



Improving calibration over texts by providing standards both with and without idea-units

Marloes L. Nederhand, Huib K. Tabbers, Homaira Abrahimi & Remy M. J. P. Rikers

To cite this article: Marloes L. Nederhand, Huib K. Tabbers, Homaira Abrahimi & Remy M. J. P. Rikers (2018): Improving calibration over texts by providing standards both with and without idea-units, *Journal of Cognitive Psychology*, DOI: [10.1080/20445911.2018.1513005](https://doi.org/10.1080/20445911.2018.1513005)

To link to this article: <https://doi.org/10.1080/20445911.2018.1513005>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 21 Aug 2018.



[Submit your article to this journal](#)




Article views: 47



[View Crossmark data](#)

Improving calibration over texts by providing standards both with and without idea-units

Marloes L. Nederhand ^a, Huib K. Tabbers^a, Homaira Abrahimi^a and Remy M. J. P. Rikers^{a,b}

^aInstitute of Psychology, Erasmus University Rotterdam, Rotterdam, Netherlands; ^bRoosevelt Center for Excellence in Education, University College Roosevelt, Utrecht University, Middelburg, Netherlands

ABSTRACT

This study aims at improving calibration accuracy, which is the match between estimated performance and actual performance. In our experiment, one hundred and twenty-seven university students read texts and learned definitions. The students recalled these definitions during a test and made performance judgements. After recalling their definitions half of the students received full-definition standards, stating what the correct definition should have been. The other half of the students received idea-unit standards: The correct definition was parsed into units that had to be present. Providing standards improved calibration accuracy not only on current texts, but also on new, subsequent texts. Especially the calibration of low performing students benefitted from receiving both idea-unit and full-definition standards. Furthermore, over multiple texts, students who received idea-unit standards benefitted more than students receiving full-definition standards. This study is among the first to show the effect of standards on calibration on new texts and underscores the importance of self-testing.

ARTICLE HISTORY

Received 16 October 2017
Accepted 12 August 2018

KEYWORDS

Calibration; feedback; standards; self-assessment; metacognition; performance level

Introduction


Each course, students are confronted with vast amounts of information. Hence, they should strive for effective and durable ways of learning. To foster such learning, students must be able to estimate at what point their understanding of the course material is sufficient. When students are unable to accurately estimate their own performance, due to overestimation or underestimation, we speak of miscalibration; they show a mismatch between estimated performance and actual performance (Alexander, 2013; Lichtenstein, Fischhoff, & Phillips, 1982). Miscalibration is a widely acknowledged phenomenon, and especially prominent among low performers (Dunlosky & Rawson, 2012; Grimaldi & Karpicke, 2014; Kruger & Dunning, 1999; Rawson & Dunlosky, 2007). Because calibration influences control decisions made during learning, miscalibration causes problems both for overconfident and underconfident students. While overconfident students may assign too little time to study less-

well-known material, underconfident students may have difficulty disengaging from studying material they already mastered (Bol, Hacker, O'Shea, & Allen, 2005; Dunlosky & Rawson, 2012).

According to the cue-utilisation view of Koriati (1997), students use a variety of cues when estimating their own performance, and calibration accuracy depends on the predictive validity of the cues used. To improve calibration accuracy, an intervention should help students using better, more valid cues when estimating their performance. An effective way to do so is by giving students the opportunity to compare their own answers to a standard (i.e. the correct solution; Dunlosky & Thiede, 2013; Dunlosky, Hartwig, Rawson, & Lipko, 2011; Lipko et al., 2009; Rawson & Dunlosky, 2007). Such a standard serves as an informative cue about students' performance because it gives them insight in whether their own answer matched with the desired response.

One prominent experiment demonstrating the effect of standards on calibration accuracy was

CONTACT Marloes L. Nederhand  m.l.nederhand@essb.eur.nl

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/20445911.2018.1513005>.

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

conducted by Rawson and Dunlosky (2007). In their study, students had to read several texts in which keywords were explained. After reading the texts, students continued with a test and had to recall the definition belonging to each keyword. Half of the students received standards while estimating their performance, and were thus able to directly compare their own recall response to the correct answer. Conversely, the other half of the students had to estimate their performance without any standard present. Results showed that students that received a standard were better calibrated than students that did not receive any standards. This is a promising finding because standards are often used during self-testing: A popular learning strategy among students (Hartwig & Dunlosky, 2012).

Existing research predominantly focuses on providing standards *while* students estimate their performance: For each recall attempt made by the students, standards are present (Dunlosky et al., 2011; Dunlosky & Thiede, 2013; Lipko et al., 2009; Rawson & Dunlosky, 2007). However, this leaves the question unanswered whether comparing own answers to standards indeed teaches students to calibrate better, in such a way that these students will also show better calibration on new, subsequent texts that are similar in nature but different in content (e.g. when encountering new definitions about different topics).

It could be argued that students may indeed show better calibration accuracy on new texts after receiving standards. After receiving standards in a study by Rawson and Dunlosky (2007), students did not only receive a cue on their performance (did it match or mismatch the desired answer?), but they could also generate a cue on their calibration accuracy (did their initial performance estimate match the outcome as scored with the standard present?). In turn, students could use both cues to make better judgements on new, subsequent texts. For example, when students become aware that they generally are overconfident, they may lower their estimates. Although this may seem intuitively plausible, empirical findings that support this assumption are scarce.

