Improving Calibration Accuracy Through Performance Feedback

Marloes L. Nederhand

Improving Calibration Accuracy Through Performance Feedback

Marloes L. Nederhand

Cover by istockphoto – www.istockphoto.com
Lay-out by M. L. Nederhand
Printing by Ridderprint BV – www.ridderprint.nl

ISBN: 978-94-6375-138-4

The research presented in this dissertation was funded by a Research Excellence Initiative grant from Erasmus University Rotterdam awarded to the Educational Psychology section.

Erasmus University Rotterdam

The research was conducted in the context of the Interuniversity Center for Educational Sciences.

ico

Improving Calibration Accuracy Through Performance Feedback

Verbeteren van kalibratieaccuratesse door prestatiefeedback

Proefschrift

ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam

op gezag van de rector magnificus
Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
donderdag 22 november 2018 om 13:30 uur

door
Marloes Lisanne Nederhand
geboren te Rotterdam

Erasmus University Rotterdam

Promotiecommissie

Promotor
Prof. dr. R.M.J.P. Rikers

Overige leden
Prof. dr. F. Paas
Prof. dr. K. Scheiter
Dr. S. Mamede

Copromotor
Dr. H.K. Tabbers

Contents

Chapter 1

General Introduction

We all monitor our performance on a daily basis. As a manager, you may consider whether you have succeeded in motivating your employees to work on a new project. As a teacher, you may consider whether your explanation has really supported the understanding of your students. And as a student, you may wonder whether you have studied sufficiently to pass an exam. Although we all make such considerations regularly, decades of research show that many people are unable to provide accurate estimates of their performance. In a variety of domains, people regularly provide inflated estimates that do not represent their actual performance (Dunlosky & Lipko, 2007; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999; Lichtenstein & Fischhoff, 1977; Sanchez & Dunning, 2017; Sheldon, Dunning, & Ames, 2014). For instance, physicians often misjudge the accuracy of their diagnoses (Davis et al., 2006; Friedman et al., 2005); bankers generally overestimate the profitability of their investments (Glaser & Weber, 2007); and when asked about their schoolwork, over 95% of high school and 90% of college students report that they score equal to or higher than their peers (Chevalier, Gibbons, Thorpe, Snell, & Hoskins, 2009; Thorpe, Snell, Hoskins, & Bryant, 2007).

The notion that people have difficulty in estimating their own performance is problematic as being unable to do so is linked to underachievement (Bol, Hacker, O'Shea, & Allen, 2005; Dunlosky & Rawson, 2012; Hacker, Bol, Horgan, & Rakow, 2000; Kornell & Bjork, 2008; Metcalfe & Finn, 2008; Nietfeld, Cao, & Osborne, 2006). The relationship between monitoring and performance is illustrated in the metacognitive model of Nelson and Narens (1990). Nelson and Narens propose that monitoring and performance are interrelated by a flow of information between a meta-level and an object-level. Whereas the object-level describes the actual performance (what one is actually doing), the meta-level describes a representation of the object level. For example, a student's learning process is described in the object-level, and by monitoring herself (is her knowledge sufficient to pass the exam?), the learning process reaches the meta-level. If she decides her knowledge is indeed satisfactory, this metacognitive judgement can, in turn, influence the object-level through control; hence, she may stop studying. Being unable to accurately monitor one's own performance can therefore lead to individuals failing to realize they should, for example, change ineffective learning strategies or ask for help. Indeed, Dunlosky and Rawson (2012) showed that students who fail to adequately estimate their own performance underachieve.

The importance of being able to monitor one's own performance has increased considerably, especially in education, where students of all levels are increasingly in charge of their learning trajectory (Trilling & Fadel, 2009; Wolters, 2010). Given that these students are shown to be largely incompetent in estimating their own performance (Kruger & Dunning, 1999; Sanchez & Dunning, 2018; Sheldon et al., 2014), and given that inaccurate performance judgements are related to underachievement (Dunlosky & Rawson, 2012), a better understanding of how to improve students' performance estimates is required. The first aim of the studies in this dissertation is therefore to investigate if, and how, students can be supported to learn how they can provide better estimates of their own performance. Furthermore, because inaccurate performance estimates do not solely depend on external support but may also relate to individual differences between students, the second aim of this dissertation is to examine how differences in performance level and more general experience with the task at hand affect both the quality of performance estimates and the effect of the support given. The third and final aim of this dissertation is to test the effects of feedback and individual differences in an ecological valid school setting.

Measuring performance estimates

When investigating how to improve the estimates students make of their own performance, one metric that is used frequently is calibration accuracy. Calibration accuracy is a term that is used to describe how well performance estimates match actual performance (Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1990). It is measured using the following formula in which the absolute difference between an estimated performance score (e_i) and actual performance score (p_i) are calculated:

$$\text{Absolute calibration accuracy} = \frac{1}{N}\sum_{i=1}^{N} |e_i - p_i|\cdot$$

To illustrate, imagine two students estimating their exam grades. Student A thinks that she will obtain 8 of the 10 points, while student B thinks she will obtain 6 of the 10 points. Both students actually obtained 7 points, and hence, both are equally miscalibrated (i.e., both estimates differ 1 point from the actual performance). Although this measure of calibration accuracy provides insight into the mismatch between estimated and actual performance, the direction of the mismatch remains unclear. This direction is indicated as bias and describes whether students are overconfident or underconfident (Schraw, 2009). To measure bias, the following formula is used:

$$\text{Bias index} = \frac{1}{N}\sum_{i=1}^{N} (e_i - p_i)\cdot$$

Again, e_i refers to the estimated performance score, and p_i refers to the actual performance score. Note that the only difference between the bias index and absolute calibration accuracy is that the difference between estimated and actual performance is not absolute when computing bias scores. When looking at the bias scores in the example described previously, we now see that whereas both student A and B were equally miscalibrated, student A showed an overconfident bias, while student B showed an underconfidence bias. Note that when mean bias scores are calculated, negative and positive bias scores can cancel each other out. It is therefore possible that students show a zero bias score on average, but are still miscalibrated. Using a combination of absolute calibration accuracy scores and bias scores is therefore recommended.

Cues to support calibration accuracy: the role of performance feedback

The studies in this dissertation aim to examine how calibration accuracy can be improved. However, to investigate this question, it is important to examine why students are estimating their own performance poorly in the first place. According to the cue-utilization framework of Koriat (1997), students use a variety of cues when providing a performance estimate. For example, they think about how much information they recalled (Baker & Dunlosky, 2006), how fluently this information came to mind (Finn & Tauber, 2015), and how familiar the test items appeared (Metcalfe & Finn, 2012). However, students find it hard to select the right cues when estimating their performance, leading to poor calibration accuracy and little improvement over time (Thiede, Griffin, Wiley, & Anderson, 2010). Illustrating this problem, Foster, Was, Dunlosky, and Isaacson (2017) showed that when students were asked to estimate their exam grade during a course, the students' estimated exam grade was anchored on previously provided performance estimates, while such prior estimates were not predictive for actual exam performance. By failing to switch to a more valid cue, many students did not improve their calibration accuracy over time.

Thus, to ensure accurate calibration, students need to be assisted in using cues that are predictive of their performance (Thiede et al., 2010). A promising way of supporting students to use better cues is by simply providing them with such a cue: feedback on the quality of their actual performance. Indeed, providing performance feedback has been found to improve calibration accuracy (Bol & Hacker, 2012; Koriat, 1997; Labuhn, Zimmerman, & Hasselhorn, 2010; Lipko et al., 2009; Nietfeld et al., 2006). Among the first to show the beneficial effect of performance feedback on calibration accuracy were Rawson and Dunlosky (2007). In their study, students learned and recalled definitions from a text. After each recall attempt, students were asked to judge the quality of their recall attempt on a 3-point scale (0 = incorrect; 0.5 = partially correct; 1 = correct). While estimating their performance, half of the students were provided with performance feedback in the form of the correct definition (standard). The other half of the students were required to make an estimate without any standard present. Rawson and Dunlosky showed that students who could compare their recall attempt to the standard had significantly better calibration accuracy than the group of students who did not receive a standard. Thus, the standard served as an extra (valid) cue, helping students to provide better performance estimates.

Receiving performance feedback thus helps students to become more aware of their actual performance level and to obtain more insight in the accuracy of their performance estimates. So far, however, the literature has predominantly focused on providing performance feedback *while* students estimated their performance (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Dunlosky, Rawson, & Middleton, 2005; Lipko et al., 2009; Rawson & Dunlosky, 2007). This approach leaves unanswered how their calibration accuracy will be affected when encountering a similar task where feedback is not immediately present. For example, in the experiment of Rawson and Dunlosky (2007), would students who received standards when estimating the quality of their recalled definitions, also be better calibrated when they had to learn and recall a new set of definitions? This question is important, because in many daily situations, and for many tasks, performance feedback is not immediately available. To provide better performance estimates on new tasks that are similar in structure but different in content, students could use the feedback they received on previous tasks. For example, students who were overconfident on previous tasks, may become more conservative on subsequent tasks.

Although this reasoning is intuitively plausible, experimental evidence showing the benefits of receiving performance feedback on calibration accuracy on new tasks is lacking. Furthermore, the quasi-experimental studies that have been conducted show mixed results. For example, in the previously described study by Foster et al. (2017), students failed to improve their calibration accuracy even after 13 feedback moments (i.e., students took 13 tests and received 13 grades) because they continued to anchor their judgements on prior estimates instead of on prior grades. Yet, other researchers such as Miller and Geraci (2011), Labuhn et al. (2010) and Callender, Franco-Watkins, and Roberts (2015) showed that providing students with feedback about their performance did help them to become better calibrated over subsequent tests. However, because these studies did not use an experimental design, and many variables were varied simultaneously, generalization of the results is problematic. This indicates the need for experimental research on whether providing feedback can indeed be used to help enhance calibration accuracy in such a way that students also show better calibration accuracy on subsequent tasks where feedback is absent.

Individual differences in calibration accuracy

The studies of Hacker et al. (2000), Nietfeld et al. (2006), and Hacker (2008) touch upon a second gap in the literature: individual differences in calibration accuracy and its improvement. Low and high performers have been found to calibrate differently (Kruger & Dunning, 1999), and this difference potentially impacts how they improve their calibration accuracy after feedback is given (Hacker, Bol, & Bahbahani, 2008; Hacker et al., 2000; Nietfeld et al., 2006). Studies have shown the existence of differences in calibration accuracy and bias among students from different performance levels (Ehrlinger et al., 2008; Kruger & Dunning, 1999). Whereas high performers (i.e., the 25% best performing students) are generally well calibrated but somewhat underconfident, low performers (i.e., the 25% poorest performing students) show large miscalibration and overconfidence (Ehrlinger et al., 2008; Kruger & Dunning, 1999). This effect is called the Dunning-Kruger effect (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Miller & Geraci, 2011b; Pennycook, Ross, Koehler, & Fugelsang, 2017), and its occurrence has most often been explained by using the argument originally provided by Kruger and Dunning (1999): low performing students suffer from a double “curse” (Sanchez & Dunning, 2017). The first curse of low performers is that they have a knowledge deficiency, which leads to their second curse: because of this knowledge decifiency, students have difficulty in discriminating between good and poor performance. In other words, low performers simply cannot recognize what is correct or incorrect, thereby leading to inaccurate judgements of their performance.

Typically, studies on the Dunning-Kruger effect operationalize performance level by dividing students into different categories based on their task performance (Bol et al., 2005; Hacker et al., 2000; Kruger & Dunning, 1999; Miller & Geraci, 2011b; Nietfeld et al., 2006). Although the literature on this topic is scarce, some studies have also found that task experience is related to calibration accuracy. For example, experienced investment bankers calibrated better than inexperienced ones because the former had more knowledge of their previous portfolio benefits (Glaser & Weber, 2007). Furthermore, students from higher grades have been shown to calibrate better than students from lower grades (Lockl & Schneider, 2002; Van der Stel & Veenman, 2010), as metacognitive awareness has been found to develop until adulthood (Paulus, Tsalas, Proust, & Sodian, 2014; Weil et al., 2013). Hence, individual differences in calibration accuracy appear not to be confined to students of different performance level groups, but may also appear among individuals that differ in other, more fundamental aspects, such as age and task experience.

Whereas the effect of performance level on calibration accuracy has often been shown, and even seems to hold among groups that differ substantially in terms of experience (Glaser & Weber, 2007), experimental studies that aim to improve calibration have rarely examined the role of performance level. However, this variable has the potential to moderate the effect of feedback on calibration accuracy. On the one hand, low performers have more room to improve their performance estimates than high performers, given their poorer calibration accuracy (e.g., Kruger & Dunning, 1999) and their more frequent use of invalid cues (Gutierrez de Blume, Wells, Davis, & Parker, 2017; Thiede et al., 2010). Providing low performers with performance feedback that does serve as a valid cue should therefore be especially helpful. On the other hand, Stone (2000) has argued that low performers may encounter more difficulty in understanding or incorporating feedback correctly, leading them to benefit less. To complicate things even more, the few studies that did include performance level in their analyses showed largely mixed results. For example, Hacker et al. (2000) and Nietfeld et al. (2006) showed that only high performers improved their calibration accuracy after receiving feedback, Miller and Geraci (2011) found that low performers benefited as well, but Hacker et al. (2008) found that low performers' calibration became even worse after feedback had been given. As low performers are the ones that are most poorly calibrated and most overconfident, they would especially need support in making better estimates. If the proposed intervention with performance feedback would be less useful for them, this would signal the need for adaptive support for low performers. Hence, research on improving calibration accuracy should take performance level into account.

To fully understand differences in metacognitive awareness among performance level groups, however, Miller and Geraci (2011) proposed that measuring calibration accuracy and bias may not be sufficient. By making a distinction between functional confidence (i.e., students' performance estimates) and subjective confidence (i.e., how much confidence students assigned to their estimates), Miller and Geraci showed that while low performers were functionally overconfident (their estimated grades were higher than their actual ones), low performers were not subjectively overconfident. In fact, they assigned little confidence to their incorrect performance estimates. Miller and Geraci therefore argued that to understand possible individual differences in metacognitive awareness, confidence judgements (so called second-order judgements, SOJs) may need to be taken into account; besides asking students to estimate their grade, they should indicate how confident they are of this estimate. To provide further insight into the effects of performance level on calibration accuracy, research would therefore benefit from including second-order judgements as an outcome measure.

Improving calibration accuracy in an authentic educational setting

Finally, in addition to the central aims of clarifying the role of performance feedback and performance level on calibration accuracy, the third aim of this dissertation is to bridge the gap between science and educational practice. If calibration accuracy on new tasks can indeed be enhanced with performance feedback, will it then be possible to design a feasible intervention using materials and feedback that are naturally available in an everyday classroom setting?

In school, the most common type of performance feedback is outcome feedback (e.g., grades). Because outcome feedback only shows students' overall level of performance, it is evidently less informative than feedback that also contains the correct answers, such as standards. However, outcome feedback is still considered useful when improving calibration as it can help students to become more aware of the difference between estimated and actual performance, which in turn can encourage students to examine why their performance was better or worse than expected (Nelson & Narens, 1990; Zimmerman, 2000). The potential of outcome feedback is promising because schools face constraints on time and money available. Interventions involving outcome feedback, a type of feedback that is naturally present, could therefore be easily implemented.

However, studies have shown that merely providing outcome feedback does not directly lead to enhanced calibration (Bol et al., 2005; Foster et al., 2017; Hacker et al., 2008; Huff & Nietfeld, 2009). It seems that instead of merely providing students with outcome feedback, students need to be encouraged to use feedback to improve calibration (Hacker et al., 2000; Miller & Geraci, 2011a; Nietfeld et al., 2006). Unfortunately, how this needs to be done exactly remains unclear, as previous studies have delivered mixed results. Sometimes, students improved their calibration accuracy after receiving an outcome feedback intervention (Callender et al., 2015; Hacker et al., 2000; Huff & Nietfeld, 2009; Miller & Geraci, 2011a), whereas in other studies, no improvement was found (Bol et al., 2005; Foster et al., 2017), or calibration accuracy even worsened (Hacker et al., 2008). Perhaps a major reason for these mixed results is that a systematic experimental approach is often lacking—different variables were manipulated at once, making it hard to generalize the results among studies and to identify the effective element in each feedback intervention. Hence, with the aim of constructing a feasible and easily implemented intervention for a school, the final purpose of this dissertation is to systematically investigate how students can be supported to use the outcome feedback received in class to improve their calibration accuracy.

Research questions and overview of the studies in this dissertation

Taken together, the goal of this dissertation is to provide a better understanding of whether calibration accuracy on new tasks could be enhanced with the help of performance feedback, both for high and low performers, in both laboratory and school settings. This dissertation had two research questions:

1. Does providing performance feedback help students to enhance their calibration accuracy in such a way that they will also show better calibration accuracy on new tasks, both in the laboratory and in a classroom setting?
2. Does the effectiveness of performance feedback to improve calibration accuracy on new tasks depend on performance level?

To answer the research questions, five studies are included in this dissertation, each described in a separate chapter.

The first two chapters, Chapter 2 and Chapter 3, build on the studies by Rawson and Dunlosky (2007) and Dunlosky et al., (2011). These studies showed that student calibration accuracy can be improved with performance feedback (i.e., performance standards) on a text reading task. Their results showed that standards strongly benefitted the performance estimates of the students. However, Rawson and Dunlosky (2007) and Dunlosky et al. (2011) did not investigate effects on new tasks. Thus, Chapter 2 and 3 aim to examine this question: Does providing standards also improve calibration accuracy on a new task?

Chapter 2 describes an experiment in which we tested whether providing students with performance standards (i.e., the correct answer) improved their calibration on a subsequent new task. Students had to read the same texts as presented by Rawson and Dunlosky (2007) about a variety of topics and were requested to learn four definitions in each text. After reading each text, students had to recall all four definitions and estimate the quality of their definition on a three-point scale (incorrect, partially correct, correct). After providing this initial performance estimate, half of the students received a standard (i.e., the correct definition) and again scored their performance. Calibration accuracy of students both with and without standards was compared on new texts. Furthermore, the differential effect among high and low performers was examined.

