standard group showed less bias than students in the control group (M = .06, SD = .11 and M = .13, SD = .16, respectively). Furthermore, results showed a main effect of performance level $F(116) = 21.51$, $p < .001$, $\eta^2 = .271$. Low performers showed the most bias (M = .17, SD = .14), followed by medium performers (M = .08, SD = .12) and high performers, showed a negligible bias (M < .01, SD = .10). There was no significant interaction between standards and performance level $F(116) = 1.37$, $p = .259$, $\eta^2 = .023$.

3.2 | Effect of standards on calibration accuracy on subsequent tasks

To investigate whether providing standards improved calibration accuracy on subsequent tasks when standards were not available anymore, we ran a two-way ANOVA, with standards (Yes vs. No) and performance level (Low vs. Medium vs. High) as independent variables and calibration accuracy without standards present on five critical texts as the dependent variable (see Table 2 for descriptives). Note that on the first text, students in the standard group had not received any standards yet before providing their postdiction without standards present. We therefore excluded the calibration score of the first critical text from our analysis.

Our results showed a main effect of providing standards, $F(116) = 7.17$, $p = .008$, $\eta^2 = .043$. Students in the standard group calibrated more accurately on subsequent tasks without standards present than students in the no-standard group (see also Figure 2). Our results also showed a main effect of performance level, $F(116) = 20.56$, $p < .001$, $\eta^2 = .195$. Follow-up t tests showed that medium performers calibrated better on subsequent tasks than low performers $t(80.95) = 2.51$, $p = .014$, $d = .53$ and that high performers calibrated better than medium performers $t(72) = 4.17$, $p < .001$, $d = .97$, meaning that they also calibrated better than low performers by transitive property. Figure 2 shows that low performers seem to have benefitted the most from receiving standards, followed by medium performers,

TABLE 2 Calibration accuracy without standard present

	Standards								
	No			Yes			Total		
Performance level	*N*	*M* (*SE*)	95% *CI*	*N*	*M* (*SE*)	95% *CI*	*N*	*M* (*SE*)	95% *CI*
Low	24	.34 (.02)	[.30, .38]	24	.27 (.02)	[.24, .31]	48	.31 (.01)	[.28, .33]
Medium	17	.28 (.01)	[.25, .31]	21	.25 (.01)	[.21, .28]	38	.26 (.01)	[.24, .28]
High	19	.20 (.02)	[.16, .24]	17	.19 (.02)	[.15, .22]	36	.19 (.01)	[.17, .22]
Total	60	.28 (.01)	[.25, .31]	62	.24 (.01)	[.22, .26]	122	.26 (.01)	[.24, .28]

Note. This table displays scores of calibration accuracy without standards present. Students scoring below the 33th percentile belong to the group of low performers. Medium performers are students who scored between the 33th and 66th percentile. Finally, students scoring above the 66th percentile belong to the last group: high performers. Calibration accuracy scores without standards present are shown from Text 2 till Text 6.

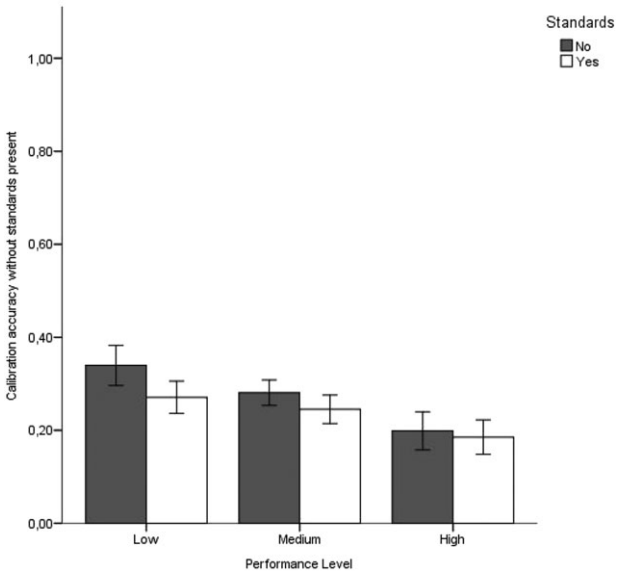

FIGURE 2 This graph displays the effects of standards and performance level on calibration accuracy without standards present (i.e., calibration accuracy on subsequent tasks) ranging from 0 to 1 (note that the lower the score, the better the match between estimated performance and actual performance)