A recent study by Nederhand, Tabbers, and Rikers (2017) aimed to bridge this gap in the literature by analysing whether comparing own answers to standards indeed improves calibration accuracy on new, subsequent texts where standards were not immediately available. To investigate this, Nederhand et al. conducted an experiment similar to Rawson and

Dunlosky's (2007) study in which students had to read several texts outlining specific keywords. In a test, students had to recall the definition of each keyword and were requested to estimate their level of performance. Half of the students received standards while making their estimation—they could directly compare their own recall to the correct definition, and then decide on the credit of their answer. Results showed, in line with Rawson and Dunlosky (2007) that directly comparing an answer to a standard led to better calibration accuracy: The credit given by the student better matched actual obtained credit. More importantly, Nederhand et al. also found that by receiving standards, students became better calibrated on subsequent texts about a different topic, where standards were not yet present. Results even showed a trend of a learning curve: The more standards received, the better the calibration accuracy became.

When improving calibration accuracy by providing standards, however, the type of standard can matter. The full-definition standards that were used in the studies by Rawson and Dunlosky (2007) and Nederhand et al. (2017) can be seen as a non-elaborated form of feedback because such standards only provide students with the correct answer. In other words, no explanation is provided as to *why* the answer is correct. For example, the standard for the definition of *proactive interference* was “proactive interference is when information stored in memory interferes with learning of new information”. In this case, students were free to compare their own answer to this standard and were not given any guidance about how they should do so. As a result, many students still made mistakes comparing their answer to the standards (Rawson & Dunlosky, 2007). In an attempt to reduce these comparison mistakes, Dunlosky et al. (2011) decided to provide more detailed standards in which students were given idea-units, signalling elements that had to be present in the recall response to obtain full credit. For example, the definition of “proactive interference” was presented in the following way: “Proactive interference is when (1) information stored in memory (2) interferes with learning (3) of new information”. Hence, when receiving idea-unit standards, students received a second cue: They learned about the criterion that would be used to score the recall response. Knowledge about this criterion helped students to better score the response, and by doing so, allowed them to obtain more valid insight in whether their

performance was accurate or not. Dunlosky et al. (2011) showed that, compared to full-definition standards, receiving idea-unit standards led indeed to better calibration among students.

According to Koriat (1997), calibration accuracy depends on the cues that are used when estimating one's own performance. The better and more valid the cues that students use, the better their calibration accuracy will be. Hence, following the reasoning that adequate cue use leads to better calibration, providing students with extra detailed standards, such as idea-unit standards, should further help students to improve their calibration—both on the current text where the standards are immediately present, as well as on subsequent texts.

Present study

With the present experiment, we investigated whether students improved their calibration on a subsequent text when receiving either full-definition or idea-unit standards. We defined the following hypotheses:

- (1) Calibration accuracy with standards present will be better when idea-unit standards are provided than full-definition standards (cf. Dunlosky et al., 2011).
- (2) Full-definition standards will help students improve their calibration accuracy over subsequent texts (cf. Nederhand et al., 2017), but the improvement will be bigger when providing idea-units.

In our experiment, half of our students received full-definition standards, stating the correct definition. The second group of students received idea-unit standards, which specifically stated what elements should be present for a definition to be considered correct. Since previous research has shown that the performance level of students can influence the use of feedback and calibration accuracy (Hacker, Bol, Horgan, & Rakow, 2000; Nietfeld, Cao, & Osborne, 2006), we also explored the effect of recall performance in our study.

Method

Participants and design

One hundred and twenty-seven first-year psychology students participated in this study. The students

had a mean age of 19.76 ($SD = 2.71$) and 11.80% of the students reported to be male and 88.20% indicated to be female. Students received course credit for their participation and provided informed consent.

The experiment conformed to a 2 Idea-unit (Yes vs. No) \times 3 Performance level (Low vs. Average vs. High) design. Half of the students received full-definition standards ($N = 64$). The remaining students ($N = 63$) also received additional guidance how to use these standards (i.e. idea-units). Based on students' test performance (i.e. how many definitions were correctly recalled), we defined three performance level groups in each standard group to facilitate interpretation of our findings. This concerned a group of low performing students, with students scoring below the 33th percentile ($N = 42$); a group of average performing students, with students scoring between the 33th and 66th percentile ($N = 47$); and a group of high performing students with students scoring above the 66th percentile ($N = 38$). See Table 1 for descriptives of recall performance.

Materials

Texts

For our experiment, we used the materials of Rawson and Dunlosky (2007). These materials consisted of seven texts (six critical texts and one example text) from textbooks of undergraduate courses, on subjects such as communication and family studies. In each text, four key terms were presented in capital letters, followed by a definition of each keyword (e.g. EMBLEMS are gestures that represent words or ideas). All texts were translated into Dutch by De Bruin, Kok, Lobbestael, and De Grip (2017) and ranged between 273 and 303 words. See Appendix for an example text. The texts were specifically designed to be equal in difficulty. To check whether students indeed did not differ in their performance between the various texts, we conducted a univariate ANOVA in which we included Text order as independent variable and Recall performance as dependent variable. Indeed, there were no differences in performance accuracy between any of the texts $F(5, 756) = 0.25$, $p = 0.941$, $\eta_p^2 = 0.002$. Hence, all texts can be considered equal in difficulty.