In Chapter 3, the question of whether varying the type of standard, as shown by Dunlosky et al. (2011), influenced calibration accuracy differently is investigated. Largely similar to the design of the study described in Chapter 2, students read several texts and in each text learned four definitions. After recalling each of the definitions during a test, students made a performance judgement and estimated whether their recall attempt was incorrect, partially correct, or correct. After providing this first performance estimate, students received either a full definition standard showing the correct answer, or an idea-unit standard, in which the correct answer was parsed into parts. Each of these parts had to be present for students to receive full credit. The question of whether providing students with extra guidance would further enhance calibration accuracy and would lead to a steeper learning curve was also tested. Furthermore, the difference between low and high performers was again investigated.

In Chapter 2 and 3, performance level was operationalized at the task level: the best and worst performing participants were compared to each other. Chapter 4 focuses on larger experience differences: calibration differences between board-certified medical specialists and second-year medical students. In the study described in Chapter 4, specialists and students solved medical cases and estimated whether they thought they adequately solved each individual case. Only half of the medical specialists and students received feedback (i.e. the correct diagnosis) after a case. Differences in calibration accuracy on new clinical cases were tested between (1) the group that received feedback versus no feedback, and (2) between medical specialists and medical students.

Chapter 5 and 6 aim to investigate calibration accuracy in a more authentic educational setting. Chapter 5 describes an observational study that presented a baseline measure of calibration accuracy, bias, and second-order judgements of students in secondary school and university. After doing their exam, students received a form on which they could estimate their obtained grade and rate the confidence they had in this estimate (i.e., second-order judgements). In addition to describing a baseline measure, the difference in calibration accuracy was compared between students in university and secondary school. Moreover, it was studied whether both university and secondary school students aligned their confidence judgement to the accuracy of their estimates.

Chapter 6 describes the final study of this dissertation. A feedback intervention was systematically implemented in a Dutch secondary school during one school year. Students were asked to estimate their grade after each exam and to estimate how confident they were in their estimate. Students were divided into three groups, differing in the level of support: the first group of students only estimated their grade; the second group of students had to calculate the difference between their estimated and actual grade; and the last group had to reflect on how they estimated their performance and on explanations for differences between their estimated and actual performance. Besides investigating differences between the intervention groups, this final study also included performance level and examined how this interacted with the effect of feedback.

In the final chapter, Chapter 7, a summary and discussion of the main findings is presented and theoretical and practical implications of the studies described are discussed.

Chapter 2

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

This chapter is under revision as:
Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels.

An earlier version of this chapter has received the Best paper award, of the Interuniversity Center for Educational Research (ICO).

Abstract

This experimental study explores whether feedback in the form of standards not only helps students in giving more accurate performance estimates on current tasks, but also on new, similar tasks, and whether performance level influences the effect of standards. Using the set-up from Rawson and Dunlosky (2007), we provided 122 first-year psychology students with 7 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 as well as on new, similar tasks, when standards were not available yet. Also, with or without standards, high performers calibrated better than low performers. So, standards help students in learning to calibrate better, regardless of performance level.

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).

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 not only improve calibration accuracy on the current task, but also on subsequent, similar tasks, when standards are not available anymore.

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.

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.

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. Based on 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*. However, the process of comparing own answer to a standard may also provide students with a cue about the quality of their *estimate* of performance. If students recognize the difference between the estimate they gave with the standard, and the estimate they would have given without the standard present, this could serve as an extra cue when making estimates on new tasks. For example, if students recognize that they would 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, but also on a similar subsequent task without a standard present.

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. 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.

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.

Method

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 percent females and 15.6 percent 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) x 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. Table 1 displays the performance accuracy of the percentile groups.

Table 1.
Test performance scores

	Standards								
	No			Yes			Total		
Performance	*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.

Materials

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

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.

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, 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 percent) 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.

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.

Performance estimates

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 won’t be able*) to 10 (*definitely will be able*).

Post-diction 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 *post-diction without standard present*.

Post-diction 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 *post-diction with standard present*.

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).[1] 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.

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.

Calibration accuracy without standards present. For each text, *calibration accuracy without standards present* was calculated as the absolute difference between *post-dictions 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).

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 *post-dictions with standards present* and actual obtained credit, averaged over the four definitions. Again, bias scores were calculated by taking the (non-absolute difference) between *post-dictions without standards present* and actual obtained credit (cf. Dunlosky & Thiede, 2013; Schraw, 2009).

[1] 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).

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 *post-diction without standard present* before they could continue to the next key term. 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 *post-diction 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.

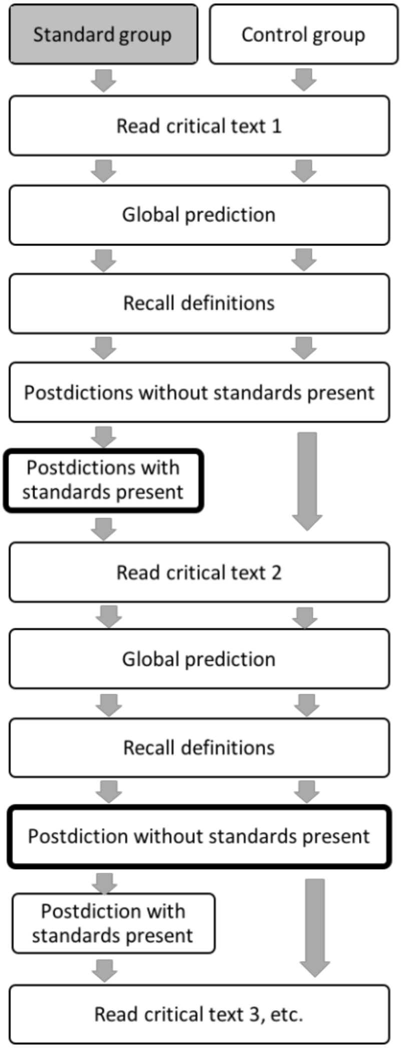

Standard group
Control group
Read critical text 1
Global prediction
Recall definitions
Postdictions without standards present
Postdictions with standards present
Read critical text 2
Global prediction
Recall definitions
Postdiction without standards present
Postdiction with standards present
Read critical text 3, etc.

Figure 1. A graphical display of the experimental procedure.

Our procedure differs in two ways from that of Rawson and Dunlosky (2007). First, students in our standard-group also provided post-dictions 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 *post-dictions without standards present* of the no-standard group, to the *post-dictions 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 *post-dictions 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.

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.

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 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, η^2 = .221, replicating the findings of Rawson and Dunlosky (2007).

Secondly, we explored whether low and high performers would benefit equally from receiving standards. We found a non-significant interaction effect between Standards and Performance Level, *F*(116) = 1.13, *p* = .325, η^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, η^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 medium performers, *p* = .002. 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, η^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, η^2 = .271. Low performers showed the most bias (*M* = .17, *SD* = .14), followed by medium performers (*M* = .08, *SD* = .12) and high performers only showed a negligible bias (*M* < .01, *SD* = .10). There was no significant interaction between standards and performance level *F*(116) = 1.37, *p* = .259, η^2 = .023.

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 post-diction 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, η^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, η^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. There was again no significant interaction effect between Performance Level and Standards, *F*(116) = 1.27, *p* = .285, η^2 = .015.

Table 2.
Calibration accuracy without standard present

	Standards								
	No			Yes			Total		
Performance	*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 3 shows the bias scores of all performance level groups. Results showed a main effect of standard group *F*(116) = 6.35, *p* = .013, η^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, η^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). Finally, high performers showed the least bias (*M* = -.02, *SD* = .13). There was no significant interaction between standards and performance level *F*(116) = 1.41, *p* = .248, η^2 = .024.

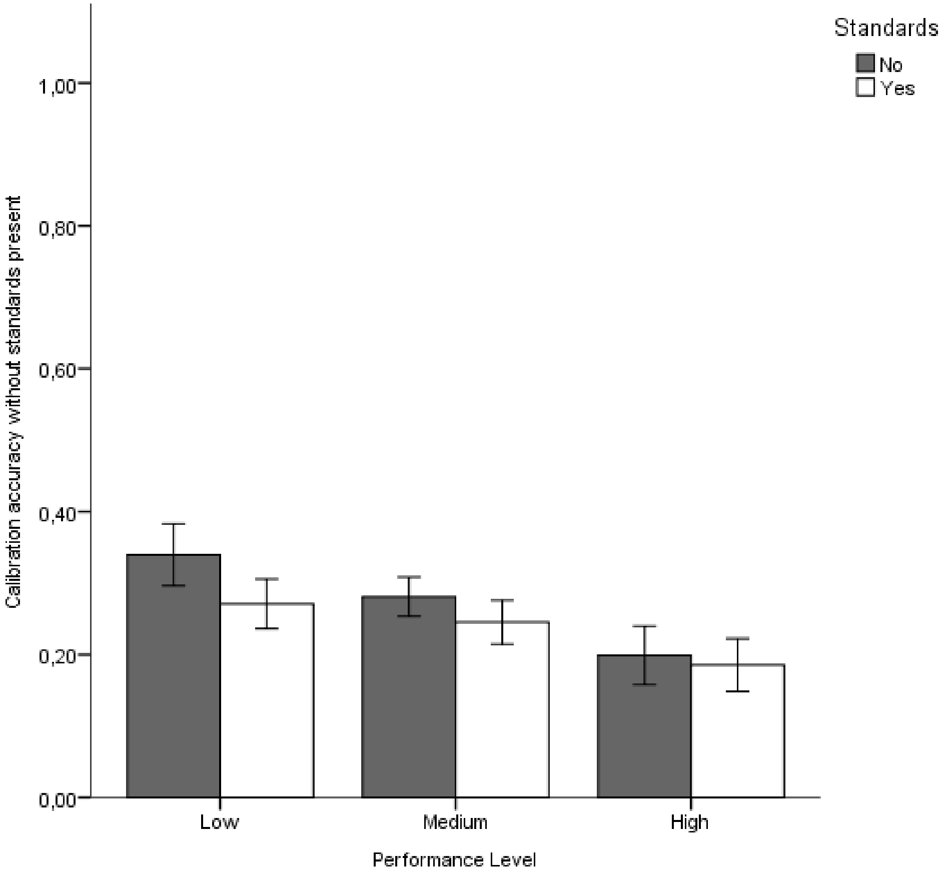


Figure 2. This graph displays the effects of standards and performance level on *calibration accuracy without standards present* (i.e., calibration accuracy on subsequent tasks) ranging from 0 to 1 (note that the lower the score, the better the match between estimated performance and actual performance).

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, $F(116) = 3.27$, $p = .073$, $\eta^2 = .025$.

- Insert Figure 4 around here -

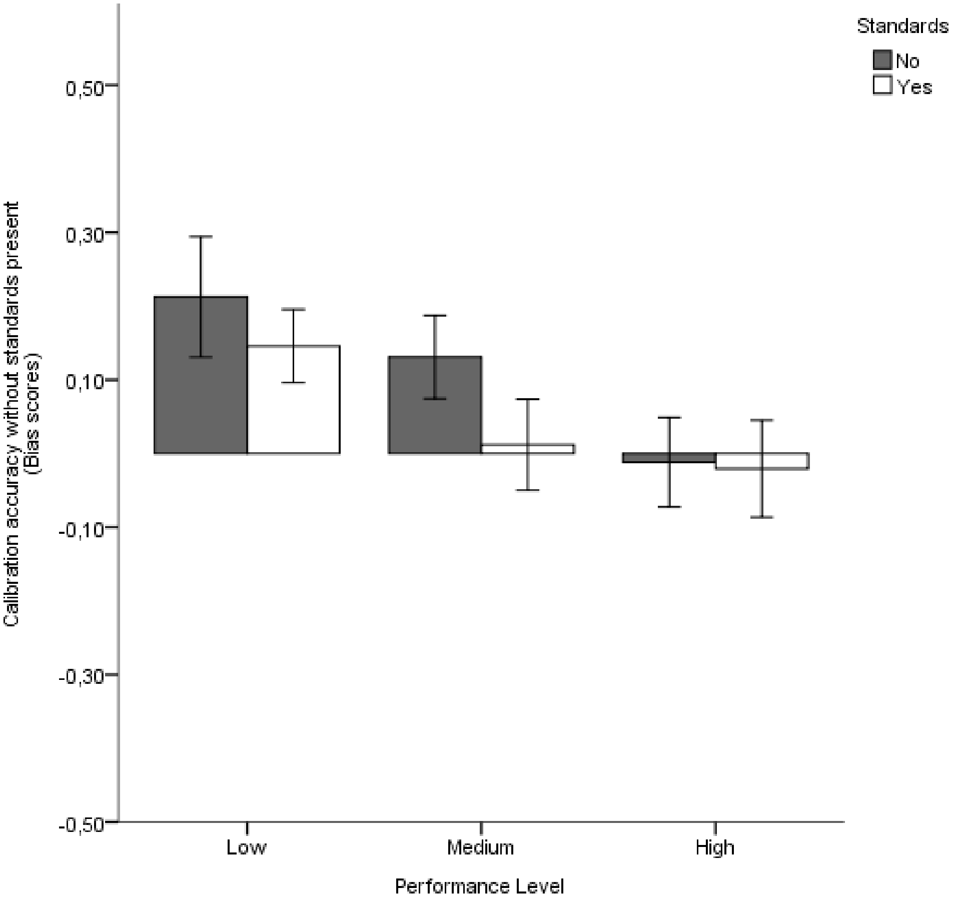


Figure 3. This graph displays the effects of standards and performance level on the bias scores (from -1 to +1) of *calibration accuracy without standards present* (i.e., calibration accuracy on subsequent tasks). Note that the closer to zero, the better the match between estimated performance and actual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

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.

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.

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.

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 new tasks. 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 new tasks. Furthermore, our results are promising for educational practice: students can learn from standards and are capable of adjusting their calibration accordingly, even when they are confronted with new tasks.

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 post-dictions (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 post-dictions can be improved, a recent study by Foster et al. (2017) showed that even after thirteen exams, students were unable to *predict* their next exam grade. Indeed, our results show that although standards improve post-diction 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): while 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 lets students not only compare their own answer to the correct answer, but also shows 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 and colleagues (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 both low and high performers benefit equally well 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).

Finally, it is important to note 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, η_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).

Conclusion

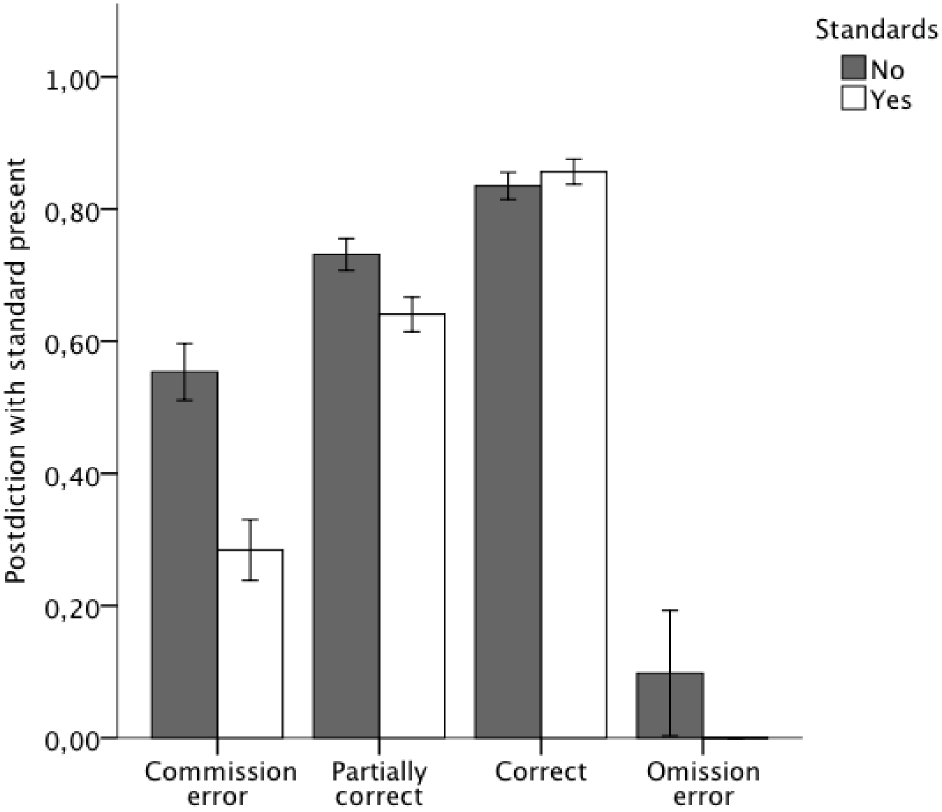
Our study is the one of the first to investigate the role of performance level when students receive standards to improve their calibration accuracy on textual recall tasks. We have shown that providing standards improves calibration accuracy for all performance levels—although low performers show more 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.

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 – Different types of judgement errors



Chapter 3

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

This chapter is published as:
Nederhand, M.L., Tabbers, H.K., Homaira, A., & Rikers, R.M.J.P. (2018). Improving calibration over texts by providing standards both with and without idea-units. *Journal of Cognitive Psychology*. doi:10.1080/20445911.2018.1513005

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.

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).

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 et al., 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 et al., 2005; Dunlosky & Rawson, 2012).

According to the cue-utilisation view of Koriat (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 performance feedback such as standards (i.e., the correct solution; Dunlosky, Hartwig, Rawson, & Lipko, 2011; Dunlosky & Thiede, 2013; 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 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, 2013a; 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 (Submitted) 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. (Submitted) 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 (1998), calibration accuracy depends on the cues that are used when estimating ones 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., Submitted), 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 et al., 2000; Nietfeld et al., 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) x 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.