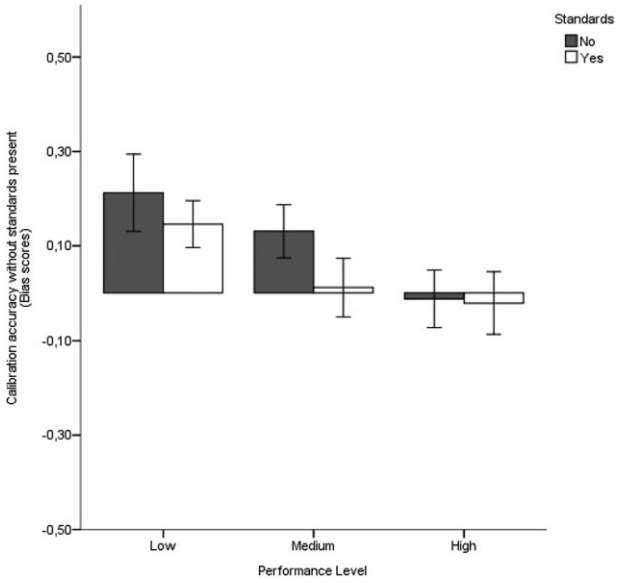

FIGURE 3 This graph displays the effects of standards and performance level on the bias scores (from −1 to +1) of calibration accuracy without standards present (i.e., calibration accuracy on subsequent tasks). Note that the closer to zero, the better the match between estimated performance and actual performance

whereas high performers do not appear to benefit much. However, the interaction effect between performance level and standards was not statistically significant, $F(116) = 1.27$, $p = .285$, $\eta^2 = .015$.

Figure 3 shows the bias scores of all performance level groups. Results showed a main effect of standard group $F(116) = 6.35$, $p = .013$, $\eta^2 = .052$. Students in the standard group ($M = .05$, $SD = .14$) showed less bias than students in the control group ($M = .12$, $SD = .18$). Results also showed a main effect of performance level $F(116) = 20.21$, $p < .001$, $\eta^2 = .258$ following a similar pattern as with calibration accuracy with standards present. Low performers were biased the most ($M = .18$, $SD = .16$), followed by medium performers ($M = .07$, $SD = .14$), and finally, high performers showed the least bias ($M = -.02$, $SD = .13$).

Again, when looking at Figure 3, there appears to be an interaction. Both low and medium performers seem to decrease in overconfidence when receiving standards, whereas high performers do not seem to change in bias scores. However, again the interaction between standards and performance level was not significant, $F(116) = 1.41$, $p = .248$, $\eta^2 = .024$.

To further explore the effect of standards on calibration accuracy on new tasks, we looked at the improvement of calibration accuracy over texts. Figure 4 shows that in the standard condition, calibration accuracy seems to improve linearly, whereas in the no-standard condition, calibration accuracy seems to remain more or less equal. To test this interaction pattern, we used a mixed-design ANOVA, with text (Text 1 until 6) and standards (Yes vs. No) as independent variables and calibration accuracy without standards present as the dependent variable. The within-subject contrast showed, however, no significant linear interaction effect between text and standards, $F(116) = 3.27$, $p = .073$, $\eta^2 = .025$.

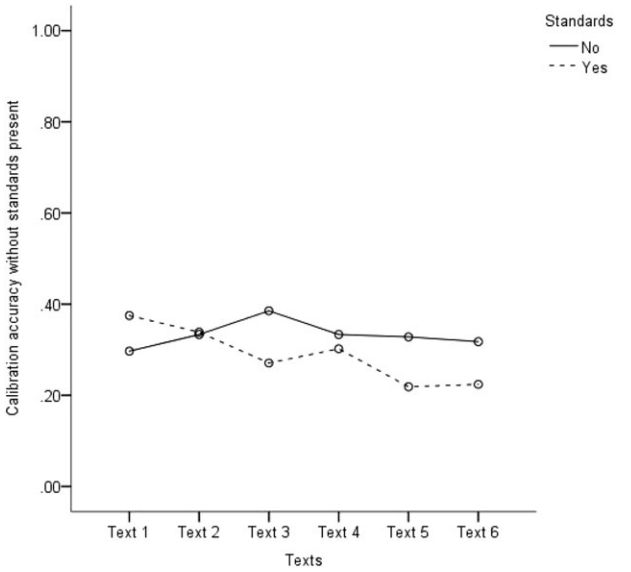

FIGURE 4 This graph displays the effect of standards on the calibration accuracy without standards present (ranging from 0 to 1) over texts (note that the closer the score is to zero, the better the match between estimated performance and actual performance)

3.3 | Effect of standards and performance level on predictions

Finally, although the measure of global predictions was not central to our hypotheses, we still analyzed the effect of standards on students' global prediction accuracy for archival purposes. We ran a two-way ANOVA, with standards (Yes vs. No) and performance level (Low vs. Medium vs. High) as independent variables and global prediction accuracy on five critical texts as the dependent variable. We excluded the prediction of the first critical text from our analysis, because students in the standard group had not yet received any standards at that time yet.