Scoring

Students had to read each text and needed to recall the definitions belonging to each keyword. Student

Table 1. Recall performance scores.

Performance level	Idea-unit standards								
	No			Yes			Total		
	<i>N</i>	<i>M</i> (<i>SE</i>)	95% <i>CI</i>	<i>N</i>	<i>M</i> (<i>SE</i>)	95% <i>CI</i>	<i>N</i>	<i>M</i> (<i>SE</i>)	95% <i>CI</i>
Low	21	0.35 (0.02)	[0.32, 0.38]	21	0.37 (0.02)	[0.34, 0.40]	42	0.36 (0.01)	[0.33, 0.39]
Average	25	0.58 (0.02)	[0.55, 0.61]	22	0.54 (0.02)	[0.51, 0.57]	47	0.56 (0.01)	[0.54, 0.58]
High	18	0.79 (0.02)	[0.75, 0.82]	20	0.75 (0.02)	[0.72, 0.78]	38	0.77 (0.01)	[0.76, 0.79]
Total	64	0.56 (0.02)	[0.52, 0.61]	63	0.55 (0.02)	[0.50, 0.59]	127	0.63 (0.01)	[0.60, 0.65]

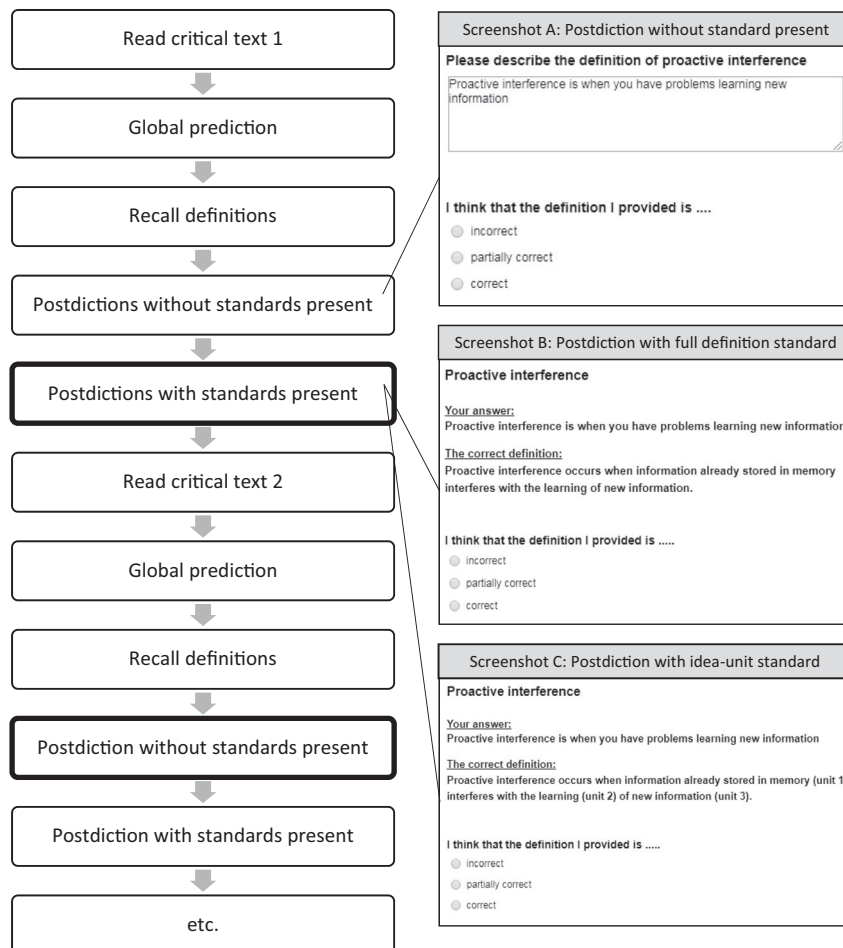
recall was scored with a scoring grid (cf. De Bruin et al., 2017). Recalled definitions could receive full (1 point), partial (0.5 point) or no credit (0 point), depending on the correct number of critical elements present. A random selection of the entire data set (11%) was scored by a second independent rater. The intraclass correlation for single measures was 0.86, with a 95% confidence interval from 0.81 to 0.90.

Procedure

The procedure within the two experimental conditions (i.e. full-definition standards and idea-unit

standards) was identical and is depicted in Figure 1. Using the online software application Qualtrics, a computer presented all materials and recorded the students' responses. As part of the experiment, students first were provided with an example text to familiarise themselves with the materials and the procedure. Subsequently, the six critical texts were presented one by one (please see Appendix for an example text). The order in which both the texts and the definitions within each text were presented was fully randomised.