Table 1.
Recall performance scores

	Idea-unit standards								
	No			Yes			Total		
Performance	*N*	*M* (*SE*)	*95% CI*	*N*	*M* (*SE*)	*95% CI*	*N*	*M* (*SE*)	*95% CI*
Low	21	.35 (.02)	[.32, .38]	21	.37 (.02)	[.34, .40]	42	.36 (.01)	[.33, .39]
Average	25	.58 (.02)	[.55, .61]	22	.54 (.02)	[.51, .57]	47	.56 (.01)	[.54, .58]
High	18	.79 (.02)	[.75, .82]	20	.75 (.02)	[.72, .78]	38	.77 (.01)	[.76, .79]
Total	64	.56 (.02)	[.52, .61]	63	.55 (.02)	[.50, .59]	127	.63 (.01)	[.60, .65]

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 key word (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 the 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 $F(5, 756) = 0.25$, $p = .941$, $\eta_p^2 = .002$. 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 key word. Student 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 .86, with a 95% confidence interval from .81 to .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 A 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 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. (Submitted) as closely as possible, but was not of further interest in the current study. Then, students continued with a recall test, in which one key word 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 postdiction 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 postdiction, 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 postdiction 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

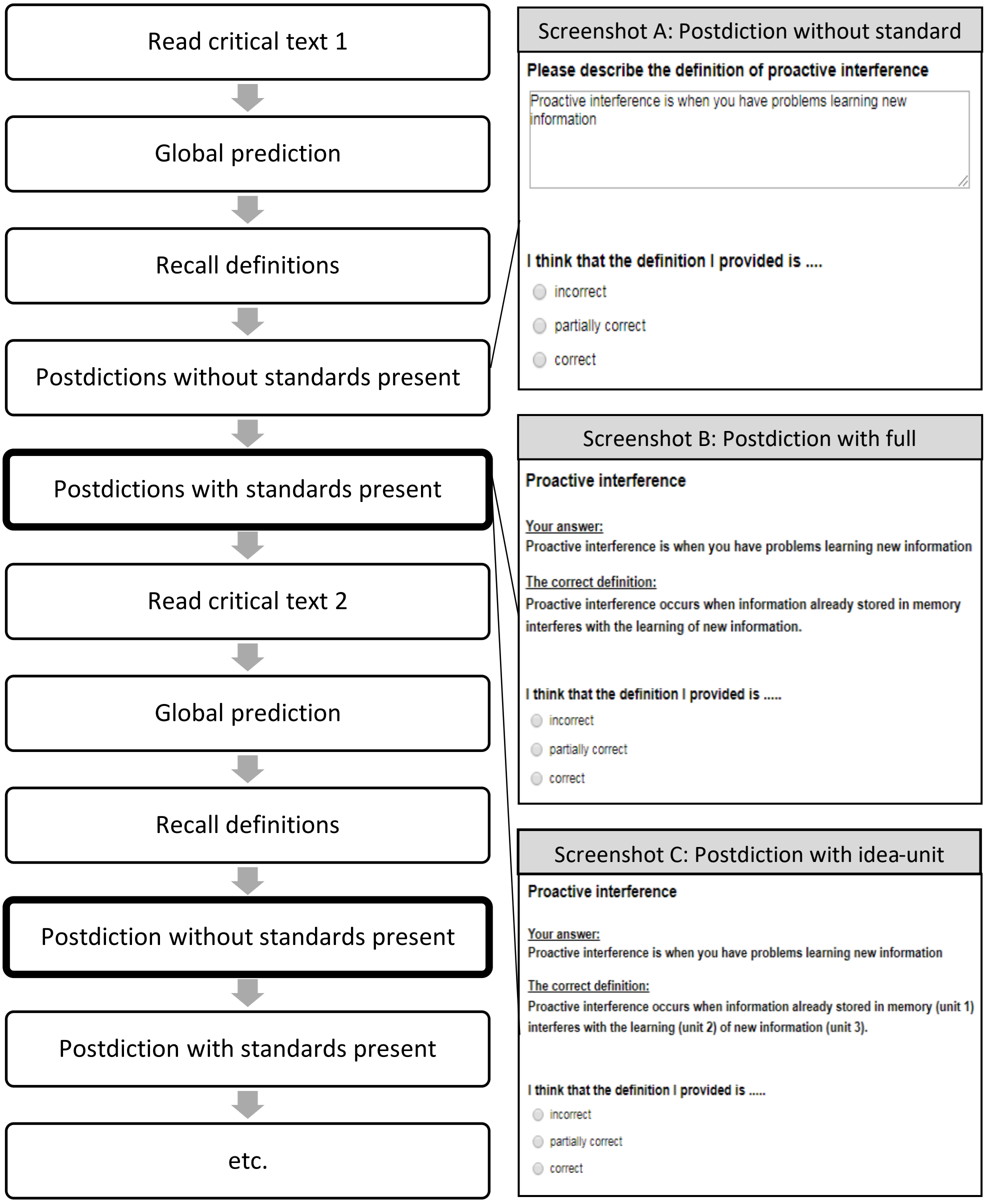
Calibration accuracy. Two types of calibration accuracy were calculated. The first type, calibration with standards present, was calculated by the absolute difference between postdictions with standards present and actual performance, as scored by the experimenter. The second type of calibration accuracy, calibration without standards present, was calculated by the absolute difference between postdictions 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., postdictions 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.


Read critical text 1
Global prediction
Recall definitions
Postdictions without standards present
Postdictions with standards present
Read critical text 2
Global prediction
Recall definitions
Postdiction without standards present
Postdiction with standards present
etc.
Screenshot A: Postdiction without standard
Screenshot B: Postdiction with full
Screenshot C: Postdiction with idea-unit

Please describe the definition of proactive interference

Proactive interference is when you have problems learning new information

I think that the definition I provided is ....

○ incorrect

○ partially correct

○ correct

Proactive interference

Your answer:
Proactive interference is when you have problems learning new information

The correct definition:
Proactive interference occurs when information already stored in memory interferes with the learning of new information.

I think that the definition I provided is .....

○ incorrect

○ partially correct

○ correct

Proactive interference

Your answer:
Proactive interference is when you have problems learning new information

The correct definition:
Proactive interference occurs when information already stored in memory (unit 1) interferes with the learning (unit 2) of new information (unit 3).

I think that the definition I provided is .....

○ incorrect

○ partially correct

○ correct

Figure 2. The procedure of the current experiment, including screenshots from the postdiction estimates.

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* = .380, *d* = 0.148, between students in the full-definition standard group (*M* = .55, *SD* = .26) and students in the idea-unit group (*M* = .51, *SD* = .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* = .35, *SD* = .22) and students in the idea-unit group (*M* = .35, *SD* = .22), *t*(125) = 0.11, *p* = .910, *d* < .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 .05 was used.

Calibration accuracy with standards present: idea-units versus 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, β = -.23, *t*(123) = -2.97, *p* = .004. Students who received idea-unit standards while estimating their performance were better calibrated (*M* = .20, *SD* = .09) than students who received full-definitions standards (*M* = .24, *SD* = .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, β = -0.42, *t*(123) = -5.36, *p* < .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* = .16, *SD* = .07) calibrated better than average performers (*M* = .23, *SD* = .08), *p* = .002, and low performers (*M* = .26, *SD* = .10), *p* < .001. The difference between average performers and low performers was not significant *p* = .146.

There was no interaction effect between Idea-units and Recall performance, β < .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, β = -0.27, *t*(123) = -3.81, *p* < .001 on Bias scores with standards present. Students in the idea-unit group (*M* = .06, *SD* = .14) were significantly less overconfident than students in the full-definition group (*M* = .14, *SD* = .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 β = -0.57, *t*(123) = -8.12, *p* < .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* = .20, *SD* = .14) overestimated themselves more than average performers (*M* = .11, *SD* = .14), *p* = .003. Average performers, in turn, overestimated themselves more than high performers (*M* = -.01, *SD* = .12), *p* = .001. Again, results did not show an interaction effect between Idea-units and Recall performance β = 0.21, *t*(122) = 0.71, *p* = .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, β = -0.08, *t*(758) = -2.07, *p* = .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., Submitted). In contrast to the main effect of Texts, there was no main effect of Idea-unit standards β = -0.50, *t*(758) = -1.38, *p* = .168. However, the interaction between Idea-units and Texts was significant β = -0.19, *t*(758) = -2.15, *p* = .032. Students receiving idea-unit standards showed a stronger learning curve over texts—they improved their calibration accuracy more (see also Figure 2).

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

		Calibration without standard present						Calibration with standard present					
		Full definition standard			Idea-unit standard			Full definition standard			Idea-unit standard		
Performance level	Text	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
Low	1	.45	.20	21	.43	.27	21	.33	.19	21	.24	.16	21
	2	.40	.23	21	.39	.23	21	.33	.20	21	.30	.24	21
	3	.37	.17	21	.30	.13	21	.30	.16	21	.20	.17	21
	4	.35	.17	21	.36	.16	21	.24	.16	21	.21	.17	21
	5	.38	.16	21	.24	.15	21	.27	.18	21	.23	.16	21
	6	.41	.21	21	.29	.16	21	.28	.18	21	.23	.20	21
	Total	.39	.19	126	.34	.20	126	.29	.18	126	.24	.18	126
Medium	1	.36	.24	25	.32	.17	22	.27	.22	25	.24	.15	22
	2	.25	.18	25	.34	.19	22	.23	.17	25	.24	.17	22
	3	.29	.18	25	.29	.17	22	.25	.17	25	.13	.13	22
	4	.28	.22	25	.29	.21	22	.27	.22	25	.22	.17	22
	5	.32	.19	25	.24	.17	22	.22	.17	25	.19	.15	22
	6	.31	.19	25	.28	.10	22	.27	.18	25	.22	.15	22
	Total	.30	.20	15	.29	.17	132	.25	.19	15	.21	.16	132
High	1	.22	.14	18	.28	.18	20	.21	.15	18	.17	.19	20
	2	.24	.16	18	.21	.14	20	.19	.17	18	.11	.12	20
	3	.21	.12	18	.29	.20	20	.14	.1	18	.18	.17	20
	4	.19	.16	18	.26	.15	20	.13	.12	18	.14	.14	20
	5	.28	.23	18	.23	.15	20	.16	.19	18	.12	.13	20
	6	.30	.18	18	.23	.12	20	.25	.16	18	.18	.11	20
	Total	.24	.17	18	.25	.16	12	.18	.15	18	.15	.15	120

We also explored the role of recall performance. Results showed a significant main effect of Recall performance on calibration without standards present β = -0.45, *t*(758) = -5.70, *p* < .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* = .25, *SD* = .16) calibrated better than average performers (*M* = .30, *SD* = .19) *t*(508) = -3.26, *p* = .001, and that average performers calibrated better than low performers (*M* = .36, *SD* = .20) *t*(532) = -4.05, *p* < .001. Thus, regardless of what type of standards students received, high performers calibrated better without standards present than low performers.

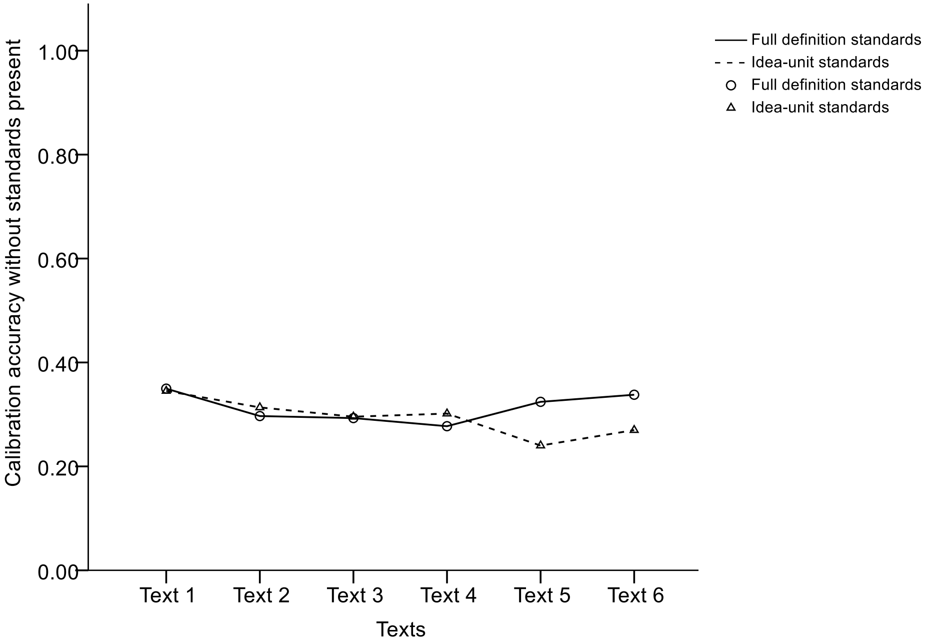


Figure 2. Calibration accuracy without standards present over texts by standard group.

There was a significant interaction between Recall performance and Texts β = 0.32, *t*(758) = 2.36, *p* = .018. Figure 3 depicts the curves of the different performance level groups over time. Whereas low performers *b* **=** -.02, *t*(758) = -2.82 [-.03, -.01], *p* = .005 and average performers *b* **=** -.01, *t*(758) = -2.08 [-.02, <-.01], *p* = .038 improved their calibration over texts, calibration of high performers remained stable *b* < .01, *t*(758) = 0.18 [-.01, .01], *p* = .856. However, the three-way interaction between Idea-units, Recall performance, and Texts was not significant β = 0.01, *t*(754) = 0.30, *p* = .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.

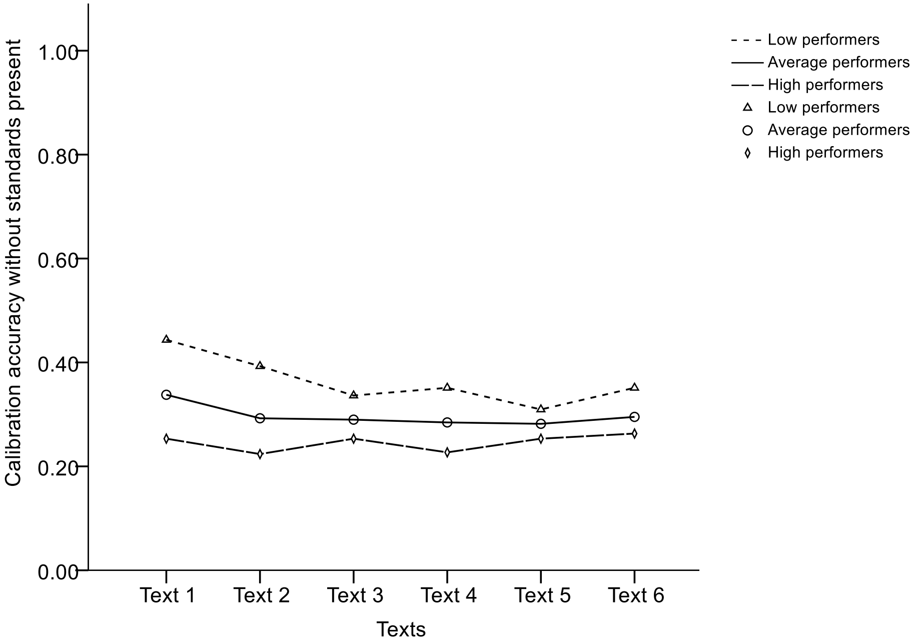


Figure 3. Calibration accuracy without standards present over texts by performance level group.

Discussion

In the current study, we investigated whether providing students with performance feedback in the form of 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., Submitted). 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, et al., 2005; Lipko et al., 2009; Nederhand et al., Submitted; 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., Submitted) 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 et al., 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, Was and Dunlosky (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, $F = .34$, $p = .34$. 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 et al., 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 et al., 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.

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

Chapter 4

The effect of performance standards and medical experience on diagnostic calibration accuracy

This chapter is published as:
Nederhand, M.L., Tabbers, H.K., Splinter, T.A.W., & Rikers, R.M.J.P. (2018). The Effect of Performance Standards and Medical Experience on Diagnostic Calibration Accuracy. *Health Professions Education*. doi:10.1016/j.hpe.2017.12.008

An earlier version of this chapter has received a Graduate School Award for PhD Excellence in 2015, by the Erasmus Graduate School of Social Science and the Humanities.

Abstract

Purpose. Medical doctors do not always calibrate accurately in terms of their diagnostic performance, which means that their evaluation of their diagnosis differs from their actual performance. Inaccurate calibration can lead to maltreatment and increased health care costs. This study was conducted to investigate whether calibration accuracy can be improved among both board certified medical specialists and medical students by providing them with a simple form of feedback (i.e., performance standards). We expected that performance standards would enhance calibration accuracy. Furthermore, we expected that medical specialists would overall be better calibrated than medical students.

Method. Medical specialists (n = 42) and medical students (n = 43) diagnosed three clinical cases and rated their own performance, after which they did or did not receive standards (i.e., the correct diagnoses). All participants were then tested: they had to diagnose three new cases and had to rate their performance without receiving diagnostic feedback.

Results. In support of our hypotheses, findings indicate that both students and specialists who received performance standards calibrated better than students and specialists who did not receive standards. Furthermore, medical specialists calibrated better than medical students.

Discussion. This study shows that providing simple forms of feedback constitute effects on calibration accuracy on new tasks.

Acknowledgements

The authors wish to thank all participants for their participation in our study. Many thanks to Liesbeth Dirksen-de Tombe and Lennart Rem for their valuable help and Amanda Binnendijk and John Tjen for their help during the training. Thank you Matthijs de Hoog for approving testing during the training sessions. Furthermore, the authors wish to thank Eugène Custers for providing us with the clinical cases. Finally, the authors like to dedicate their study to the memory of Prof. Dr. Splinter, one of our co-authors, who recently passed away after many years of illness. He has been a source of inspiration to all of us.

Introduction

Professionals often experience difficulties in adequately estimating the quality of their own diagnostic performance: they tend to overestimate themselves (Davis et al., 2006; Friedman et al., 2005; Martin, Regehr, Hodges, & McNaughton, 1998; Meyer, Payne, Meeks, Rao, & Singh, 2013). Providing an accurate estimate of one's performance is referred to as calibration accuracy (Dunlosky & Thiede, 2013a; Lichtenstein et al., 1982). Being able to calibrate accurately is especially important in a dynamic domain as medicine, where insights about proper treatment quickly change, and doctors have to make sure that they keep up with these developments to ensure the best treatment for their patients. However, in complex jobs, like that of a medical professional, calibration is difficult due to the little detailed and immediate performance feedback that is provided in their work (Ericsson, 2015). Consequently, such contexts increase the susceptibility to make decisions based on incorrect or outdated knowledge or skills, resulting in inefficient or ineffective treatment, increased health care costs, and most importantly, harm to the patient (Berner & Graber, 2008; Blendon et al., 2002).