Our results did not show main effects of standards, $F(116) = 0.139$, $p = .710$, $\eta_p^2 = .001$, nor of performance level, $F(116) = 1.12$, $p = .328$, $\eta_p^2 = .019$. We did find a significant interaction effect however, $F(116) = 5.55$, $p = .005$, $\eta_p^2 = .087$. Follow-up t tests showed that low performers in the standard group predicted their global performance better ($M = .20$, $SD = .07$) than low performers in the no-standard group ($M = .27$, $SD = .10$), $t(46) = 2.51$, $p = .016$, $d = .72$. Interestingly, however, medium performers receiving standards predicted their own performance worse ($M = .24$, $SD = .08$) than medium performers who did not receive standards ($M = .18$, $SD = .05$), $t(36) = -2.69$, $p = .011$, $d = .90$. Prediction accuracy of high performers who received standards ($M = .23$, $SD = .14$) did not differ from prediction accuracy of high performers in the no-standards group ($M = .21$, $SD = .06$), $t(34) = -0.61$, $p = .545$, $d = 0.20$.

We also examined the prediction bias scores of our intervention groups by using an ANOVA with prediction bias scores as dependent variable and performance level and standards as independent variable. Results showed that standards significantly influence students' bias scores, $F(116) = 15.59$, $p = < .001$. Students receiving standard were more underconfident ($M = -.08$, $SD = .18$) than students not receiving standards ($M = .03$, $SD = .19$). This means that, by receiving standards, students seem to lower their performance estimates. Consequently, when looking at prediction bias scores for low performers, we see that low performers in the no-standard group showed overconfidence ($M = .16$, $SD = .19$), whereas students in the standard group lowered their overconfidence and showed negligible bias ($M = .05$, $SD = .12$). However, when students show negligible bias already, as is the case for medium performers ($M = .01$, $SD = .13$), lowering their judgements after receiving standards leads to underconfidence ($M = -.16$, $SD = .14$). Consequently, we think that standards made medium performers too cautious, causing underconfidence and worse calibration accuracy, as seen in the previous paragraph. Again, as with our prior tests on calibration accuracy, high performers were not affected by the standards. Although descriptives showed that high performers became somewhat more underconfident when receiving standards ($M = -.17$, $SD = .18$) than without standards ($M = -.13$, $SD = .12$), this difference did not reach statistical significance, $t(34) = .73$, $p = .468$.

4 | DISCUSSION

In this study, we investigated whether students can learn to calibrate better by receiving standards. We hypothesized that providing standards while students made a performance estimate would improve their calibration accuracy (cf. Rawson & Dunlosky, 2007). We also explored whether high performers would benefit more from receiving standards than low performers. Furthermore, we investigated whether providing standards could improve calibration accuracy on similar, subsequent tasks when these standards were not immediately available, and we explored whether this was the case for both high and low performing students.

4.1 | Calibration accuracy with standards present

We investigated whether providing students with standards would enhance calibration accuracy as Rawson and Dunlosky (2007) found. Our results indeed show that the calibration accuracy of students who receive standards while estimating performance is better than the calibration accuracy of students who do not receive such standards. Our results thus support the positive effect of standards on calibration, as shown in previous studies (Dunlosky et al., 2011; Dunlosky & Thiede, 2013; Lipko et al., 2009; Rawson & Dunlosky, 2007) and are in line with findings of Koriat (1997) that students experience difficulties to estimate their own performance when standards (i.e., valid cues) are unavailable.

Additional to discussing the absence of standard hypothesis, Rawson and Dunlosky (2007) stated that students are limited in their competence to use standards. They did not, however, specify whether some students may be more limited than others. In our study, we explored whether performance level would influence the effect of

standards. On the one hand, low performers may fail to benefit from receiving standards because they understand these standards less well than high performers. On the other hand, low performers have more room for improvement as shown by their poor calibration (e.g., Ehrlinger et al., 2008; Kruger & Dunning, 1999). These low performers could therefore especially benefit from receiving standards (i.e., more valid cues) when estimating their performance. Our results show that both high and low performers improve their calibration accuracy after receiving standards—refuting the hypothesis that low performers are less able to adequately use standards. These are promising findings because it means that providing students with a standard will help them become better calibrated, regardless of their initial performance level.