When finishing reading each text, students made a global prediction by answering the following

**Figure 1.** The procedure of the current experiment, including screenshots from the post-diction estimates.

question: How well will you be able to complete a test on this material? Students provided their answer on a scale from 0 (definitely won't be able) to 10 (definitely will be able). This global prediction measure was included to follow the procedure of Dunlosky et al. (2011) and Nederhand et al. (2017) as closely as possible, but was not of further interest in the current study. Then, students continued with a recall test, in which one keyword was presented a time, and students had to provide the definition they considered belonged to this keyword. Since in each text, four key terms were presented, students had to recall four corresponding definitions. The key terms were presented in a random order. Immediately after recalling each definition, students provided a post-diction without standard present by indicating how much credit their answer should receive (no credit = 0 points; partial credit = 0.5 points; full credit = 1 point). In Figure 1, Screenshot A depicts the screen in which students typed their response and provided their performance judgement. After providing this post-diction, students received a standard. Standards were presented together with the recalled definition of the student to facilitate a comparison between the two. Depending on the experimental group students were in, either a full-definition standard (Figure 1, Screenshot B) or an idea-unit standard (Figure 1, Screenshot C) was provided. While comparing their own answer to the standard, all students provided a post-diction with standard present, by again scoring their answer's credit (no credit = 0 points; partial credit = 0.5 points; full credit = 1 point). This procedure was repeated for each of the six critical texts. In total, the experiment lasted for about an hour.

Analyses

Calculating measurements

Calibration accuracy. Two types of calibration accuracy were calculated. The first type, *calibration with standards present*, was calculated by the absolute difference between post-dictions with standards present and actual performance, as scored by the experimenter. The second type of calibration accuracy, *calibration without standards present*, was calculated by the absolute difference between post-dictions without standards present and actual performance as scored by the experimenter. Scores ranged from 0 to 1. Scores of 0 represented perfect calibration accuracy and scores of 1

represented a complete mismatch between estimated and actual performance. Each text involved the recalling of four definitions and thus the calculation of four calibration scores. Subsequently, we calculated a mean calibration score per text based on these four calibrations.

Bias scores. With regards to calibration with standards present, we also calculated bias scores to investigate the level of overconfidence or underconfidence. For each text, bias scores were calculated as the mean difference between estimated performance (i.e. post-dictions both with and without standards) and actual performance (Dunlosky & Thiede, 2013; Schraw, 2009). In contrast to the absolute calibration accuracy measurement, bias score differences were relative and could thus range from -1 to 1. A negative score indicated underconfidence (actual performance was higher than estimated performance) and a positive score indicated overconfidence (actual performance was lower than estimated by the student).

Statistical analyses

Our first research question was whether differences in type of standards influenced *calibration accuracy with standards present*. To investigate this question, we conducted a regression analysis in SPSS with Calibration accuracy with standards present as dependent variable, Idea-unit standards (yes vs. no) added to Model 1 and Performance level added to Model 2. In a second regression analysis, we included Bias scores as dependent variable to gain insight in whether the miscalibration was caused by either overconfidence or underconfidence.

More importantly, we analysed our second research question, whether the presence of idea-units influenced *calibration accuracy without standards present* over the six critical texts. To test the learning curves of calibration accuracy over time, we conducted a linear regression analysis in SPSS with PROCESS (Hayes, 2013). Calibration on subsequent texts was our dependent variable and (1) Texts (i.e. Time); (2) Idea-units; (3) and Performance level were our independent variables.

Results

Before running any of the analyses, we checked whether there were any a-priori differences in recall performance or calibration without standards

present between students in our two standard groups. Results showed a non-significant difference in Recall performance level on the first critical text, $t(125) = 0.88$, $p = 0.380$, $d = 0.148$, between students in the full-definition standard group ($M = 0.55$, $SD = 0.26$) and students in the idea-unit group ($M = 0.51$, $SD = 0.29$). Furthermore, results showed no significant differences in calibration without standards present on the first critical text between students in the full-definition standard group ($M = 0.35$, $SD = 0.22$) and students in the idea-unit group ($M = 0.35$, $SD = 0.22$), $t(125) = 0.11$, $p = 0.910$, $d < 0.001$. In the following sections, our hypotheses on calibration accuracy both with and without standards present are tested. In all our analyses, a significance level of 0.05 was used.

Calibration accuracy with standards present: Idea-units vs. full-definitions

Table 2 presents the calibration accuracy scores over texts as a function of the two standard groups (full-definition standard vs. idea-unit standard) and the three performance level groups. First, we investigated whether *calibration with standards* differed for students in the full-definition or the idea-unit group. In line with previous findings (Dunlosky et al., 2011) and in support of Hypothesis 1, we found a main effect of Idea-units on calibration with standards present, $\beta = -0.23$, $t(123) = -2.97$, $p = 0.004$. Students who received idea-unit standards while estimating their performance were better calibrated ($M = 0.20$, $SD = 0.09$) than students who received full-definitions standards ($M = 0.24$, $SD = 0.09$). This means that our study confirms the earlier findings of Dunlosky et al. (2011).

Results further showed an overall main effect of Recall performance, $\beta = -0.42$, $t(123) = -5.36$, $p < 0.001$. The negative coefficient indicated that the higher student performance, the lower their miscalibration. To further examine the effects between high, average, and low performers we also divided our students into three performance level groups (see Table 1). Bonferroni pairwise comparisons showed that high performers ($M = 0.16$, $SD = 0.07$) calibrated better than average performers ($M = 0.23$, $SD = 0.08$), $p = 0.002$, and low performers ($M = 0.26$, $SD = 0.10$), $p < 0.001$. The difference between average performers and low performers was not significant $p = 0.146$.