Previous studies have therefore argued that medical professionals can greatly benefit from becoming better calibrators to increase their performance (Eva & Regehr, 2005). To date, however studies on how to improve calibration accuracy in medicine are scarce (Eva & Regehr, 2005). In the current study we aim to address this gap in the literature by investigating how calibration accuracy can be increased among medical professionals. We examined the effect of providing a simple form of performance feedback: performance standards. Studies in educational science have shown that such standards effectively help individuals become more aware of their performance (Dunlosky et al., 2011; Nederhand, Tabbers, & Rikers, Submitted; Rawson & Dunlosky, 2007). However, whereas existing studies both within and outside medicine predominantly focused on students' calibration who bring little experience to the task at hand, the present study involves participants with high levels of experience. So besides using medical students as participants, this study also investigates calibration accuracy of board certified medical specialists.

Improving calibration accuracy in medicine

Miscalibration is explained by the notion that individuals generally have difficulty estimating their performance when valid cues are absent (Koriat, 1997; Rawson & Dunlosky, 2007). To improve calibration accuracy, individuals should therefore be given the opportunity to compare their own performance to a standard (Dunlosky & Thiede, 2013a; Lipko et al., 2009; Rawson & Dunlosky, 2007). By comparing one's own performance to a standard, individuals gain insight in the match or mismatch of their estimated performance and actual performance (Martin et al., 1998; Sargeant et al., 2010). This awareness, in turn, improves calibration accuracy (Butler & Winne, 1995; Koriat, 1997; Rawson & Dunlosky, 2007; Zimmerman, 2000).

If providing standards indeed helps to enhance calibration accuracy, we might expect that individuals who receive standards also show better calibration accuracy on similar future tasks. That is, the capability to calibrate may transfer to new situations, where standards are not immediately available. In support of this assumption, two recent studies among psychology students showed that providing standards indeed improved calibration on a subsequent task where standards were not present (Nederhand, Tabbers, Abrahimi, & Rikers, Accepted; Nederhand, Tabbers, & Rikers, Submitted). To date, however, studies on how standards improve calibration in medicine are lacking. So based on findings outside medicine we ask the question whether providing medical professionals with standards would improve calibration on subsequent diagnostic tasks.

Experience differences in calibration accuracy

In studies on student calibration, top-performing students are typically compared to low-performing students (Kruger & Dunning, 1999). These studies showed that students who perform well are generally also better calibrated; low performers, however, show poor calibration and tend to overestimate their performance. Although these studies provide insights in calibration accuracy between students who differ from each other in terms of task performance, generalization to groups that differ more substantially in experience remains unclear. For example, whereas task performance of medical students within the same cohort can vary, their general diagnostic experience is relatively similar. This raises the question whether comparing an unexperienced group of medical students to highly experienced physicians—who have received much more feedback on the accuracy of their medical diagnoses over the years—leads to the same results.

To the knowledge of the authors, there is only one study conducted in medicine on calibration in which different experience levels were included. Friedman et al. (2005) investigated the calibration accuracy of medical students, internal medicine residents, and internists. Contrary to their expectation, results showed that instead of the internists, medical students were the most accurate calibrators. However, Friedman and colleagues questioned the validity of their findings because students had much more difficulty solving the cases than the residents and internists. They argued that it would have been better if the participants were challenged with less difficult cases, making the diagnostic task equally understandable for both the students and the specialists. The question therefore remains whether medical specialists would calibrate better than medical students when both groups are provided with diagnostic cases both the specialists and students can solve.

Present study

The present study investigates whether providing performance standards can enhance calibration accuracy on subsequent diagnostic tasks among medical specialists and medical students, and whether medical specialists calibrate better than medical students. The specialists and students were randomly divided in two groups: The first group received performance standards (the correct diagnoses) after diagnosing three clinical case, while the second group did not receive any standards. Subsequently, all participants received three entirely new cases and then their calibration was tested again. However, this time no standards were provided to both groups. To make sure the diagnostic tasks were equally understandable for both the students and the specialists, all participants were provided with general clinical cases, instead of cases from one type of specialty (e.g., internal medicine as in the study of Friedman et al.). The cases used in our study have been shown to be suited for both students and specialists (Custers, Boshuizen, & Schmidt, 1996; Custers, 1995).

We expected that providing standards on the first three cases would enhance calibration accuracy on the new cases. Furthermore, based on findings that high or experienced performers calibrated better than low or inexperienced performers, we expected medical specialists to calibrate better than the students.

Note the difference between improving diagnostic performance and calibration accuracy. Providing standards unlikely will improve diagnostic performance because this type of feedback is non-elaborate. We therefore do not expect to find any effect on diagnostic performance. Improving calibration accuracy is, however, considered a first step to eventually improve (diagnostic) performance: being aware of a mismatch between estimated performance and actual performance causes control behavior (Dunlosky & Rawson, 2012), such as requesting additional tests or asking help from colleagues if poor performance is detected. So the main focus of the current study is on calibration accuracy: knowing whether your diagnosis is accurate or not.

Method

Participants and design

Eighty-five participants were recruited, 43 second-year medical students and 42 medical specialists. The medical specialists (30 males, 12 females) were all board certified in their specialty and had a mean age of 44.73 (*SD* = 7.61). They were specialized in more than 20 different medical domains, such as internal medicine, neuroscience, or cardiology. Their mean years of clinical experience was 11.83 (*SD* = 7.95). The specialists were approached and tested during a professional training program by the Erasmus Medical Center and did not receive any compensation for their participation. The second-year medical students (18 males and 25 females) studied at the Erasmus University Medical School and were recruited during, and tested after, one of their lectures. Their mean age was 20.88 (*SD* = 2.40). By participating, the students could join a lottery to win four small prizes. All participants provided informed consent and the study was approved by our institutional review board.

Materials and procedure

Participants were randomly assigned to a group that received standards and one group that did not receive standards. The standard group consisted of 44 participants (22 specialists and 22 students) and the no-standard group consisted of 41 participants (20 specialists and 21 students). Participants received a total of six clinical cases. Each case consisted of a short clinical scenario in which the following information was given: A patient's medical history, present complaints, physical examination, and additional investigations (e.g., lab data, ECGs). Cases were presented in a booklet in which each case was printed on a separate page with some space intentionally left blank where the participants could write down their diagnosis and provide their performance estimate (see Appendix). With the exception of receiving standards or not, the procedure for the two groups was the same. All participants received a booklet with the cases that also contained a short introduction. After reading the introduction, participants continued with diagnosing the first three cases one by one. It was stressed that the participants should be as specific as possible in their diagnoses. After providing a diagnosis, participants rated the confidence they had in their diagnosis (i.e., their performance estimate) on a 10-point Likert scale (1: 'very unconfident' to 10: 'very confident').

Depending on the group they were in, participants received a performance standard on the next page after writing down each diagnosis. The standards consisted of the confirmed diagnosis for each of the first three cases, respectively: Aneurism of the aortic artery (threatening rupture); herpes zoster; and nervous abdominal pain. The participants in the control group (i.e., no-standard) were not informed about the correct diagnosis, but had to do a filler task (i.e., answer the question how familiar they were with the case).

After the intervention, all participants were tested on three completely new cases. The confirmed diagnoses associated with the test cases were: meningitis or encephalitis as a complaint of mumps; kidney stones (colic); and epidural hematoma. All cases have been used in previous studies (Custers et al., 1996; Custers, 1995). During the test phase, there were no differences between both conditions: all participants diagnosed the new cases without any standards, and had to provide a performance estimate after each diagnosis. The three test cases were used to investigate whether having received standards leads to improved calibration accuracy on new cases where these standards are missing.

Analysis

Diagnostic accuracy was scored by comparing the diagnoses of the participants with the confirmed diagnosis for each case (Custers et al., 1996; Custers, 1995). Diagnoses were scored on a three-point scoring grid (0 = incorrect, .5 partly correct, and 1 = fully correct) and the scoring was double checked by a medical specialist and a professor of internal medicine.(cf. Schmidt et al., 2014) There were no differences between raters.

A difficulty when using Likert scales is that participants do not tend to use the extreme response options (Landy & Conte, 2009). Because reluctance to use extreme response options leads to miscalibration in our experiment, we adjusted the Likert scale. Low confidence scores (1 to 3), were given the value of 0 and high confidence scores (8 to 10) were given the value of 1. The confidence scores that indicate average confidence in an answer (4 to 7), were given a value of .5. This converted the confidence scale into a three-point scale so it would correspond to the three-point diagnostic performance scale, enabling us to calculate calibration accuracy.

To calculate the calibration accuracy of the participants, the absolute difference between performance and confidence was calculated, and the average of the three difference scores was used as the calibration accuracy score (Lipko et al., 2009; Pajares & Graham, 1999; Pajares & Miller, 1997; Rawson & Dunlosky, 2007; Schraw, Potenza, & Nebelsick-Gullet, 1993). Perfect calibration accuracy was thus indicated by a score of 0 (perfect match between confidence and performance scores in all three cases) and strongest miscalibration had a score of 1 (largest mismatch between confidence and performance scores in all three cases).

We analyzed our data with IBM SPSS Statistics, version 23 (IBM, New York). To compare the feedback and experience conditions, a 2x2 univariate analysis of variance (ANOVA) was conducted with Standards (Yes vs. No) and Experience level (Specialists vs. Students) as independent variables and diagnostic performance and calibration accuracy as dependent variables. A significance level of .05 was set for all analyses.

Results

Diagnostic performance

We first checked diagnostic performance (a score between 0-1) of both specialists and students, and between the two experimental conditions. We tested whether medical specialists showed better diagnostic performance than the students over the total of six cases. Results showed that diagnostic performance differed significantly between medical specialists (*M* = .91, *SD* = .12) and medical students (*M* = .66, *SD* = .18), *F*(81) = 55.81, *p* < .001, η_p^2 = .408. Medical specialists solved more cases correctly than medical students. It is important to note, however, that although students' diagnostic performance was significantly lower than that of the specialists, their level of performance was still high (66%). So, as intended the cases used in our study were solvable for both specialists and students.

The diagnostic performance over all six cases did not differ between the group that received standards on the first three cases (*M* = .79, *SD* = .20) and the group that did not receive any standards (*M* = .79, *SD* = 0.20), *F*(81) = .03, *p* =.858, η_p^2 < .001. There was no statistically significant interaction on diagnostic performance between standards and experience level, *F* < 2.

Calibration accuracy among specialists and students

To test whether medical students and medical specialists differed in terms of their calibration accuracy, we analyzed whether there was a main effect of experience on calibration accuracy over all six clinical cases. The main effect of experience was statistically significant, *F*(83) = 46.32, *p* < .001, η_p^2 = .358, with medical specialists having a better calibration accuracy as indicated by the lower mean score on calibration accuracy (*M* = .19, *SD* = .14) than the students (*M* = .39, *SD* = .12). This result supports our hypothesis that specialists calibrate better than students.

Enhancing calibration by receiving standards

Furthermore, we analyzed whether providing standards was associated with better calibration accuracy on the last three test cases. The main effect of standards was statistically significant, $F(1,81) = 4.00$, $p = .049$, $\eta_p^2 = .05$ (see Table 1 for descriptives). Participants in the standard group calibrated better on the new test cases ($M = .18$, $SD = .17$) than participants who did not receive standards ($M = .26$, $SD = .22$). Our hypothesis that providing standards can improve calibration accuracy on subsequent tasks (i.e., calibration becomes closer to zero) is therefore supported. Finally, we tested the interaction between experience level and performance standards. This interaction was not statistically significant, $F(1,81) = .69$, $p = .41$, $\eta_p^2 = .01$. Standards similarly improved calibration accuracy for both specialists and students.

Table 1.
Mean calibration scores in the test phase. The range of the calibration scores is from 0 (perfect calibration) to 1 (no calibration)

		Calibration scores	
	n	*M* (*SE*)	95% CI
Standards			
2nd-year students	22	0.22 (0.04)	[0.14, 0.30]
Medical specialists	22	0.14 (0.04)	[0.06, 0.22]
Total	44	0.18 (0.03)	[0.13, 0.23]
No Standards			
2nd-year students	21	0.33 (0.04)	[0.25, 0.41]
Medical specialists	20	0.18 (0.04)	[0.10, 0.27]
Total	41	0.26 (0.03)	[0.19, 0.33]

Discussion

With this study we investigated the effect of feedback (i.e., performance standards) on calibration accuracy of board certified medical specialists and medical students. We hypothesized that medical specialists would overall calibrate better than medical students because the specialists have received over the years much more feedback on the accuracy of their medical diagnoses. Furthermore, research shows that standards can be used to enhance calibration accuracy because such standards help individuals to become aware of the (mis)match between their own performance and the required performance (Koriat, 1997; Rawson & Dunlosky, 2007). We therefore predicted that standards in the form of a correct diagnosis would enhance diagnostic calibration accuracy. Our results confirm both hypotheses and hence have several educational and theoretical implications.

Calibration accuracy among specialists and students

We tested whether medical specialists are more accurate calibrators than students. In support of our expectation, medical specialists calibrated better than medical students. The specialists had a mean of eleven years of experience with treating patients. Consequently, they had many years of experience with diagnosing patients and monitoring their diagnoses. Our results show that this experience helps specialists to adequately estimate their performance on clinical cases that are not necessarily in their own domain of expertise.

Although there are many studies that argue it is important to individually differentiate in calibration accuracy (e.g., Kruger & Dunning, 1999), little attention has been paid to differences in experience. When studies did include experience level, studies often focused on small variations. For example, students from different grades were compared (García, Rodríguez, González-Castro, González-Pienda, & Torrance, 2016). The current study therefore adds to the existing literature by included two groups that substantially differ in their clinical experience.

Standards to improve calibration accuracy

Besides investigating the effect of experience on calibration accuracy, we also tested the effect of standards. Findings indicate that receiving standards is associated with better calibration on new tasks (i.e., clinical cases). The results of this study are therefore in line with Nederhand et al. (Accepted), and among the first to show that providing standards can help individuals to also enhance their calibration on subsequent new tasks. Our results also show that standards helped students equally well as specialists to enhance their calibration accuracy.

As intended, whereas standards affected calibration accuracy, participants did not show better diagnostic performance after receiving standards. This is because using non-elaborated forms of feedback are little effective at improving performance directly (Archer, 2010). However, because standards are proven effective to enhance calibration, (diagnostic) performance can indirectly be improved (Dunlosky & Rawson, 2012). For instance, when a physician knows he or she performs poor (i.e., calibration is accurate while but diagnostic performance is low), he or she can take steps to overcome this poor performance by for example asking extra help. In other words, the calibration accuracy helps them to take steps that will ultimately improve their diagnostic performance. Vice versa, better calibration accuracy among physicians that already perform well is also beneficial. For example, if a physician performs very well but is unaware of that, he or she will request too many additional tests, costing both the hospital and patient time and money. It is therefore promising that our study shows that even simple forms of feedback help to improve calibration accuracy.

Limitations and future directions

While our study provides new insight in how to improve calibration accuracy in medicine, it also has some limitations. Although there are many theoretical and empirical reasons to assume that accurate calibration also enhances diagnostic performance, we did not investigate whether our participants would indeed engage in corrective actions after they received standards. An important direction for future research is therefore to investigate the steps that are taken after a mismatch between estimated performance and actual performance is detected.

A second limitation in our study is that although we made a first attempt to measure the longitudinal effect of standards, our diagnostic test cases followed directly on the first three diagnostic cases in which participants received standards. Our design thus provides insight in whether the capability to calibrate accurately can transfer to new cases, but the effect over time still remains largely unclear. For example, when we would have asked our participants to diagnose new cases one week after our intervention, would there still have been differences between groups? Future research could investigate this question. Related to this issue, the optimal amount of feedback could further be explored. For example, it is not unlikely that, instead of our three feedback moments, more feedback is needed to constitute effects over time.

Finally, we used both board certified medical specialists and medical students as participants in our study to investigate large experience differences. As intended, the students had some degree of expertise as they were able to solve the clinical cases. In other words, although students clearly differed from medical specialists in terms of diagnostic experience, they were no full novices on the task. It is therefore important to mention here that it remains unclear whether standards would also help full novices improve their calibration and our results must be treated with caution when generalizing to a group of novices.

Conclusion and implications

Because medicine is a dynamic domain, life-long-learning of clinicians is necessary. Being able to improve oneself continuously requires self-monitoring—clinicians have to be aware of their own performance quality so they can discriminate between things that go well things that have to be (further) developed. The current study highlights an under-studied topic in medical education: Although many studies on calibration accuracy were conducted with the aim to generalize to clinical practitioners, this group is hardly ever included as participants. We have shown that even a relatively simple feedback intervention in the form of correct diagnoses can help both medical specialists and medical students to improve their calibration accuracy on new diagnostic tasks (i.e., their awareness of their actual diagnostic performance).

Appendix

Description casus

01. Man, age 47, married, 3 children
02. Occupation: storekeeper
03. Medical history: bronchitis at age 30
04. Had his leg broken 6 years ago, as a consequence of a car accident
05. Four years ago: treated with medicaments for kidney stones
06. Some of his relatives are known to have coronary disease and diabetes mellitus
07. His wife rings up, asks the physician for an immediate visit:
08. Just like a few years ago, her husband is rolling across the room because of the pain
09. He is also vomiting almost continuously
10. When the physician arrives, the pain has just subsided. The patient is sitting on the sofa and recovering a bit
11. He complains about having had a convulsive abdominal pain abreast of the navel, at the left side
12. The pain is radiating to his groin
13. The pain emerges very suddenly, and then gradually subsides. During an attack he almost can't stand it
14. Earlier that day he had already seen some blood in his urine, but had had no pain at the time
15. He reports having measured 37.8° (Centigrade) temperature

What diagnosis would you give on basis of the previous information?

How confident are you that your diagnosis is correct?
Please encircle your estimation.

Very unconfident									*Very confident*
1	2	3	4	5	6	7	8	9	10

Chapter 5

Metacognitive awareness as measured by second-order judgements among university and secondary school students

This chapter is submitted as:
Nederhand, M.L., Tabbers, H.K., De Bruin, A.B., & Rikers, R.M.J.P. Metacognitive awareness as measured by second-order judgements among university and secondary school students.