4.2 | Performance standards and calibration accuracy on subsequent tasks

Knowing that students calibrate better when a standard is present is a first important step. However, until now, it has been unclear whether standards also help students to better calibrate on similar tasks without receiving standards. Although theory (Koriat, 1997; Zimmerman, 2000) and previous studies gave rise to such an assumption (Hacker et al., 2000; Nietfeld et al., 2006), this effect had not been investigated before in a controlled laboratory experiment.

Our results show that providing students with standards can indeed improve calibration accuracy on new, subsequent tasks when a standard is not available. Students that have read a text, and made an estimate of their recall performance based on a standard, seem to learn from this experience. On the recall task from the next text, these students also provide a more accurate performance estimate, even though this text is about a different topic than the previous one, and the students have not (yet) received any standard when estimating their performance. A possible explanation for this finding can be found in the cue utilization model of Koriat (1997). Providing students with standards and asking them to give a performance estimate allows them to compare this estimate with their original performance estimate, given without a standard. This gives the students extra help in the form of a valid cue about the quality of their original estimate. This cue can, in turn, help them improve their calibration accuracy on subsequent tasks (Koriat, 1997; Zimmerman, 2000). This study therefore is one of the first to show that the beneficial effect of standards on calibration accuracy also transfers to subsequent similar tasks.

Interestingly, the interaction between performance level and standards did not reach statistical significance. Our results thus conflict with the hypothesis that low performers would benefit less from performance feedback than high performers (Hacker et al., 2000; Nietfeld et al., 2006) and run counter to the "Matthew effect" that high performers would actually benefit the most compared with low performers (Merton, 1968; Otto & Kistner, 2017). In fact, when looking at Figures 2 and 3, low performers even seem to show the strongest improvement in their calibration accuracy and a decreased bias. However, this interaction pattern was not statistically significant, so further research with even a more powerful design is needed to determine whether this conclusion is warranted or not. In sum, our results look promising, by showing that students—including those in need of an intervention to improve their calibration accuracy—do actually benefit from receiving standards.

4.3 | Limitations and future directions

Although our experiment provides valuable insights in the role of performance level and standards on calibration accuracy, it also had some limitations. As Nelson and Narens (1990) discussed, there are many types of judgements students can make when estimating their performance, and studies focusing on the match between estimated performance and actual performance use different types of judgements. For example, some researchers focus on Judgements of Learning or predictions, before completing a task (e.g., Foster, Was, Dunlosky, & Isaacson, 2017), whereas others focus on postdictions, after completing a task (e.g., Nietfeld et al., 2006). It is important to stress that interventions aimed at improving postdictions (i.e., estimates after completing a task) cannot always be generalized to other types of judgements, such as predictions (i.e., estimates before completing a task) and vice versa. For example, although previous studies found that postdictions can be improved, a recent study by Foster et al. (2017) showed that even after 13 exams, students were unable to predict their next exam grade. Indeed, our results show that although standards improve postdiction accuracy, the effects are different when correcting prediction accuracy—medium and high performers started underestimating themselves when receiving standards. This result is also shown in a study by De Bruin et al. (2017): whereas low performers benefitted from extra feedback, high performers became more underconfident. In addition, such findings underscore the importance of including performance level as a variable when studying interventions to improve calibration accuracy: high and low performers may not always benefit the same way.

Our study also shows that even simple forms of standards can already help to enhance calibration accuracy. It must be noted, however, that the standards used are a limited form of feedback. For example, students do not see how they should have scored their answer. Especially, low performers might benefit from such extra guidance as they struggle the most with estimating their performance. A suggestion for future research would therefore be to use more extended types of feedback that not only let students compare their own answer to the correct answer but also show them how they should have scored their own definitions. A type of standard that could offer this extra guidance could be the idea-unit standards used by Dunlosky et al. (2011). In such an idea-unit standard, all elements of the standard that have to be present to receive full credit are specifically defined.