There was no interaction effect between Idea-units and Recall performance, $\beta < 0.01$, $p = 1.00$ —

the effect of idea-units did not differ as a function of performance level and hence was equal for low, average and high performers.

Bias scores

There was a significant effect of Idea-units, $\beta = -0.27$, $t(123) = -3.81$, $p < 0.001$ on Bias scores with standards present. Students in the idea-unit group ($M = 0.06$, $SD = 0.14$) were significantly less overconfident than students in the full-definition group ($M = 0.14$, $SD = 0.16$). These results again confirm the findings of Dunlosky et al. (2011): Idea-units help to diminish overconfidence.

Furthermore, there was a significant effect of Recall performance $\beta = -0.57$, $t(123) = -8.12$, $p < 0.001$. The negative coefficient showed that the high student performance, the lower the miscalibration of these students. Bonferroni pairwise comparisons showed that low performers ($M = 0.20$, $SD = 0.14$) overestimated themselves more than average performers ($M = 0.11$, $SD = 0.14$), $p = 0.003$. Average performers, in turn, overestimated themselves more than high performers ($M = -0.01$, $SD = 0.12$), $p = 0.001$. Again, results did not show an interaction effect between Idea-units and Recall performance $\beta = 0.21$, $t(122) = 0.71$, $p = 0.482$.

Calibration accuracy without standards present (calibration on subsequent texts)

More importantly, we investigated the effect of standards on calibration accuracy without standards present on subsequent texts: Does calibration become better after more standards are received? Table 2 shows the descriptives of calibration accuracy without standards present for each text as a function of the two standard groups (full-definition standard vs. idea-unit standard) and the three performance level groups.

Results showed a significant main effect of Texts, $\beta = -0.08$, $t(758) = -2.07$, $p = 0.039$. Regardless of the experimental group students were in, calibration accuracy without standards present improved slightly over texts as shown by the negative slope (cf. Nederhand et al., 2017). In contrast to the main effect of Texts, there was no main effect of Idea-unit standards $\beta = -0.50$, $t(758) = -1.38$, $p = 0.168$. However, the interaction between Idea-units and Texts was significant $\beta = -0.19$, $t(758) = -2.15$, $p = 0.032$. Students receiving idea-unit standards showed a stronger learning curve over texts—they

Table 2. Calibration accuracy with and without standards present over texts.

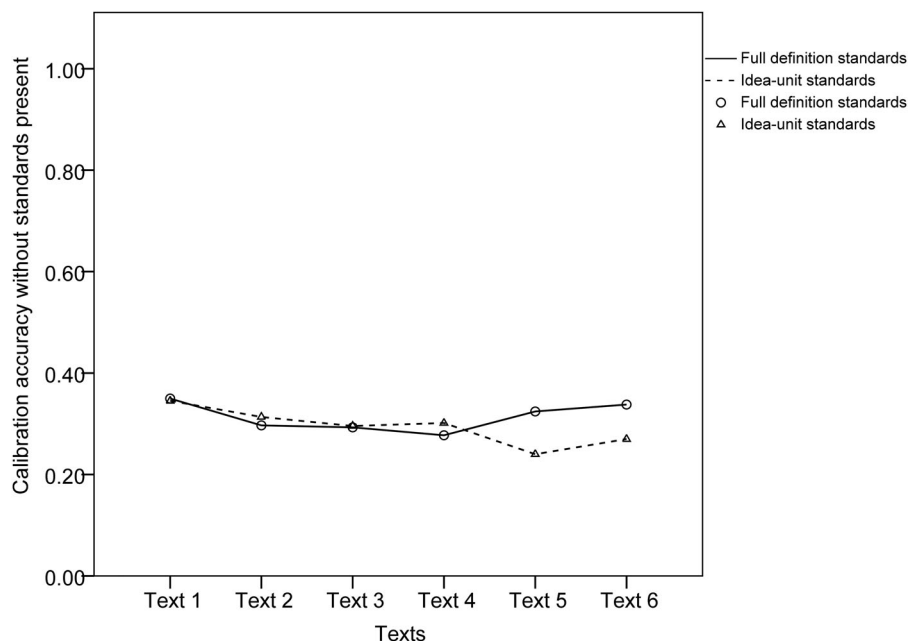
Performance level	Text	Calibration without standard present						Calibration with standard present					
		Full-definition standard			Idea-unit standard			Full-definition standard			Idea-unit standard		
		<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>
Low	1	0.45	0.20	21	0.43	0.27	21	0.33	0.19	21	0.24	0.16	21
	2	0.40	0.23	21	0.39	0.23	21	0.33	0.20	21	0.30	0.24	21
	3	0.37	0.17	21	0.30	0.13	21	0.30	0.16	21	0.20	0.17	21
	4	0.35	0.17	21	0.36	0.16	21	0.24	0.16	21	0.21	0.17	21
	5	0.38	0.16	21	0.24	0.15	21	0.27	0.18	21	0.23	0.16	21
	6	0.41	0.21	21	0.29	0.16	21	0.28	0.18	21	0.23	0.20	21
	Total	0.39	0.19	126	0.34	0.20	126	0.29	0.18	126	0.24	0.18	126
Medium	1	0.36	0.24	25	0.32	0.17	22	0.27	0.22	25	0.24	0.15	22
	2	0.25	0.18	25	0.34	0.19	22	0.23	0.17	25	0.24	0.17	22
	3	0.29	0.18	25	0.29	0.17	22	0.25	0.17	25	0.13	0.13	22
	4	0.28	0.22	25	0.29	0.21	22	0.27	0.22	25	0.22	0.17	22
	5	0.32	0.19	25	0.24	0.17	22	0.22	0.17	25	0.19	0.15	22
	6	0.31	0.19	25	0.28	0.10	22	0.27	0.18	25	0.22	0.15	22
	Total	0.30	0.20	150	0.29	0.17	132	0.25	0.19	150	0.21	0.16	132
High	1	0.22	0.14	18	0.28	0.18	20	0.21	0.15	18	0.17	0.19	20
	2	0.24	0.16	18	0.21	0.14	20	0.19	0.17	18	0.11	0.12	20
	3	0.21	0.12	18	0.29	0.20	20	0.14	0.10	18	0.18	0.17	20
	4	0.19	0.16	18	0.26	0.15	20	0.13	0.12	18	0.14	0.14	20
	5	0.28	0.23	18	0.23	0.15	20	0.16	0.19	18	0.12	0.13	20
	6	0.30	0.18	18	0.23	0.12	20	0.25	0.16	18	0.18	0.11	20
	Total	0.24	0.17	108	0.25	0.16	120	0.18	0.15	108	0.15	0.15	120