Abstract

When compared to high performers, low performers generally have more difficulty to accurately estimate their own performance. This has been explained by low performers being both unskilled and unaware about their performance. However, Miller and Geraci (2011) found that low performing university students also assigned less confidence to their estimates (i.e., second-order judgments, SOJs), indicating some metacognitive awareness of their poor calibration. The current study examined whether the relationship between calibration accuracy and confidence in performance estimates is more general, and exists irrespective of performance level, not only for university students but also for secondary school students. We asked both university students and secondary school students to estimate their exam grade after taking their exam, and to provide a second-order judgement). The results showed that for university students, poor calibration accuracy was indeed accompanied by low confidence scores, independent from performance level, confirming our hypothesis. For secondary school students however, calibration accuracy was unrelated to confidence scores, suggesting a less developed metacognitive awareness.

Introduction

Many students are generally overconfident about their own performance (De Bruin, Kok, Lobbestael, and De Grip 2017; Dunning, Johnson, Ehrlinger, and Kruger 2003; Ehrlinger and Dunning 2003; Kruger and Dunning 1999; Pennycook, Ross, Koehler, and Fugelsang 2017; Sanchez and Dunning 2018). Not being able to accurately estimate the level of one's performance can have far-reaching implications, as performance estimates influence control decisions made during the learning process (Nelson and Narens 1990). In the case of overconfidence, students may prematurely end studying because of an incorrect conviction they have mastered the materials (Bol, Hacker, O'Shea, and Allen 2005; Dunlosky and Rawson 2012). Especially low-performing students seem to suffer from inaccurate self-monitoring, mostly from overconfidence. High performing students are often much better calibrated, although they may show some underconfidence. This difference in over- and underconfidence among low and high performers is dubbed the Dunning-Kruger effect (Dunning et al. 2003; Kruger and Dunning 1999). The most dominant explanation for the Dunning-Kruger effect, originally stated by Kruger and Dunning in 1999, is that low performers lack metacognitive awareness because of a 'double curse': They lack the skills to conduct a task, and because of this incompetence, they are also unaware of the level of their performance.

But are students who show large miscalibration really obvlious to their own performance? Miller and Geraci (2011) argued that focusing merely on performance estimates gives an incomplete image of students' metacognitive awareness. Instead, they argued that calibration research should focus on both functional and subjective confidence. Miller and Geraci defined functional confidence as students' performance estimate, and subjective confidence as the confidence judgement assigned to a given performance estimate. In their study, Miller and Geraci asked university students not only to estimate their grade before taking an exam, but also to provide a 'second-order judgement' (SOJ; see Dunlosky, Serra, Matvey, and Rawson 2005): rating the confidence they had in their estimate. In line with previous research, Miller and Geraci found that low performers showed a larger mismatch between their estimated grade and actual grade than high performers. At the same time however, low performers reported less confidence in the accuracy of their estimates than high performers. According to Miller and Geraci, low performers thus seemed to have a certain level of metacognitive awareness. Following up on the research of Miller and Geraci (2011), Händel and Fritzsche (2016) investigated the alignment between SOJs and calibration accuracy on item-by-item judgements and global judgements (i.e., judgements regarding performance on an entire task). Händel and

Fritzsche also found a positive relation between performance level and SOJs. Low performers, who were found to be poorly calibrated, were again less confident in their performance estimates than high performers.

Hence, the studies of Miller and Geraci (2011) and Händel and Fritzsche (2016) showed that low performers seem to be aware of their poor calibration accuracy. This finding had implications for the explanation of the Dunning-Kruger effect, because qualifying low performers as fully unaware of their performance seemed to be incorrect. However, in a follow-up study, Fritzsche, Händel, and Kröner (2018) reanalyzed the data from Händel and Fritzsche (2016), and found that low performing students seemed to have a general tendency to provide low confidence judgements. This finding raises the issue whether it is performance level that determines whether students provide high or low subjective confidence scores, or whether it is really metacognitive awareness. If the latter were the case, you would expect subjective confidence to be directly related to calibration accuracy, with students high on miscalibration providing low SOJs, and students low on miscalibration showing high confidence, irrespective of performance level. In that case, students would be metacognitively aware, with confidence judgements aligned to their calibration accuracy.

However, none of the previous studies actually examined the direct relation between calibration accuracy on the one hand, and SOJs on the other, independent from performance level. Furthermore, previous studies have been exclusively conducted among university students (Fritzsche et al. 2018; Händel and Fritzsche 2016; Miller and Geraci 2011). In general, adult students have been found to be more accurately calibrated than teenage students, such as secondary school students (Lockl and Schneider 2002; Van der Stel and Veenman 2010). A possible reason for this difference is that metacognitive awareness is still developing during childhood and adolescence (Paulus, Tsalas, Proust, and Sodian 2014), after which it plateaus in adulthood (Weil et al. 2013). Just as in university, Dunning-Kruger effects have been found for younger students as well (e.g., Finn and Metcalfe 2014; Labuhn, Zimmerman, and Hasselhorn 2010) in terms of functional overconfidence. However, it is yet an open question whether the findings on university students' subjective confidence can be generalized to younger secondary school students, or that secondary school students are less metacognitively aware, and thus show less alignment between their subjective confidence and actual calibration accuracy.

Present study

The current study thus aimed to answer two research questions: (1) are SOJs related to calibration accuracy, independent from performance level?; and (2) does the negative relation between SOJs and miscalibration hold for both university students and secondary school students?

To investigate the relation between subjective confidence and calibration accuracy, we conducted two studies in which students were asked to estimate their exam grade and to provide a confidence score of their estimate (subjective confidence as measured by a SOJ) after taking their exam. In the first study, we examined the alignment between SOJs and calibration accuracy among university students who took an exam in educational psychology (cf. Miller and Geraci 2011). The second study was conducted among secondary school students who took exams in three different subjects: French, German, and Mathematics.[1] For each exam, we investigated the alignment between calibration accuracy and SOJs.

Based on Miller and Geraci, we expected a significant negative relation between SOJs and calibration accuracy. That is, higher miscalibration would be associated to lower confidence. Furthermore, because developmental differences might affect metacognitive awareness, we expected that secondary school students would be less metacognitively aware and thus the relation between calibration accuracy and SOJs would be less strong in this sample.

Finally, we checked whether we could replicate the association between performance level and SOJ that Miller and Geraci (2011) found. Similar to Miller and Geraci, we divided our students into different quartiles based on performance level, to examine the differences in SOJs among high and low performers. Based on the assumption that low performers indeed calibrate worse than high performers (cf. the Dunning-Kruger effect), we expected low performers to also provide lower confidence scores.

Method

Participants

Two-hundred-and-ninety-four second-year psychology students and 388 secondary school students from the same urban area in the Netherlands were recruited for this study. Confidentiality of the performance estimates was ensured to all students, and they all provided informed consent that their estimates and grades could be used for this study. Ethical approval was obtained for this study by the Ethical Committee of the Department of

[1] The secondary school students participated in an intervention study in which we trained their calibration accuracy over time. The data reported in this study consist of their pre-measurement scores, before they took part in the intervention.

Psychology, Education, and Child Studies at our university. Two-hundred-and-fifty-one university students provided informed consent (85.37% response rate). The secondary school students provided informed consent together with their parents, leading to a final sample of 302 students (77.84% response rate).

Materials

Exams. We used students' regular exams to measure calibration accuracy. The university students had an end-of-term exam in educational psychology, comparable to the exams used in most research on calibration accuracy with university students (Bol and Hacker 2001; Hacker, Bol, Horgan, and Rakow 2000; Miller and Geraci 2011; Nietfeld, Cao, and Osborne 2006). During their academic year in which they took 8 exams, they could resit a maximum of 2 exams. The secondary school students had regular end of term exams in French, German, and Math. None of the students were allowed to resit unless they were sick when the exam was administered.

Procedure

The procedure was the same for all students. After taking their exam, students received a form on which they estimated the grade they thought they would obtain on a 10-point scale with decimal points (i.e., most common scoring scale in the Netherlands with 10 representing the highest possible score and 1 the lowest). Students also gave a SOJ by indicating the confidence they had in their performance estimate on a five-point scale, ranging from not confident to highly confident (cf. Händel and Fritzsche 2016; Miller and Geraci 2011).

Analyses

Calculating measurements

Calibration accuracy was calculated by taking the absolute difference between estimated exam grade and actual exam grade (Dunlosky and Thiede 2013; Schraw 2009). This means that the higher the difference between the estimated and actual grade, the higher the miscalibration. We also calculated bias scores to investigate the direction of the miscalibration (i.e., over- or underconfidence). Bias scores consisted of the difference between estimated grade and actual grade (Dunlosky and Thiede 2013; Schraw 2009).

Statistical analyses

Our research question was whether subjective confidence was aligned with calibration accuracy scores. To test this question, we investigated Pearson correlations between SOJs and calibration accuracy for each exam. As a check, we also investigated whether we could replicate the findings of Miller and Geraci (2011). To do so, we divided the university students into four quartiles based on their exam performance. We tested whether SOJs differed between low performers and high performers (cf. Miller and Geraci 2011) using an ANOVA with SOJ as dependent variable and performance level as independent variable.

Results

Table 1 shows the descriptives of the variables under study: Bias scores, calibration accuracy, SOJs, and final exam grades. In the following sections, our hypotheses on calibration accuracy are tested. In all our analyses, a significance level of .05 was used.

As a first check, we examined whether we could replicate the findings of Miller and Geraci (2011), by investigating whether possible differences in subjective confidence within our samples could be explained by performance level differences. Table 2 presents the means and standard deviations of the four different performance level groups on calibration accuracy, bias, SOJs, and performance, both for university students on their Educational Psychology exam and for secondary school students for their French, Math and German exam. We performed an ANOVA with SOJ as dependent variable and performance level (i.e., the performance quartiles) as independent variable (cf. Miller and Geraci 2011). Results were not significant: subjective confidence ratings did not differ between the performance quartiles for the university students, $F(3, 251) = 0.07$, $MSE = 0.05$, $p = .978$, $\eta_p^2 = .001$, nor for the secondary school students on the courses French, $F(3, 289) = 1.01$, $MSE = 0.70$, $p = .387$, $\eta_p^2 = .010$, Math, $F(3, 287) = 0.42$, $MSE = 0.82$, $p = .741$, $\eta_p^2 = .004$, and German $F(3, 298) = 0.44$, $MSE = 0.56$, $p = .725$, $\eta_p^2 = .004$. This indicates that, contrary to Miller and Geraci’s findings, subjective confidence scores in our samples did not differ between high and low performing students.

Table 1.

Descriptives of the university and secondary school students

	N	*M*	[min, max]	*SD*
University students				
Bias	251	-0.67	[-4.00, 2.10]	1.10
Calibration	251	1.05	[0.00, 4.00]	0.74
SOJ	251	2.98	[1.00, 5.00]	0.86
Final grade	251	7.00	[1.90, 9.70]	1.32
Secondary school students[1]				
French				
Bias	295	-0.19	[-4.70, 3.10]	1.26
Calibration	295	1.02	[0.00, 4.70]	0.75
SOJ	295	3.23	[1.00, 5.00]	0.83
Final grade	295	6.67	[3.00, 9.70]	1.45
Math				
Bias	291	-0.27	[-5.10, 4.80]	1.57
Calibration	291	1.24	[0.00, 5.10]	1.00
SOJ	291	3.34	[1.00, 5.00]	0.90
Final grade	291	6.86	[3.00, 10.00]	1.47
German				
Bias	302	0.58	[-2.10, 3.20]	1.18
Calibration	302	1.07	[0.00, 3.20]	0.76
SOJ	302	3.32	[1.00, 5.00]	0.75
Final grade	302	6.32	[3.00, 9.40]	1.36

Note. This Table presents the bias scores, on a scale of -10 to 10; calibration scores, on a scale of 0 to 10; SOJ scores, on a scale of 1 to 5; and performance scores, on a scale of 1 to 10, in which a score of ≥5.5 indicates that students passed the exam.

[1] Secondary school students had to take three exams (French, German, Math), but due to sickness or personal reasons, students sometimes failed to be present at all three exams. The *N* therefore varies between subjects.

Relation between calibration accuracy and second-order judgements among university students

More importantly, however, we examined the direct relation between calibration accuracy and SOJs, irrespective of the level of performance. Results among university students showed a significant negative correlation between subjective confidence (SOJ) and calibration accuracy $r = -.197$, $p = .002$, in line with our hypothesis. Figure 1 shows that less confident students had higher miscalibration, and vice versa: confident students were shown to be better calibrated.

Relation between calibration accuracy and second-order judgements among secondary school students

Thirdly, we tested whether the direct relation between calibration accuracy and SOJs also holds among secondary school students. In contrast to university students, results showed no significant correlation between SOJs and calibration accuracy among secondary school students for French *r* = -.029, *p* = .616; Math *r* = .036, *p* = .539; and German *r* = .059, *p* = .304. Hence, in line with our expectation, secondary school students' calibration accuracy and SOJs did not seem to be directly related: more miscalibration was not accompanied by lower confidence scores.

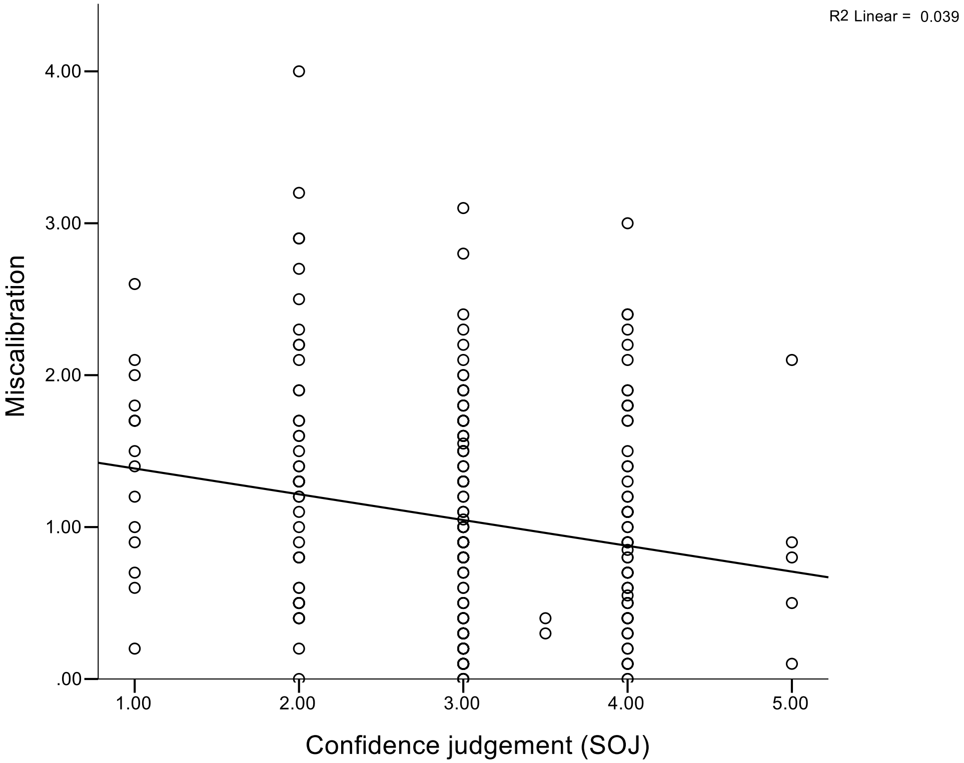


Figure 1. The negative relation between second-order judgements and miscalibration among university students. Low confidence judgements are related to high miscalibration, whereas high confidence judgements are related to lower miscalibration

Table 2.
Performance quartiles among the university and secondary school students

		Bias	Calibration	SOJ	Performance
Quartiles	*N*	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
		University students			
1	68	0.37 (0.89)	0.79 (0.54)	2.96 (1.00)	5.37 (0.84)
2	63	-0.53 (0.81)	0.73 (0.62)	2.97 (0.90)	6.69 (0.26)
3	66	-1.03 (0.72)	1.09 (0.64)	3.02 (0.83)	7.61 (0.25)
4	54	-1.70 (0.80)	1.70 (0.80)	3.00 (0.67)	8.69 (0.46)
		Secondary school students			
French					
1	82	0.63 (1.22)	1.14 (0.76)	3.17 (0.94)	4.90 (0.73)
2	70	0.04 (1.22)	0.97 (0.73)	3.17 (0.92)	6.22 (0.24)
3	70	-0.52 (0.98)	0.84 (0.71)	3.20 (0.71)	7.26 (0.33)
4	73	-1.00 (0.91)	1.11 (0.78)	3.38 (0.72)	8.53 (0.48)
Math					
1	74	0.80 (1.44)	1.36 (0.93)	3.42 (1.03)	4.92 (0.84)
2	82	-0.05 (1.40)	1.11 (0.84)	3.24 (0.92)	6.55 (0.31)
3	51	-0.70 (1.27)	1.07 (0.96)	3.39 (0.80)	7.39 (0.16)
4	84	-1.16 (1.39)	1.36 (1.19)	3.35 (0.81)	8.55 (0.56)
German					
1	77	1.31 (1.04)	1.45 (0.83)	3.19 (0.83)	4.77 (0.51)
2	74	1.20 (0.92)	1.29 (0.79)	3.39 (0.70)	5.57 (0.26)
3	81	0.26 (0.95)	0.78 (0.59)	3.31 (0.77)	6.85 (0.38)
4	70	-0.52 (0.76)	0.75 (0.53)	3.41 (0.67)	8.22 (0.51)

Note. This Table presents the bias scores, on a scale of -10 to 10; calibration scores, on a scale of 0 to 10; SOJ scores, on a scale of 1 to 5; and performance scores, on a scale of 1 to 10, in which a score of ≥5.5 indicates that students passed the exam.

Discussion

The current study focused on the relation between calibration accuracy and subjective confidence (as measured by SOJs) among university and secondary school students. We tested the hypothesis that students, regardless of their performance level, are metacognitively aware of their calibration accuracy. To do so, we examined the correlations between subjective confidence and miscalibration. Furthermore, we tested whether a relation between subjective confidence and miscalibration could be found both among university students and secondary school students.