Furthermore, although low performers appear to benefit at least as much as high performers from receiving standards when postdicting

their performance, they do not become calibrated equally well. Our results show that overall, high performers remain significantly better calibrated than low performers when receiving standards (i.e., low performers make more mistakes comparing their own answer to the correct answer). It is possible that high performers were better at judging whether their own recalled definitions matched the standards or not, because they were more able to identify the critical elements that should have been present to receive credit. Future research could investigate whether providing students with extra guidance how to use standards—such as when providing full definition standards with idea units (i.e., all critical elements a definition consists of are specified, Dunlosky et al., 2011)—diminishes the difference in calibration accuracy between low and high performers (i.e., mistakes due to misunderstanding are minimalized).

Another direction for future research regards the number of estimates that are made by the students. In the current study, students in the standard group provided an extra estimate compared with students in the no-standard group. It is possible that making such an additional estimate would have impacted their calibration accuracy. To the knowledge of the authors, however, no studies have shown that estimating performance without receiving feedback leads to enhanced calibration accuracy (Bol et al., 2005; Foster et al., 2017; Lipko, Dunlosky, & Merriman, 2009). Hence, although we encourage future research on this topic, we deem it unlikely that the number of judgements provided by our students could explain our findings.

A final remark is that good monitoring alone is not sufficient to improve performance. Students should also use the monitoring to control their learning by, for example, rereading or selecting better learning strategies (Butler & Winne, 1995; Fernandez & Jamet, 2017; Koriat, 2012; Metcalfe, 2009; Nelson & Narens, 1990; Tuysuzoglu & Greene, 2015). If students use better control strategies, this should help them to gain more content knowledge, which will eventually be reflected in better task performance. Interestingly, the data of our study already seem to indicate that providing standards leads to better performance. Note that there were no a priori performance differences on the first critical text (after the practice text) between students in the standard group (M = .59, SD = .24) and no-standard group (M = .58, SD = .26), $t(120) = -0.12$, p = .905. However, we made a comparison of average task performance on the five following critical texts between students that did not receive standards versus students who did receive standards. To do so, we ran an ANOVA with calibration without standards present on the five texts as dependent variable and standards as independent variable. Results show a main effect of standards on task performance, $F(116) = 24.16$, $p < .001$, $\eta_p^2 = .172$. So, it seems that only after receiving standards on Text 1, students in the standard group started to perform better. Future research could complement our findings by investigating in more detail if, and how, standards can influence subsequent study behavior. When doing so, it may be informative to take cognitive load into account as well, as research suggests that this could interfere with monitoring and improvement of performance (Raaijmakers, Baars, Paas, Van Merriënboer, & Van Gog, 2018; Van Gog, Kester, & Paas, 2011).

4.4 | Conclusion

Our study is the one of the first to investigate the role of performance level when students receive standards to improve their calibration accuracy on textual recall tasks. We have shown that providing standards improves calibration accuracy for all performance levels—although low performers show more miscalibration than high performers, both when receiving and not receiving standards. Furthermore, it is the first study to show that providing standards can also improve calibration accuracy on subsequent tasks. This is a promising finding that has implications for both theory and educational practice.

ACKNOWLEDGEMENTS

This research was funded by a Research Excellence Initiative grant from Erasmus University Rotterdam awarded to the Educational Psychology section. We would like to thank all participants for their participation in our study. Many thanks to Anique de Bruin for providing us with the translated version of the materials of Rawson and Dunlosky (2007).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ORCID

Marloes L. Nederhand https://orcid.org/0000-0001-7388-6381
Remy M.J.P. Rikers https://orcid.org/0000-0002-4722-1455

REFERENCES

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, *24*(2013), 1–3. https://doi.org/10.1016/j.learninstruc.2012.10.003

Baker, J. M. C., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review*, *13*(1), 60–65. https://doi.org/10.3758/BF03193813

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. *Journal of Experimental Education*, *73*(4), 269–290. https://doi.org/10.3200/JEXE.73.4.269-290

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

De Bruin, A. B. H., Kok, E., 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*(1), 21–43. https://doi.org/10.1007/s11409-016-9159-5

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology*, *64*(3), 467–484. https://doi.org/10.1080/17470218.2010.502239

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

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

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, *24*, 58–61. https://doi.org/10.1016/j.learninstruc.2012.05.002

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*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Fernandez, J., & Jamet, E. (2017). Extending the testing effect to self-regulated learning. *Metacognition and Learning*, *12*(2), 131–156. https://doi.org/10.1007/s11409-016-9163-9

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). 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), 1–19. https://doi.org/10.1007/s11409-016-9158-6