improved their calibration accuracy more (see also Figure 2).

We also explored the role of recall performance. Results showed a significant main effect of Recall performance on calibration without standards present $\beta = -0.45$, $t(758) = -5.70$, $p < 0.001$. Again, the negative coefficient indicates that the better students performed, the lower their miscalibration was. Follow-up t -tests showed that high performers ($M = 0.25$, $SD = 0.16$) calibrated better than average performers (M

$= 0.30$, $SD = 0.19$) $t(508) = -3.26$, $p = 0.001$, and that average performers calibrated better than low performers ($M = 0.36$, $SD = 0.20$) $t(532) = -4.05$, $p < 0.001$. Thus, regardless of what type of standards students received, high performers calibrated better without standards present than low performers.

There was a significant interaction between Recall performance and Texts $\beta = 0.32$, $t(758) = 2.36$, $p = 0.018$. Figure 3 depicts the curves of the different performance level groups over time.

**Figure 2.** Calibration accuracy without standards present over texts by standard group.

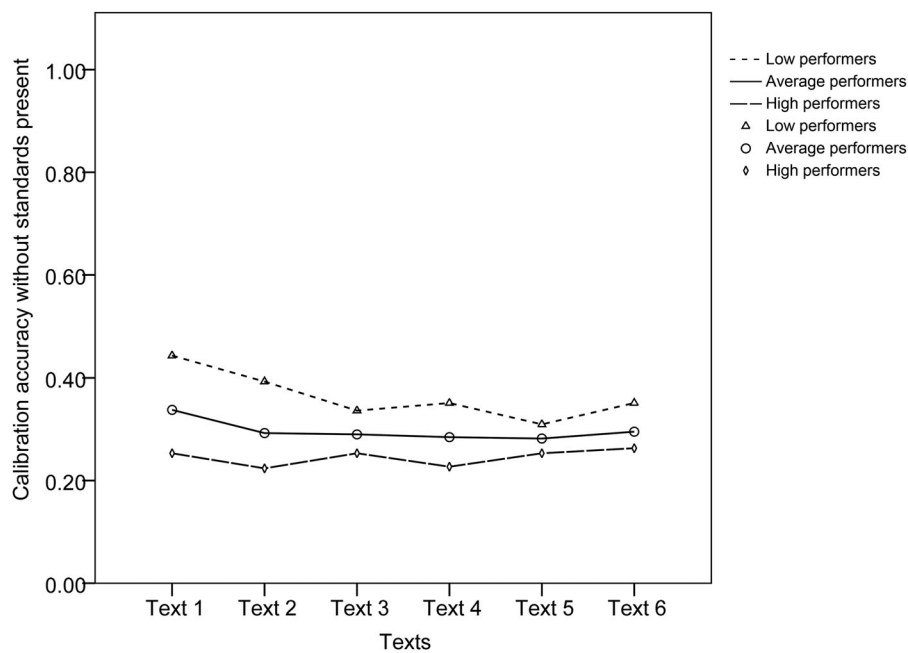


Figure 3. Calibration accuracy without standards present over texts by performance level group.

Whereas low performers $b = -0.02$, $t(758) = -2.82$ $[-0.03, -0.01]$, $p = 0.005$ and average performers $b = -0.01$, $t(758) = -2.08$ $[-0.02, -0.01]$, $p = 0.038$ improved their calibration over texts, calibration of high performers remained stable $b < .01$, $t(758) = 0.18$ $[-0.01, 0.01]$, $p = 0.856$. However, the three-way interaction between Idea-units, Recall performance, and Texts was not significant $\beta = 0.01$, $t(754) = 0.30$, $p = 0.765$. So, although low and average performers improved their calibration accuracy the most over time, it did not matter much what type of standards they received.