Calibration accuracy and SOJs among university students

In line with our expectations, university students' subjective confidence is directly related to calibration accuracy, meaning that their confidence judgements are somewhat aligned to the actual quality of their performance estimates. Our results extend previous research findings on subjective confidence (Händel and Fritzsche 2016; Miller and Geraci 2011). While the studies of Miller and Geraci (2011) and Händel and Fritzsche (2016) were initially aimed at investigating subjective confidence among different performance level groups only, our study shows that their findings generalize to students with poor calibration accuracy in general. That is, the current study shows that students assign less confidence to their judgements when they are less well calibrated. Hence, regardless of performance level, students show some awareness of their calibration accuracy by providing high or low SOJs. This indicates that metacognitive awareness as measured by SOJs does not (only) seem to depend on performance level, as can be concluded from the findings from Miller and Geraci and Händel and Fritzsche, but rather seems to hinge on calibration accuracy.

In line with the above reasoning that SOJs are more dependent on calibration accuracy than on performance level per se, we do not find any differences in SOJs between different performance quartiles in our sample. However, this also means that we do not directly replicate the findings from Miller and Geraci (2011). A possible explanation may be that in our study, the relation between performance level and calibration accuracy is different from previous studies. Although we find that on average, high performers are underconfident while low performers are (slightly) overconfident, in line with the typical Dunning-Kruger effect, our results show that overall, underconfidence is a larger 'problem' than overconfidence: 3 out of the 4 performance quartile groups show underconfidence. Furthermore, unusual for calibration research, pairwise comparisons even show that the 25% best performing students calibrate slightly worse (M = 1.70, SD = 0.80) than the 25% lowest performing students (M = 0.79, SD = 0.54), p < .001. It is yet unclear why high-performing students in our sample are so underconfident and thus miscalibrated, compared to students from previous studies. Nevertheless, our data confirm our hypothesis that confidence level is directly related to calibration accuracy. This further strengthens the generalizability of the idea that university students base their confidence judgements on the accuracy of their performance estimate.

Calibration accuracy and SOJs among secondary school students

Whereas university students' subjective confidence is related to calibration accuracy, the secondary school students' confidence judgements do not relate to their actual calibration accuracy. This finding seems to be rather robust, as we do not find a relationship in either of the three exams (French, German, and Math). So it seems that our findings with university students do not straightforwardly generalize to secondary school students. This is in line with the notion that metacognitive awareness in secondary school is still under development (Schneider 2008; Weil et al. 2013). Whereas university students seem to have some metacognitive awareness as they assign little confidence to incorrect estimates, secondary school students seem less able to recognize poor calibration accuracy. This also has an interesting practical implication: when the aim is to improve students' calibration accuracy, secondary school students may first need to be stimulated to reflect on the quality of their performance estimates. Helping secondary school students to become more aware of the poor alignment between their confidence scores and their calibration accuracy, may set the stage for improvements in their overall performance monitoring. Interventions that aim to enhance calibration accuracy in secondary school may therefore focus not only on functional confidence, but on subjective confidence as well.

Limitations and future directions

Our study provides new insight in the relation between calibration accuracy and SOJs, both among university and secondary school students. However, our study has also some limitations. Although we find differences in the relation between confidence judgements and calibration accuracy, direct comparisons between the two samples are confounded by many variables. For example, in university, students' exam grades decide whether they will pass or fail the course and can thus be considered a high-stake exam. In secondary school, the grades add up to one another to form a grand mean at the end of the year. At the same time, however, university students can do a resit whereas secondary school students cannot. It is possible that such differences influence the judgements students make, and hence, their subjective confidence scores. In the current study, we cannot rule out this possibility because we used the exams naturally present in the classrooms. To get more insight in this topic and to enable better comparison between secondary and university students, future research should further examine the role of the task and its weighing on the relation between SOJs and calibration accuracy.

That we do not find significant differences in subjective confidence scores among students of different performance quartiles also contradicts the recent finding that low performers have a general tendency to provide lower confidence judgements than high performers (Fritzsche et al. 2018). It is important to note that our study differs from the study of Fritzsche et al. (2018) in several aspects. Perhaps most importantly, we ask students their exam grade while Fritzsche et al. asked students to estimate performance while making multiple choice questions. Their calibration accuracy was therefore based on a dichotomous scale (correct-incorrect), whereas in our study, a continuous scale was used. To facilitate comparison between studies, future research should focus on the question whether differences in scaling impact confidence judgements and calibration accuracy.

Conclusion

To conclude, this study shows that university students' calibration scores are related to their second-order judgements: independent from performance level, students with poor calibration accuracy show less confidence in their estimates and vice versa. Secondary school students, however, provide confidence judgements that are not related to their actual calibration accuracy. Hence, to further understand metacognitive awareness, we encourage future research to take into account measures of both functional and subjective confidence (SOJs). Furthermore, as younger students show less metacognitive awareness, they may benefit less from an intervention that only targets their functional overconfidence, but might instead flourish under an intervention that also aims to align their subjective confidence.

Chapter 6

Outcome feedback and reflection to improve calibration of high school students: A longitudinal study

A version of this chapter is submitted as:
Nederhand, M.L., Tabbers, H.K., Jongerling, J., & Rikers, R.M.J.P. Outcome feedback and reflection to improve calibration of high school students: A longitudinal study.

An earlier version of this chapter has received a Graduate School Award for PhD Excellence in 2017, by the Erasmus Graduate School of Social Science and the Humanities.

Abstract

The potential of outcome feedback such as grades has been suggested before, but research falls short at examining the conditions under which this feedback effectively improves calibration accuracy. This study examined how students' estimates of their own performance (i.e., calibration accuracy) could be improved during an entire school year. Within a secondary school, we implemented an intervention in which three groups of students received different levels of support in estimating their performance on their French exams. For each exam, the first group practiced with making performance estimates, the second group had to estimate their performance and calculate differences between their estimates and actual outcome, and the third group had to reflect on their estimates and also on reasons for a possible mismatch. We expected that increasing support would relate to more improvement in calibration accuracy. Results showed that having students estimate their exam performance during a school year helped them to improve their calibration accuracy, regardless of whether support was increased. The findings contribute to the discussion of how students can be supported to use outcome feedback when improving their calibration accuracy.

Acknowledgements

This research was funded by a Research Excellence Initiative grant from the Erasmus University Rotterdam. We owe many thanks to the immense support of the board and teachers of the school in which we conducted our study. We specifically mention Wilma Lambregts, Peter van Wijk, Floris Yperlaan, and Christiaan van der Ven. Furthermore, we would like to thank all the students who participated in this study. In addition, we would like to thank Homaira Abrahimi for helping us with the data input.

Introduction

Students generally tend to provide performance estimates that are far off their actual performance (Kruger & Dunning, 1999; Sheldon, Dunning, & Ames, 2014). This can be the cause of problems while studying, as students may not recognize the need to change ineffective learning strategies or fail to ask for help (Dunlosky & Rawson, 2012; Nelson & Narens, 1990). Especially since students become increasingly in charge of their own learning at all levels of education (Trilling & Fadel, 2009; Wolters, 2010), scholars have urged for more understanding of how to improve the judgements of own learning and performance (De Bruin & van Gog, 2012). The quality of such judgements are defined by calibration accuracy: the match between estimated and actual performance (Alexander, 2013; Lichtenstein & Fischhoff, 1977). For example, a student who thinks s/he has obtained an A on an exam and indeed scored an A, is perfectly calibrated.

Providing performance feedback is essential to help students become aware of their miscalibration (Butler & Winne, 1995; Stone, 2000; Zimmerman, 2000). A form of performance feedback often provided in education is outcome feedback. Students take many exams and receive lots of outcome feedback in the form of grades from their secondary school years onwards. This should enable students to improve their calibration accuracy over time, but recent findings have showed that this improvement fails to occur even after many feedback moments (Foster, Was, Dunlosky, & Isaacson, 2017). Is providing outcome feedback ineffective or do students need perhaps more guidance to adequately use and interpret outcome feedback, and what should this guidance consist of? By providing insight into these questions, this study contributes to the understanding of how students can be supported to improve their calibration accuracy after receiving outcome feedback. To achieve this, we conducted a year-long longitudinal study to test various interventions with outcome feedback within a secondary school-setting.

The effectiveness of outcome feedback to improve calibration accuracy

Koriat (1997) argued that when estimating performance, students use cues such as how much information they can recall (Baker & Dunlosky, 2006) or how familiar the test items are to them (Metcalfe & Finn, 2012). The quality of the performance estimates depends on the validity of these cues. Instead of using valid cues, however, students often use unreliable and irrelevant cues when estimating their own performance (Baker & Dunlosky, 2006; Gutierrez de Blume, Wells, Davis, & Parker, 2017; Serra & DeMarree, 2016; Thiede, Griffin, Wiley, & Anderson, 2010), limiting their ability to improve calibration accuracy over time. This is especially illustrated by Foster et al. (2017) who asked students to estimate their performance during an educational psychology course. Although students made performance estimates on thirteen exams throughout the course, they did not show any improvement in calibration accuracy over time. This lack of improvement was due to invalid cue use: students anchored their estimates on their prior performance estimates, but these prior performance estimates were not predictive of their future performance. Foster and colleagues noted that it would have been more helpful if students would anchor their performance estimates on their actual prior performance. In other words, students should have attended to the outcome feedback after each exam, and should have used this feedback when making a new estimate on a subsequent exam.

Outcome feedback is generally not considered to be the most beneficial form of feedback (Nicol & Macfarlane-Dick, 2006)—it does not inform students about how well they performed on each individual question or task, what the correct answers should have been, and what students can do to improve their performance on new tasks. However, at the same time, outcome feedback can be used to make students aware of their level of miscalibration by giving them the opportunity to compare their initial estimate to their actual performance (Butler & Winne, 1995; Labuhn, Zimmerman, & Hasselhorn, 2010; Zimmerman, 1990). Furthermore, a practical benefit of outcome feedback is that it is naturally given in many educational settings in the form of grades and implementing an intervention with outcome feedback is therefore feasible for schools.

Because of the practical advantages of using outcome feedback in a classroom setting, its effect on calibration accuracy has been investigated in several studies (Bol & Hacker, 2001; Foster et al., 2017; Hacker, Bol, & Bahbahani, 2008; Hacker, Bol, Horgan, & Rakow, 2000; Nietfeld, Cao, & Osborne, 2006). These studies have showed that providing only outcome feedback does not seem to be sufficient (Bol, Hacker, O'Shea, & Allen, 2005; Brown, Andrade, & Chen, 2015; Foster et al., 2017; Huff & Nietfeld, 2009). For outcome feedback to be effective at improving calibration accuracy, it seems necessary to actively encourage students to use the outcome feedback and to actively focus on the (mis)match between their estimated and actual performance (Brown et al., 2015; Hacker et al., 2000; Miller & Geraci, 2011; Nietfeld et al., 2006). However, until this day, it is hard to decide on how students should be encouraged.

There is only a limited number of studies investigating calibration accuracy in a classroom setting, but these studies also manipulated different variables simultaneously, making it hard to compare interventions and to distinguish ineffective from effective elements. For example, to encourage students to attend to their performance and calibration, Hacker et al. (2000), Huff and Nietfeld (2009), and Nietfeld et al. (2006) provided their students with training in their monitoring skills. That is, besides estimating their performance, students had to reflect on how well they understood the materials, identify their strengths and weaknesses, and reflect on why their estimates did or did not correspond to their actual performance. In contrast, Miller and Geraci (2011) and Callender et al. (2015), only asked students to reflect on the mismatch between estimated performance and actual performance, and provided students with rewards for accurate calibration. Hence, the question remains whether the effects were due to the different reflection prompts, rewards, or even the practice of estimating one's own performance. Furthermore, most studies have been conducted in an educational psychology course (Foster et al., 2017; Nietfeld et al., 2006), during which students were sometimes specifically taught about overconfidence (Callender, Franco-Watkins, & Roberts, 2015; Hacker et al., 2000; Miller & Geraci, 2011). Following such a course may in itself lead to an effect as well. Consequently, when aiming to improve calibration accuracy with outcome feedback, many questions remain unanswered as to how to do so.

Improving calibration with outcome feedback with increasing support

To systematically examine how we could improve calibration accuracy in a secondary classroom setting with outcome feedback, we distinguished between different levels of feedback support. First, to help students become aware of the accuracy of their performance estimates, it is necessary that students actually judge their performance. Second, after receiving outcome feedback, students should realize whether there is a mismatch between their initial estimate and their actual performance. Third, when students recognize that their estimate is inaccurate, it would help if students reflect on the cause of their miscalibration, in order to be able to improve their performance estimates on future tasks (De Bruin, Kok, Lobbestael, & De Grip, 2017; Hacker et al., 2000; Nietfeld et al., 2006).

Estimating performance

The first element that we distinguished is estimating performance. Improvements in calibration accuracy can only be tested when initial performance estimates can actually be compared to the final performance. It is, however, possible that making performance estimates in itself may already improve calibration accuracy. For example, Sadler (1989) stated that, to improve monitoring, students need 'evaluative experience' (p. 135). After which Boud (2013) proposed that to help students become better calibrated, they need to engage in monitoring practice, like any other expertise would require (e.g., Ericsson, Krampe, & Tesch-Romer, 1993). Supporting the assumption that practicing with making performance estimates could benefit calibration accuracy, such practice was included in successful interventions that enhanced calibration accuracy after providing outcome feedback (Hacker et al., 2000; Miller & Geraci, 2011; Nietfeld et al., 2006). However, the effect of practicing could often not be independently established, and the two studies that did look at the sole influence of practice with estimating performance, showed that such practice did not have any effect on calibration accuracy (Bol et al., 2005; Foster et al., 2017). Furthermore, the studies described above were conducted among university students in an educational psychology course which may create generalization problems to a secondary education context. Younger students are often less metacognitive aware than adult university students (Paulus, Tsalas, Proust, and Sodian 2014; Weil et al. 2013). On the one hand, secondary school students may benefit more from practice because they may have more room for improvement. However, given that they may be less metacognitive aware, secondary school students may not benefit from such a small intervention. Therefore, the question remains whether practicing may be beneficial for secondary school students to improve their calibration accuracy. To provide further insight in the effect of practicing, we included a group of students in the current study who solely practiced with estimating their own performance.

Comparing performance estimates to actual performance

For outcome feedback to be effective at improving calibration, the feedback must make students more aware of the mismatch between their estimated and actual performance. When students are aware of this mismatch, they obtain an important cue of their calibration accuracy. Hence, in an attempt to increase metacognitive awareness, a variety of studies have encouraged students to focus on the match between their estimated and actual grades (Callender et al., 2015; Hacker, Bol, & Bahbahani, 2008; Miller & Geraci, 2011, Nietfeld et al., 2006), or have directly provided students with comparison scores by subtracting the actual performance score from the estimated score (Miller & Geraci, 2011; Nietfeld et al., 2006; Stone & Opel, 2000). However, results on the effectiveness of such comparison scores are mixed (Callender et al., 2015; Huff & Nietfeld, 2009; Miller & Geraci, 2011; Nietfeld, 2009; Stone & Opel, 2000). Callender et al. (2015) showed that students who could compare their estimates to the final outcomes were better calibrated on the final exam of the course than students who did not have this opportunity. In contrast, Hacker and colleagues (2008) also included comparison scores in interventions that showed no improvement. Given these mixed findings and a lack of research among secondary school students, the question remains whether providing students with comparison scores would be an effective way to help them improve their calibration accuracy. To investigate the added effects of comparison scores on calibration accuracy, we therefore included a group of students who, besides estimating their performance, also calculated comparison scores between their estimate and actual outcome.

Reflecting on the outcome and miscalibration

When providing grades, students can signal a mismatch between their estimated and actual performance. Consequently, it could be helpful if students had the opportunity to reflect on possible reasons for this miscalibration. Doing so could help them to become more aware of the (invalid) cues they used, and this awareness could, in turn, help them to look for ways to improve their estimates (Nelson & Narens, 1990; Zimmerman, 2000). Given the potential of reflection after receiving outcome feedback, reflection prompts have been included in several studies (De Bruin et al., 2016; Hacker et al., 2000; Huff & Nietfeld, 2009; Nietfeld et al., 2006). For example, Hacker et al. (2000) encouraged their students to reflect on possible reasons for a mismatch between performance estimates and actual outcomes, and how such a mismatch could be prevented in the future. Although this intervention was associated with enhanced calibration on the final exam, results showed that this improvement only occurred for high performing students. Differences between high and low performers after reflection were also found in a study by Hacker et al. (2008), in which reflection on the mismatch between estimated and actual performance made calibration even worse for low performing students. So, whereas reflection was beneficial to improve calibration accuracy (De Bruin et al., 2016; Hacker et al., 2000; Huff & Nietfeld, 2009; Nietfeld et al., 2006), its results were not straightforward, were dependent on performance levels, and again, were solely examined in university student samples. Hence, the question remains whether providing students with reflection prompts could help them to improve their calibration accuracy on new tasks. To test the added effect of reflection, we included a group of students who first estimated their exam grades, then calculated comparison scores, and as a final step, reflected on their calibration accuracy. Furthermore, we looked at interactions with performance level.

Present study

Providing students with outcome feedback shows potential, but the conditions under which it can effectively improve calibration accuracy remain unclear. The current study examined how outcome feedback can effectively help students improve their calibration accuracy in a secondary school setting. This setting was chosen because of practical advantages. Students have the same teacher during the year and the knowledge tested on each exam accumulates. During an entire school year, we asked students after each exam to estimate their grade and to rate the confidence they had in this estimate. We did so for the French as a foreign language course. The reason we chose this course, as indicated, was because the teacher was the same for all intervention groups, exams were administered in a controlled setting in which students took the exams at the exact same time, and scoring was decided upon beforehand based on cohort results of prior years and guidelines laid down in the books and learning materials used for the French course.