Gutierrez de Blume, A. P., Wells, P., Davis, A. C., & Parker, J. (2017). "You can sort of feel it": Exploring metacognition and the feeling of knowing among undergraduate students. *The Qualitative Report*, *22*(7), 2017–2032. Retrieved from http://nsuworks.nova.edu/tqr

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*(1), 160–170. https://doi.org/10.1037//0022-0663.92.1.160

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction*, *22*(4), 296–298. https://doi.org/10.1016/j.learninstruc.2012.01.002

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. https://doi.org/10.3758/BF03194055

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning*, *5*(2), 173–194. https://doi.org/10.1007/s11409-010-9056-2

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.023

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology*, *103*(2), 152–166. https://doi.org/10.1016/j.jecp.2008.10.002

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology. Applied*, *15*(4), 307–318. https://doi.org/10.1037/a0017599

Merton, R. K. (1968). The Matthew effect in science: The rewaed and communication systems are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning*, *6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*(26), 125–141. https://doi.org/10.1016/S0079-7421(08)60053-5

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, *1*(2), 159–179. https://doi.org/10.1007/s10409-006-9595-6

Otto, B., & Kistner, S. (2017). Is there a Matthew effect in self-regulated learning and mathematical strategy application?—Assessing the effects of a training program with standardized learning diaries. *Learning and Individual Differences*, *55*, 75–86. https://doi.org/10.1016/J.LINDIF.2017.03.005

Raaijmakers, S. F., Baars, M., Paas, F., Van Merriënboer, J. J. G., & Van Gog, T. (2018). Training self-assessment and task-selection skills to foster self-regulated learning: Do trained skills transfer across domains? *Applied Cognitive Psychology*, *32*(2), 270–277. https://doi.org/10.1002/acp.3392

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*(4–5), 559–579. https://doi.org/10.1080/09541440701326022

Schraw, G. (2009). Measuring metacognitive judgements. In *Handbook of Metacognition in Education* (pp. 439–462). https://doi.org/10.4324/9780203876428

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review*, *12*(4), 437–475. https://doi.org/10.1023/A:1009084430926

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*(4), 331–362. https://doi.org/10.1080/01638530902959927

Tuysuzoglu, B. B., & Greene, J. A. (2015). An investigation of the role of contingent metacognitive behavior in self-regulated learning. *Metacognition and Learning*, *10*(1), 77–98. https://doi.org/10.1007/s11409-014-9126-y

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology*, *25*(4), 584–587. https://doi.org/10.1002/acp.1726

Van Loon, M. H., & Roebers, C. M. (2017). Effects of feedback on self-evaluations and self-regulation in elementary school. *Applied Cognitive Psychology*, *31*(5), 508–519. https://doi.org/10.1002/acp.3347

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–40). Cambridge, MA: Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

How to cite this article: Nederhand ML, Tabbers HK, Rikers RMJP. Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Appl Cognit Psychol*. 2019;1–12. https://doi.org/10.1002/acp.3548

APPENDIX A

EXAMPLE TEXT

Gestures

Scholars who have studied body language extensively have devised a widely used system to classify the function of gestures that people use when speaking publicly. EMBLEMS are gestures that stand for words or ideas. You occasionally use them in public speaking, as when you hold up your hand to cut off applause. Emblems vary from culture to culture. The sign that stands for "a-ok" in this country refers to money in Japan, and it is an obscene gesture in some Latin American countries. ILLUSTRATORS are gestures that simply illustrate or add emphasis to your words. For example, speakers often pound on a podium to accent words or phrases. In addition, you can illustrate spatial relationships by pointing or by extending your hands to indicate width or height. Adaptors are a different group of gestures used to satisfy physical or psychological needs. SELF-ADAPTORS are those in which you touch yourself in order to release stress. If you fidget with your hair, scratch your face, or tap your leg during a speech, you are adapting to stress by using a self-adaptor. You use object-adaptors when you play with your keys, twirl a ring, jingle change in your pocket, or tap pencils and note cards. Finally, ALTER-ADAPTORS are gestures you use in relation to the audience to protect yourself. For instance, if you fold your arms across your chest during intense questioning, you may be subconsciously protecting yourself against the perceived psychological threat of the questioner. Whereas emblems and illustrators can be effective additions to a speech, adaptors indicate anxiety and appear as nervous mannerisms and should therefore be eliminated from public speaking habits.

APPENDIX B

AVERAGE POSTDICTIONS PER RECALL RESPONSE CATEGORY (CF. RAWSON & DUNLOSKY, 2007)