Discussion

In the current study, we investigated whether providing students with standards can enhance their calibration accuracy. To that end, we manipulated the amount of detail of the standards students received. Students either received full-definition standards, in which the correct answer was stated, or they received idea-unit standards that provided additional guidance, namely what elements in each definition had to be correctly recalled in order to receive full credit. Based on previous research findings, we expected that providing idea-unit standards would improve calibration accuracy more than providing full-definition standards (cf. Dunlosky et al., 2011). Recent research also shows that providing students with standards

improves their calibration on new, subsequent texts where standards are not directly available (Nederhand et al., 2017). We, therefore, expected that providing students with either full-definition standards or idea-unit standards improves their calibration accuracy on subsequent texts. However, idea-unit standards provide more cues than full-definition standards. We therefore expected to find a larger effect on calibration accuracy when providing students with idea-unit standards than full-definition standards. Our results largely confirm these hypotheses. Below, we elaborate on the theoretical and practical implications of our findings.

Calibration accuracy with standards present

To investigate whether providing idea-unit standards improved calibration more than providing full-definition standards, we compared calibration with standards present of students in the full-definition group with students in the idea-unit group. Results show that, when compared to full-definition standards, providing idea-units further improved calibration. More specifically, overconfidence decreased. This finding supports our hypothesis and is in line with research of Dunlosky et al. (2011): Providing students with an extra cue that informs them of the criterion they should use when scoring answers improves calibration accuracy. The reason for this effect is that students

generally have difficulty estimating their own performance because they use invalid cues (Koriat, 1997; Rawson & Dunlosky, 2007). Hence, previous research has extensively shown that students not receiving any standards do not show any improvement (Dunlosky et al., 2011; Dunlosky, Rawson, & Middleton, 2005; Lipko et al., 2009; Nederhand et al., 2017; Rawson & Dunlosky, 2007). When compared to full-definition standards, idea-unit standards provide students with more guidance how to score their answer, and thus with a more informative cue compared to a full-definition standard, leading to better calibration accuracy.

Calibration accuracy on subsequent texts – without standards present

Besides investigating whether providing standards while students estimate their performance helps them to become better calibrated, we were interested in whether receiving such standards also affects calibration on subsequent texts. Of specific interest were the learning curves of students. Do students indeed show a learning curve and does receiving idea-units lead to better learning?

Our results show a significant linear effect of texts: Regardless of the type of standards received, students calibrated better on the final texts than on the first texts. This is a promising finding because it means that using standards while learning definitions helps students to become more aware of their own performance, even when they are confronted with new definitions. The result is in line with previous research (Nederhand et al., 2017) showing that students who received standards calibrated better over time.

Our results also show that students receiving idea-units learn to calibrate better over texts than students receiving full-definition standards. As Figure 2 shows, this effect seems especially apparent on the final texts. At that time, students had already received multiple standards. In other words, students had received multiple cues regarding the quality of their performance and of their estimates. For example, due to the standards, students may have become aware that they consistently need to recall at least three units to get full credit. This insight could have been used when making new judgements (e.g. “I think I have recalled only two units this time ... my answer is probably partially correct”). So, it seems that practice with standards

allows students to improve their calibration accuracy due to the use of more informative cues.

An alternative explanation to the improvement may be that simply practicing with giving performance estimates may have an effect on calibration accuracy without any improved metacognitive knowledge. However, prior research showed that mere experience with estimating own performance does not seem to be effective (e.g. Bol et al., 2005; Foster, Was, Dunlosky, & Isaacson, 2017; Nederhand et al., 2017). For example, following the same procedure as in the current study, a group of students receiving standards was compared to a group of students who did not receive any standards (Nederhand et al., 2017). Although the students who did not receive standards did practice with estimating their performance 24 times (recalling 24 definitions), their calibration accuracy did not show any improvement over subsequent texts at all. Furthermore, a recent study by Foster et al. (2017) showed that students who estimated their performance on 13 consecutive exams remained overconfident, and did not seem to improve in their estimates. So it seems that repeated practice of estimating one's performance does not, by itself, lead to an improvement in calibration accuracy.

It could also be argued that the use of standards affects performance, and as students tend to overestimate themselves, this could also lead to better calibration accuracy. In our study, however, we did not find any evidence for an improvement of performance over texts, as a repeated measures ANOVA with Recall performance as dependent variable and Text order as independent variable was not significant, $F = 0.34$, $p = 0.34$. So the improvement in calibration accuracy that we found did not seem to be caused by a change in performance. However, whereas there is reason to believe practice effects may not (fully) explain the effects found in the current study, practice effects are nevertheless important to take into account for future research.

The effects of performance level

In the current study, we also explored the effect of Performance Level on calibration accuracy, both with and without standards present. High performers' calibration, both with and without standards present, was better than low performers' calibration accuracy. This is in line with previous research showing better calibration among high performers compared to low performers (Ehrlinger, Johnson,

Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999). Building on this literature, we further examined how Performance Level influenced calibration accuracy without standards present over time. Results showed a significant interaction effect between Time and Performance Level: Especially low performers become better calibrated over time. It has been suggested that low performers suffer from poor calibration accuracy because they have too little knowledge to differentiate between correct and incorrect answers (Kruger & Dunning, 1999). If so, it is indeed unsurprising that low performers benefit the most from receiving extra cues. At the same time, however, it seems strange that providing low performers with either full-definition standards or idea-unit standards would not matter: Results did not show a significant three-way interaction between Idea-units and Performance Level and Texts. As shown in the previous tests, however, the difference between Idea-units and full-definition standards was small. This means that we might have suffered from too little power to statistically show this effect.