In our intervention, we systematically varied the level of support. Our first group only estimated their exam grade (Practice group). The second group estimated their grade, but also compared this initial estimate to their actual grade (Grade comparison group). The third group also reflected on reasons for a mismatch whenever their estimate differed from their actual grade (Reflection group). We specifically focused on postdictions (cf. Hacker et al., 2000; Nietfeld et al., 2006) instead of predictions (Bol et al., 2005; Foster et al., 2017). When receiving feedback on predictions, students can easily make up excuses for a mismatch between their estimated and actual outcomes. For example, a well-known response by students is that there were too many test items on the exam that did not represent the learning materials (Bol et al., 2005). Consequently, students can more easily blame their miscalibration to the test—it is not their fault their calibration accuracy was off, the test was simply not representative of the learning materials. When compared to predictions, postdictions benefit from privileged knowledge about the test items and the nature of the test. This leaves less room for content-related excuses—students know the test content and should have incorporated any abnormalities in their estimate. Hence, we asked students to postdict their exam performance.

To test the improvement of calibration accuracy over time, we looked at three different measurements of calibration accuracy. First, we examined the absolute difference between estimated and actual performance (i.e., absolute calibration accuracy, Schraw, 2009). Second, to determine the direction of the miscalibration (i.e., overconfidence or underconfidence) we also included bias scores in our study: the non-absolute difference between estimated and actual performance (Schraw, 2009). Finally, we included a third measurement: second-order judgements (SOJs) that indicate how confident students are in their estimate (cf. Dunlosky, 2005; Miller & Geraci, 2011). To attain richer insight in metacognitive awareness among students with high and low exam performance, Miller and Geraci (2011) argued that confidence judgements need to be taken into account as students who show miscalibration can be aware of this by giving their performance estimates low confidence ratings. This was illustrated by their finding that low performing students were poorly calibrated but also gave low confidence scores to their incorrect judgements. Although research on this topic is lacking, SOJs may be relevant for intervention studies as well, as efforts to improve calibration may only be effective if students are indeed aware of their miscalibration. Furthermore, interventions that seemed unsuccessful in changing calibration accuracy may have instead influenced the second-order judgements students made. To gain richer understanding of metacognitive awareness, we therefore asked students to rate the confidence they had in their performance estimates as well (i.e., provide SOJs).

Hypotheses

Based on prior research we formulated several hypotheses. First, we expected to find a relation between the change of calibration accuracy over time and the amount of support given. Based on the limited benefits of sole practice found in the literature (Bol et al., 2005; Foster et al., 2017), calibration accuracy of students in the practice group was expected to remain stable. Second, we expected students in the grade comparison group to show an improvement in their calibration accuracy, but students in the reflection group were expected to show the strongest improvement. Third, similar to our hypotheses for calibration accuracy, we hypothesized that support would also influence bias and SOJs. We expected more support to be related to a stronger reduction in bias, and, following the reasoning that better calibration accuracy represents better metacognitive awareness, we expected more support to also affect SOJs.

In addition to the main hypotheses regarding the improvement and direction of calibration accuracy and SOJs, we measured the possible moderating effect of performance level on calibration accuracy and kept track of students' academic performance. Additionally, we investigated calibration accuracy at the beginning and the end of the year of two other courses: German (i.e., a related language course) and math (i.e., a totally different course). For these courses, students did not follow any calibration accuracy intervention.

Method

Participants

The original initial sample consisted of 261 students (49.8% female, 50.2% male) from 9 classes across grades 1 to 3 (pre-university level) at a Dutch suburban secondary school. These classes all shared the same French teacher. The mean age of these students was 14 years (from 12 to 17 years old). Informed consent was obtained from both students and their parents or caretakers. A letter was sent to their parents or caretakers in which the nature of the study was explained. Thirteen students (5.0%) indicated that they did not want to participate. Another 29 students (11.1%) indicated that they wished to participate in the study, but did not hand in informed consent and hence were excluded. The final sample of students consisted of 219 students (83.9%): 120 girls (54.80%) and 99 boys (45.20%). The mean age of these students was 14 years old (from 12 to 16 years old). The group of students participating in the study could be considered representative of the group that abstained from participation: both gender and age characteristics were similar. Table 1 shows the distribution of the students over grades and classes.

Table 1.
Distribution of students over grades and classes

		N	Boys	Girls
Grade 1				
	Class 1	21	13	8
	Class 2	28	9	19
	Class 3	25	19	6
	Total	74	41	33
Grade 2				
	Class 4	29	15	14
	Class 5	26	11	15
	Class 6	22	11	11
	Total	77	37	40
Grade 3				
	Class 7	23	6	17
	Class 8	24	7	17
	Class 9	19	9	10
	Total	66	22	44

Procedure

Figure 1 depicts the set-up of the current study. The school year was divided into four periods during which students took two French exams. The first exam of each period was administered by the teacher during a regular class meeting. All students within one class took the test at the same time and the teacher aimed to test all students from the same grade within the same week. The second exam of each period was administered during an official testing week in which students received exams for each of their school courses (French, math, German, etcetera). All classes were allocated to different classrooms and took their exam at the exact same time. So, over the year, students took eight French exams. Students in year 1 took seven exams, and only participated in the regular testing weeks from the second period onwards. Hence, in year 1 the intervention was implemented from the second period.

At the end of each exam, students in all intervention groups received a scoring form. First, students had to write down the grade they thought they had obtained for the exam on a scale from 0 to 10 with one decimal (10 represented the highest score possible). Second, students rated the confidence they had in their estimated grade on a five-point scale (1 = very unconfident; 5 = very confident, cf. Miller & Geraci, 2011). After filling out the scoring form, students put the form in an envelope and sealed it to ensure confidentiality of their estimates. Students handed both their exam and their sealed envelopes to their teacher. During the next class approximately one week later, exams with the scoring of each answer were returned to the students (exam grades had already been presented online). This was followed by a plenary discussion of the exam which occurred for all groups as this was a required part of the French lessons.

Within each year, each class was randomly assigned to one of the intervention conditions. The first group of students only practiced with making performance estimates immediately after their French exams (Practice condition). These students were not involved in any additional intervention whatsoever. In addition to making performance estimates, the second group of students had to compute mismatch scores between their estimate and their final performance when their grades were returned the week following their exam (Grade comparison condition). To do so, the initial estimate form was returned to each of the students and students received a new form on which they could calculate the difference between their estimate and actual grade. After completing their comparison, students put the comparison form together with their initial estimate in a new envelope, sealed it, and handed it in to their teacher.

Additional to computing mismatch scores, the final group of students also had to reflect on the cues they had used when estimating their performance. Furthermore, if their estimate mismatched their actual performance by at least 0.5 points they also had to reflect on reasons why (Reflection condition). The threshold of 0.5 was based on a pilot study with students from three different Dutch secondary schools (*n* = 407), in which we found that the 25% best calibrated students had a maximum deviation of less than 0.5 point. Students also indicated whether they would change their study behavior and preparation for the next French exam. Furthermore, students indicated whether they would attend one or more of the extra support hours of the French course during the next period. In these weekly support hours, the teacher provided extra explanation about the course content. Such support hours were mandatory for students who performed very poorly.

In addition to measuring calibration accuracy on the course French, we measured calibration accuracy on the German and Math courses at the beginning (pre-test) and end of the intervention (post-test, see Figure 1). After their exams German and math, students got a scoring form similar to the one used after their French exams: students estimated their grade and rated how confident they were about this estimate.

After the experiment, we provided all students with a final questionnaire, in which we asked them how they had experienced the experiment and whether they felt they had learned something about their calibration accuracy. Students were also provided with the Dutch version (Blom & Severiens, 2008) of the self-efficacy subscale from the Motivated Strategies for Learning Questionnaire (MSLQ, Pintrich & De Groot, 1990). Students in year 1 filled in the self-efficacy questionnaire after their first official exam as well.

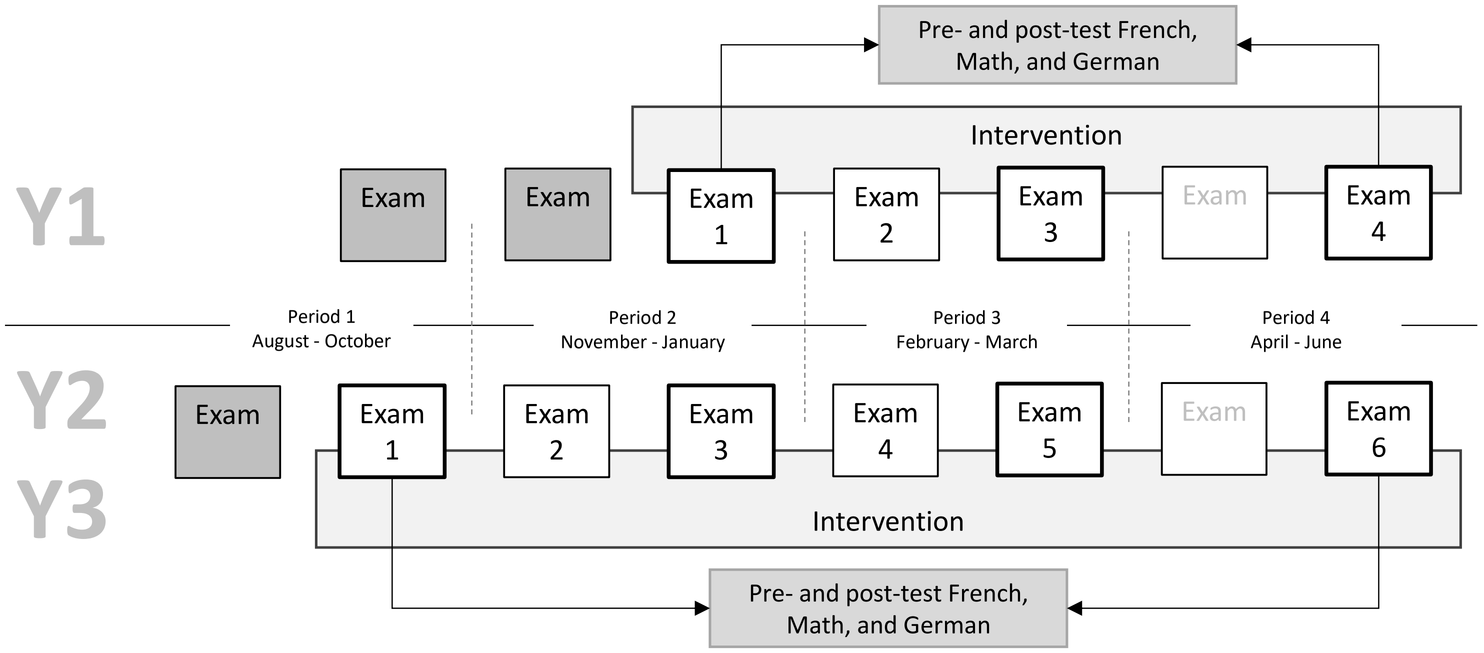


Figure 1. The design over time of the current study.

Materials

Exams

We tested the calibration accuracy on French exams. Each exam consisted of questions on vocabulary, sentence construction, and grammar. Students had to translate single words, had to fill in blanks to make sentences correct (e.g., "Chaque année, il y a un spectacle avec des ______ à la plage"), had to change verbs and sentences in a different tense and finally had to answer different questions in French (e.g., "How can you say that you will arrive around five o'clock?"). Furthermore, there was a part on either listening or reading comprehension. Students had to answer questions about the content they had just heard or read (e.g., "What is said about the History teacher?"). Furthermore, when students got multiple listing or reading vignettes, they had to be able to compare the two vignettes (e.g., "Describe what text 1 and text 2 have in common").

Within each year, students from different classes received the same exams at the same time. Grades were defined by the number of correct answers given by the students. The scoring and norm were decided upon before each exam by the teacher for the French subject based on cohort results of prior years and guidelines laid down in the method of education. The scoring norms were not adjusted during the school year and was the same for all students within the same year.

In each year, one exam was not taken into account in our analyses. This was the first exam in period 4 (exam 5 for year 1, and exam 7 for year 2 and 3). This exam did not contain a listening or reading test and was therefore not comparable to the other exams.

Intervention forms

Practice and comparison. Students could provide their estimated grade and SOJ on the practice form, and calculate differences between their estimate and actual grade on the comparison form.

Reflection. The Reflection form started similar to the Grade comparison form: students calculated the difference between their estimated and actual performance. After doing so, students had to reflect on the cues used when estimating their grades. To encourage reflection, earlier studies provided students with statements to choose from (Bol et al., 2005, Nietfeld et al., 2005, 2006). This method was favored above open reflection questions, because open questions could more easily elicit "don't know" responses (Krosnick & Pesser, 2009). At the same time, however, closed statements could evoke responses students would otherwise not have provided. We therefore decided to use a combination of open and closed questions. First, to gain more insight in the cues used by the students when estimating their performance, the first question was: ‘How did you come up with the estimate of your grade of your exam?’ All students could then choose one or more of eight statements of possible reasons for miscalibration that we found in the literature (e.g., Bol 2005; Bol & Hacker 2001). The statements were: (1) The grades I normally get for this school subject; (2) How well I prepared myself for the exam; (3) How many questions I knew the answer to; (4) How much time was left when I finished the exam; (5) How secure/insecure I was during the exam; (6) How relaxed/stressed I was during the exam; (7) How difficult I felt that the questions were; (8) The norm used by the teacher when grading the exams; (9) Different, namely … (please fill in below your answer). Second, students had to reflect on the mismatch between their actual and estimated performance by answering the following question: How do you explain the mismatch between your estimated grade and actual obtained grade? Finally, students were asked whether they would change their study behavior and preparation for the next French exam and whether they would attend the extra support hours for the French course at least once in the next semester.

MSLQ

At the end of the intervention, we administered the self-efficacy subscale from the MSLQ Dutch version (Blom & Severiens, 2008; Pintrich & De Groot, 1990). The self-efficacy questionnaire consisted of eight items (e.g., I'm confident I can do an excellent job on the assignments and tests in this course). Year one also filled out the self-efficacy questionnaire at the beginning of the intervention. We administered the self-efficacy questionnaire to attain more insight in whether a change in calibration accuracy would impact students’ self-efficacy and whether the two constructs would be related. Discussing this is beyond the scope of the current article, however.

Analyses

Outcome measures

To gain insight in the ability of students to estimate their own performance, calibration accuracy was computed as the absolute difference between students’ actual grade and their estimated grade, both on a scale from 0 to 10. Furthermore, to examine whether students showed overconfidence or underconfidence, bias scores were calculated as the signed difference between estimated grades and actual grades (Dunlosky & Thiede, 2013; Schraw, 2009). Bias score differences were relative and could thus range from -10 to +10.

Analyses

Our data could potentially show a nested structure of (1) different measurement points for each student and (2) students in classes. To test whether multilevel analysis was indeed necessary due to the nesting in the data, we first ran intercept-only models using multilevel analysis with the HLM7 software. Doing so allowed us to determine whether the outcome variables had a significant amount of variance on the second (student) and third (class) level (Hox, 2010). We tested this for our outcome variables calibration accuracy, bias scores, SOJs, and exam performance. Results showed for calibration accuracy $\chi^2(207) = 419.06$, $p < .001$; for exam performance $\chi^2(207) = 1924.27$, $p < .001$; and for SOJs $\chi^2(207) = 739.87$, $p < .001$, a nested structure on the Student Level. No Class Level effects were found for calibration accuracy $\chi^2(8) = 12.71$, $p = .122$; exam performance $\chi^2(8) = 14.27$, $p = .074$; and SOJs $\chi^2(8) = 12.56$, $p = .127$. Bias scores showed, besides a Student Level effect $\chi^2(207) = 688.58$, $p < .001$, also a significant effect on Class Level $\chi^2(8) = 23.88$, $p = .003$. This indicates that students in different classrooms differed from more from each other than students from the same class. Hence, the tests indicated that we should use a two-level multilevel design for calibration accuracy, exam performance and SOJs, and a three-level multilevel design for bias scores. To test significance, we used the 5% level. Furthermore, missing data (for example when students were absent because of illness or forgot to estimate their grade) were treated as missing at random (Hox, 2010).

Results

Before testing the hypotheses regarding calibration accuracy, bias, SOJs and performance, we first looked at the pre-test descriptives (means, medians, and standard deviations) of the intervention groups (see Table 2), and examined possible pre-test differences with HLM. Calibration accuracy of the students on the pre-test had a mean of 0.97 (median = 0.80, *sd* = 0.68). This means that student estimates were on average about one point off their actual grade. An analysis with calibration accuracy on the pre-test as dependent variable and intervention group as independent variable (Practice vs. Grade comparison vs. Reflection) showed no significant differences between the intervention groups on the pre-test on calibration accuracy, *t*'s < 2. Hence, calibration accuracy did not differ between intervention groups at the start of our intervention.

When looking at the bias scores on the pre-test, results showed that, in contrast to previous research (e.g., Kruger & Dunning, 1999), students did not show a specific tendency to be overconfident. Instead, the mean bias over all groups was -0.20 (median = -0.30, *sd* = 1.16), which indicates slight underconfidence. An analysis with bias scores on the pre-test as dependent variable and intervention group as independent variable showed no significant differences between groups *t*'s < 2. This indicates that groups, again, did not differ on the pre-test in terms of their underconfidence or overconfidence.

Second-order judgements (SOJs) on the pre-test were examined as well. The mean score indicated that students reported average confidence in most of their judgements (*M* = 3.18, median = 3.00, *sd* = 0.81). An analysis with SOJs on the pre-test as dependent variable and intervention group as independent variable showed a significant difference between the reflection group and the other two groups *b* = -0.35, *t*(193) = -2.47, *p* = .014. The reflection group (*m* = 2.99, *sd* = 0.81) was less confident than the practice group (*m* = 3.32, *sd* = 0.83) and grade comparison group (*m* = 3.27, *sd* = 0.74. There were no differences between the practice and comparison group, *b* = -0.06, *t*(193) = -0.41, *p* = .682. This indicates that only students in the reflection group generally provided lower confidence scores to their performance estimates.