Future directions

Although previous research shows that providing standards while students estimate their performance can help them to become better calibrated on the text conducted at that specific moment (Dunlosky et al., 2011; Rawson & Dunlosky, 2007), little was known about how such standards can influence calibration accuracy on subsequent texts with a different content. Whereas theory argues that transfer could take place because students can use better cues to judge their own performance (Butler & Winne, 1995; Koriat, 1997; Zimmerman, 2000), little research was conducted to prove this. With the aim to bridge that gap in the literature, the present experiment shows that providing standards helps students to become better calibrated, also on new texts where standards are not immediately present. While we show that students become better calibrated over texts, the exact metacognitive strategy that underlies this enhanced calibration remains unclear. One possibility is that students became skilled in evaluating their own answer, because comparing previous answers to standards taught them to what aspects in their own recall they should pay attention to, leading to better cue use. It is also possible that students simply started anchoring their estimate of subsequent texts on

their performance of a previous text, which may be an ineffective strategy when text difficulty varies (Geurten & Meulemans, 2017).

Although this study shows that the cues students learn to use in their estimates seem to be beneficial, as indicated by improved calibration accuracy, future research could further investigate the sources of this improvement. For example, students could be asked to think-aloud while estimating their performance, and to explain how they came up with their estimates (Gutierrez de Blume, Wells, Davis, & Parker, 2017). With such a qualitative analysis, the sources of the estimates could be clarified further.

Another suggestion for future research entails the durability of the standard effects. While the current study showed that calibration accuracy can improve over texts, the question remains how long such effects will last. Would students still show better calibration accuracy when re-entering the lab after, for example, a week? We would expect that if students indeed learn to use better cues when estimating performance, their enhanced calibration accuracy should last over time. However, if students simply learn to anchor their estimates on prior performance, the effect may wear off more quickly than when students become aware of all the cues they could use to judge their performance. Thus, to gain more insight in the duration of our effects, further research could examine the precise cues students use when making their estimates, and how this impacts calibration accuracy over time.

Another challenge in this type of research is to decide on how many texts to provide the students with. In the current experiment, it seems that students' motivation lowered or that students were fatigued at the end of the experiment: Figure 2 seems to indicate that calibration accuracy became slightly worse on the last text for students in the full-definition standard group. Furthermore, students complained that the experiment took so long and "seemed to never finish". However, providing students with less texts could diminish the learning effect of standards: Students apparently need some practice before the effect shows. To overcome boredom and fatigue while still providing students with sufficient practice, it might be beneficial to use the same number of texts, but with a better distribution over time. For example, future research could use two test sessions, in which students are provided with three texts in test session one and three texts in the second test session. Using a design in which the time between the two test

sessions is varied also would provide more insight in whether the effects of standards wear off over time.

The results of our study show practical relevance when considering that comparing own answers to standards is an intervention largely similar to self-testing. In a broader sense, this experiment, therefore, shows that self-testing can help students to become better calibrated, also when studying different materials. Over time, student calibration accuracy should improve, especially when they practice with idea-unit standards. However, we used only one type of task, i.e. memorising definitions. Importantly, caution should be exercised when generalising our findings to other types of tasks, such as problem solving.

Conclusion

This study is among the first to experimentally investigate whether students can learn how to improve calibration accuracy when judging their performance. Our results show that students indeed show a learning curve over subsequent texts and that providing students with more detailed standards leads to stronger learning curves. Furthermore, low performers, who are generally considered “at-risk” because of their poor calibration accuracy, show the strongest learning effect. These findings pave an avenue for future research that aims to further unravel the transfer effects of calibration accuracy and the role of standards on different types of texts.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was funded by a Research Excellence Initiative grant from Erasmus University Rotterdam.

ORCID

Marloes L. Nederhand  <http://orcid.org/0000-0001-7388-6381>

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, 1–3. doi:10.1016/j.learninstruc.2012.10.003

- Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, 73(4), 269–290. doi: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. doi:10.2307/1170684
- De Bruin, A. B. H., Kok, E., Lobbetael, 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. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1016/j.obhdp.2007.05.002
- 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. doi:10.1007/s11409-016-9158-6
- Geurten, M., & Meulemans, T. (2017). The effect of feedback on children's metacognitive judgments: A heuristic account. *Journal of Cognitive Psychology*, 29(2), 184–201. doi:10.1080/20445911.2016.1229669
- Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology*, 106(1), 58–68. doi:10.1037/a0033208
- 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. doi: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. doi:10.3758/s13423-011-0181-y.
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York, NY: Guilford Press.
- 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. doi:10.1037/0096-3445.126.4.349.
- 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. doi:10.1037/0022-3514.77.6.1121.
- 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.
- 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. doi:10.1037/a0017599.
- Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. J. P. (2017). *Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels*.
- 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. doi:10.1007/s10409-006-9595-6.
- 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. doi:10.1080/09541440701326022.
- Schraw, G. (2009). Measuring metacognitive judgements. In *Handbook of Metacognition in Education* (pp. 439–462). doi:10.4324/9780203876428

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.

Appendix. 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.