Table 2.
Means and standard deviations among different intervention groups

	Practice		**Grade comparison**		**Reflection**	
	M [min, max]	SD	M [min, max]	SD	M [min, max]	SD
French						
Performance	6.82 [3.50, 9.60]	1.34	6.62 [3.00, 9.30]	1.38	6.49 [3.00, 10.00]	1.38
Bias	-0.31 [-4.30, 2.90]	1.17	-0.18 [-5.50, 3.20]	1.14	-0.23 [-3.10, 3.00]	1.19
Calibration	0.96 [0.00, 4.30]	0.74	0.90 [0.00, 5.50]	0.73	0.97 [0.00, 3.10]	0.72
SOJ	3.40 [1.00, 5.00]	0.75	3.29 [1.00, 5.00]	0.83	3.11 [1.00, 5.00]	0.85
Math						
Performance	7.00 [3.00, 10.00]	1.59	6.81 [3.00, 10.00]	1.66	6.78 [3.00, 10.00]	1.65
Bias	-0.41 [-4.00, 3.70]	1.49	-0.12 [-5.10, 5.30]	1.41	-0.30 [-4.20, 4.00]	1.64
Calibration	1.24 [0.00, 4.00]	0.92	1.12 [0.00, 5.30]	0.87	1.33 [0.00, 4.20]	1.00
SOJ	3.37 [2.00, 5.00]	0.6	3.31 [1.00, 5.00]	0.97	3.14 [1.00, 5.00]	0.91
German						
Performance	6.38 [3.00, 9.50]	1.43	6.05 [3.00, 9.40]	1.32	6.17 [3.10, 9.20]	1.32
Bias	0.41 [-2.20, 3.10]	1.21	0.47 [-2.90, 3.00]	1.18	0.52 [-3.25, 4.10]	1.36
Calibration	1.05 [0.00, 3.10]	0.73	1.02 [0.00, 3.00]	0.76	1.15 [0.00, 4.10]	0.89
SOJ	3.28 [1.00, 5.00]	0.78	3.17 [1.00, 5.00]	0.97	3.17 [1.00, 4.00]	0.67

Finally, exam performance (i.e., grades) was investigated. Findings showed a mean of 6.68 (median = 6.60, *sd* = 1.36). An analysis with performance on the pre-test as dependent variable and intervention group as independent variable showed no significant differences in exam performance *t*'s < 2. Hence, intervention groups did not differ in their grades on the pre-test.

In sum, these results on the pre-test showed that intervention groups did not differ from one another on the calibration accuracy, bias, and performance. The only difference was found on SOJs: students in the reflection group were less confident in their estimates than students in the bias and practice group. To answer our research questions and hypothesis, in the next section we examined the change in calibration accuracy over time.

Improvement of absolute calibration accuracy over time

Our first hypothesis was that the level of support in using the outcome feedback was related to the improvement of calibration accuracy over time. Table 2 displays the mean in calibration accuracy over the school year. To examine calibration accuracy over time and between our intervention groups, we ran our two-level model in HLM7 as explained in the method section. Table 3 presents the model with all variables and interactions included. Results showed a significant improvement of calibration accuracy over time *b* = -0.03, *t*(215) = -2.95, *p* = .004. Note that smaller scores indicate better calibration accuracy. Hence, the negative regression coefficient shows that, over time, calibration accuracy became better. We also examined whether the effect of time was random, and thus differed between students or groups. Results showed a non-significant random effect of time: $\chi^2(213)$ = 217.16, *p* = .408. This means that the improvement was the same for students from all intervention groups. Consequently, when adding the intervention groups to the model, there were no significant interaction effects between any of the groups and time, all *t*'s < 2 (see Table 3). This means that our first hypothesis—the amount of support would influence the improvement of calibration accuracy over time—was not supported.

Change in performance estimates

The tests described above showed that calibration accuracy improved over time for all students. Note that calibration accuracy can be influenced by a change in performance level (i.e., students' score becomes more in line with their estimate), or by a change in performance estimates. To check whether the enhanced calibration was caused by actual adjustments of performance estimates, we conducted a follow-up analysis to examine the change in performance estimates. We used the same multilevel analysis with the HLM7 software as with the calibration accuracy scores. However, instead of including calibration accuracy as the outcome variable, we included 'performance estimates' as the outcome.

Results showed that there was a significant variation in performance estimates over time *b* = .05, *t*(215)=2.89, *p* = .004, indicating that students' estimates indeed changed over time. Hence, the enhanced calibration accuracy over time was accompanied by a change in performance estimates.

Performance level differences

We also examined the effect of performance level on calibration accuracy and the possible improvement over time because prior studies have showed that performance level could moderate this improvement (e.g., Hacker et al., 2000; Nietfeld et al., 2006). As already indicated by the non-significant random effect of calibration accuracy over time, performance level did not influence the improvement of calibration over time *b* = 0.01, *t*(212) = 0.89, *p* = .377—all students improved similarly. Furthermore, there was no main effect of performance level on calibration accuracy *b* = -0.07, *t*(212) = -1.38, *p* = .170. This result means that high and low performers did not differ from each other in terms of their miscalibration.

Improvement from pre-test to post-test on German and Math

Finally, we tested the improvement in calibration accuracy on courses that were not included in our intervention, i.e., German (related course) and math (unrelated course). We examined differences between pre-test and post-test calibration accuracy, as we only measured calibration before and after our intervention. We used a paired-samples *t*-test. Results showed that calibration accuracy on the German and Math courses did not significantly differ from pre-test to post-test, *t*'s < 2. This indicates that students' calibration accuracy both on German and Math did not change from pre-test (respectively *M* = 1.15, *SD* = 0.77; *M* = 1.25, *SD* = 0.96) to post-test (respectively *M* = 1.03, *SD* = 0.80; *M* = 1.19, *SD* = 0.89).

Improvement of bias scores over time

To test the change in bias scores over time, a multilevel analysis with HLM7 was planned, using a three-level model as explained in the method section. The means and standard deviations for bias scores over the school year are reported in Table 2. We expected that students' bias would decrease when students were supported in reflecting on the outcome feedback they received. However, results showed no change in bias scores over time, *b* = -0.03, *t*(8) = -0.60, *p* = .567, nor was the effect of bias scores over time random $\chi^2(205) = 232.17$, *p* = .094. This means that there were no significant differences between groups of students and how they improved their bias scores. It did not matter whether students received limited or substantial support in reflecting on the mismatch between estimated and actual performance (see also Table 3).

Performance level differences

As with calibration accuracy, we also examined the role of performance level on bias scores and its improvement over time. As already indicated by the non-significant random effect, performance level did not significantly influence how bias scores improved over time b = -0.01, t(200) = -0.23, p = .821. When examining the main effect, results showed that performance level significantly influenced bias scores b = -0.34, t(200) = -4.43, p < .001. High performance was associated with more underconfidence as indicated by the negative regression coefficients.

Improvement from pre-test to post-test on German and math

We tested the change in bias scores on courses that were not involved in our intervention: German (related) and math (unrelated). By using a paired-samples t-test, we examined differences between pre-test and post-test bias, as we only measured bias before and after our intervention. Results showed no change in bias for math from pre-test (M = -0.35, SD = 1.53) to post-test (M = -0.24, SD = 1.47) t = -.82, p = .414. The only significant change was among bias score for German t = 2.83, p = .005. Students were less overconfident on the post-test (M = 0.32, SD = 1.26) than on the pre-test (M = 0.69, SD = 1.20).

Improvement of second-order judgements over time

After investigating calibration accuracy as the absolute and relative difference between estimated and actual performance, we looked at SOJs as well (see for descriptives over the school year Table 2). Following the reasoning that students who calibrate better also have more confidence in their performance estimates, the improvement in calibration accuracy as found in our previous analysis should be accompanied by an increase in confidence over time.

Results showed no overall change in confidence judgements over time, b < -0.01, t(251) = -0.22, p = .829. The effect of confidence judgements over time was random, however, χ^2(213) = 254.04, p = .028, indicating that there were differences between (groups of) students and how they changed their confidence judgements over time. To examine whether this difference was caused by the intervention group that students were in, we examined the interaction between the intervention groups and time. Results showed that there were no significant differences between the Grade comparison group and the Practice (reference) group, $t < 2$, see also Table 3. Results showed that students in the Reflection group differed from students in the Practice and Grade comparison groups, $b = 0.06$, $t(212) = 2.12$, $p = .035$. Students in the Reflection group became more confident during the school year, whereas this change was not found for the other two groups.

Performance level differences

Again, we examined the role of performance level. We tested whether students of different performance levels changed their confidence scores differently over time. Results showed that performance level did not significantly influence how confidence scores changed over time $b = -0.01$, $t(212) = -0.33$, $p = .740$. Furthermore, when examining the main effect, results showed that performance level did not significantly influence the confidence judgements $b = 0.01$, $t(212) = 0.08$, $p = .898$. This means that students from different performance levels did not differ in how confident they were in their performance estimates.

Improvement from pre-test to post-test on German and math

Again, we tested the improvement in calibration accuracy for German (related course) and math (unrelated course) as well. By using a paired-samples *t*-test, we examined differences between pre-test and post-test calibration accuracy, as we only measured calibration before and after our intervention. Findings indicate no change in confidence judgements for math and German from pre-test (respectively $M = 3.31$, $SD = 0.92$; $M = 3.27$, $SD = 0.78$) to post-test (respectively $M = 3.24$, $SD = 0.87$; $M = 3.12$, $SD = 0.87$), $F < 1$.

Improvement of exam performance (grades) over time

Finally, we examined exam performance during the school year. To test a change in exam performance scores over time, we conducted a multilevel analysis with HLM7 using a two-level model as explained in the method section. The mean and standard deviations for exam performance are reported in Table 2. We tested whether performance improved over time and differed between different intervention groups. Our results showed that performance changed over exams $b = 0.05$, $t(215) = 3.19$, $p = .002$. During the school year, students obtained better exam grades. To test whether there were group differences in the improvement, we examined whether the effect of exam performance was random. Results showed that the effect of time on performance level was random $\chi^2(213) = 259.17$, $p = .017$ and thus differed between (groups of) students.

We investigated whether this difference could be explained by the intervention group that students were in by adding the intervention groups to the model. Results showed that differences in improvement over time were not caused by the intervention groups, *t*'s < 2 (see also Table 3). Furthermore, there were no main effects of our intervention groups, *t*'s < 2.

Improvement from pre-test to post-test on German and Math

Our final analysis consisted of a test on the change in math and German exam performance from pre-test to post-test. We used a paired-samples *t*-test. Results showed no significant change in performance for math from pre-test (*M* = 6.88, *SD* = 1.46) to post-test (*M* = 6.96, *SD* = 1.71) *t* = -.60, *p* = .549. There was also no significant increase in German performance from pre-test (*M* = 6.12, *SD* = 1.24) to post-test (*M* = 6.09, *SD* = 1.40), *t* = .23, *p* = .820.

Table 3.
Outcome variables calibration accuracy, bias, and performance

	Calibration accuracy		Bias		SOJ		Performance	
	b	*p*	*b*	*p*	*b*	*p*	*b*	*p*
Fixed part								
Level 1 (Time)								
Intercept	1.05	< .001	-0.16	.457	3.28	<.001	6.48	<.001
Time	-0.03	.004	-0.03	.567	-.001	.829	0.05	.002
Level 2 (Student)								
Practice (reference group)	1.43	< .001	2.13	.005	3.47	<.001	6.72	<.001
Grade comparison	0.12	.292	0.01	.993	-0.16	.194	-0.25	.281
Reflection	0.10	.410	-0.07	.747	-0.50	<.001	-0.44	.069
Performance level	-0.07	.170	-0.34	< .001	0.01	.898	-	-
Interactions with Time								
Practice (reference group)	-0.07	.359	0.01	.945	-0.01	.965	0.03	.144
Grade comparison	-0.05	.054	0.02	.666	0.02	.596	<0.01	.914
Reflection	-0.03	.365	0.01	.782	0.06	.035	0.03	.366
Performance level	0.01	.377	-0.01	.821	-0.01	.740	-	-

Discussion

The central aim of this article was to examine if, and how, outcome feedback can be used to improve students' calibration accuracy in secondary school. Although the potential of outcome feedback to improve calibration accuracy was widely discussed (Brown et al., 2015; Foster et al., 2017; Hacker et al., 2000; Miller & Geraci, 2011; Nietfeld et al., 2006), empirical research fell short as to how outcome feedback could effectively be used to improve calibration accuracy. In the current study, it was investigated whether increasing the support to reflect on the difference between estimated and actual outcomes would help students in secondary school to become better calibrated after receiving outcome feedback.

In line with studies showing that outcome feedback could help making students more aware of their performance (Callender et al., 2015; Hacker et al., 2000; Miller & Geraci, 2011; Nietfeld et al., 2006), results showed an improvement in calibration accuracy for each intervention group. This may be explained by an improvement of the cues used by the students (Koriat, 1997)—they expectedly shift from using less valid to more valid cues. Interestingly, supporting students in their reflection after receiving outcome feedback did not further add to this improvement. This means that even when secondary school students simply provide performance estimates, they can already improve their calibration accuracy. On the German and math courses were no intervention was implemented, no improvement in calibration accuracy was found. Furthermore, while prior research showed that performance level could influence whether students improved from getting outcome feedback, the current study did not find any differences between students. This indicates that with a very small intervention, calibration accuracy in secondary school can already effectively be improved without disadvantaging low performing students as was found by Hacker et al. (2008).

The current study makes several major contributions to the literature on calibration and self-assessment. First, this study underlines that students can improve their calibration accuracy with even a small feedback intervention. Although its potential has been discussed, much remained unclear about if and how outcome feedback could help reduce miscalibration in secondary school. Second, by including different outcome measures, we gained more insight in not only calibration accuracy, but in bias scores and SOJs as well. The current study indicates that interventions can also impact SOJs. This has implications for research on metacognitive awareness and underscores the importance to include different metacognitive outcomes to fully understand the impact of an intervention. Third, this study is conducted in a secondary school context. This is an important addition to the literature on improving calibration accuracy in classroom settings, given the predominant focus on university contexts until day.

In the following sections, the results will be discussed in relation to the existing literature. Furthermore, we will elaborate on the limitations and directions for future research.

Support to improve calibration accuracy

Our results indicate, in contrast to findings by Bol et al. (2005) and Foster et al. (2017), that calibration accuracy can be improved over time, even when students only estimate their performance. A first possibility for these conflicting findings is that our study focused on postdictions whereas Bol et al. (2005), and Foster et al. (2017), focused on predictions. As mentioned in the Introduction, students may be more likely to come up with excuses when receiving feedback on predictions. For example, students may state that the test did not adequately represent the learning materials. When providing a postdiction, however, students have more knowledge about the test, leaving less room for content-related excuses. Hence, it is possible that improvement in calibration accuracy may be harder to constitute with predictions when students, besides estimating their performance, only receive outcome feedback. For example, a study by Authors (submitted) showed that students who received feedback on their performance on a text learning task improved their postdictions over tasks, but worsened their predictions. Hence, to gain more insight in this topic, we encourage future research to further examine whether the type of performance estimates could indeed impact the effect of outcome feedback.

Another possible reason for the difference in Bol et al. (2005) and Foster et al. (2017) and the current study could lie in the setting. In university, students receive outcome feedback but often do not discuss the exam in detail unless students specifically ask for it. In Dutch secondary school, exams are typically discussed the next class. This means that students get feedback about their mistakes and what the correct answers should have been. It is plausible that this more detailed feedback may have helped students to become better calibrated, and that calculating comparison scores and reflecting on causes for miscalibration did not further add to this improvement. In other words, perhaps students became better calibrated during the year because of the detailed feedback they received after each exam. To gain more insight in this possibility, we tested calibration accuracy differences between our third-year and first- and second-year students on the first exams they took. Third-year students have had many detailed feedback moments the last two years, and if this feedback would help students to become better calibrated, we would expect the third-year students to be better calibrated than the second, and especially the first-year students, who both received much less detailed feedback. Results did not support this assumption, however. Results did not show significant differences between third-year students' calibration accuracy when compared to first-year and second-year students, *t*'s < 2. This means that all students were calibrated equally well on their first exam, refuting the hypothesis that the extended feedback in itself would have enhanced the calibration accuracy. Another way to test whether simply discussing the exams could lead to better calibration accuracy was by looking at improvements in German and math. Neither course showed improved calibration accuracy, although the content of each exam was discussed in detail afterwards. Hence, we deem it unlikely that the reason for our effects would lie solely in the secondary school setting.

A possible reason why we did not find any differences between our intervention groups could also be that students reflected on their mismatches by default, regardless of whether they were prompted to do so. At the same time, it is possible that students in the reflection group did not reflect enough, or perhaps only repeated whether they were overconfident or underconfident. To examine this, we investigated how often students reflected, and the nature of their reflection by scoring them in three categories: no reflection (e.g., "no idea"), a repetition of the situation (e.g., "my estimate was lower/higher than my actual grade") and an actual reflection (e.g., "I based my estimate on the scores I normally get for French, but this time, my grade was higher"). Over the whole year, 70.8% of the students in the reflection group were required to provide a reflection. In a little over half of the cases, students indeed provided a reflection (56.9%) compared to 12.4% who indicated they did not know how their miscalibration was caused, 23.4% who simply repeated the situation, and 7.3% who just did not reflect at all. This means that of all students who had to reflect, only half provided a meaningful reflection which may have been too little to induce a reflection group-effect. Furthermore, it seemed that motivational processes started to play a role. Some students in the reflection group reported that they started to provide lower estimates of their performance because they found it rewarding when their actual grade appeared to be higher than their initial estimate. This suggests that, although we do not see more improvement in calibration accuracy compared to the other groups, some students did become more able to control the estimates of their performance.

Although students in the Reflection group, compared to the other intervention groups, did not significantly differ in how they improved their calibration accuracy over the year, the reflection group was the only group that changed their second-order judgements. Over the school year, students in the reflection group became more confident in their performance estimates. Second-order judgements get increased attention as a different measure of metacognitive knowledge (Händel & Fritzsche, 2016; Miller & Geraci, 2011). Recent research by Authors (submitted) in which the pre-test calibration accuracy data of students in the current study were compared to calibration accuracy of university students, showed that students in secondary school are generally unable to correctly adjust confidence judgements to their performance estimates. The finding that reflecting on the mismatch between estimated and actual performance can also impact second-order judgements is therefore a promising finding and suggests that these younger students can improve their metacognitive awareness through an intervention. Note that although the reflection group does become more confident in their performance estimates, an HLM analysis with SOJ on the post-test as dependent variable and intervention group as independent variable shows that their mean SOJs do not differ from the other two intervention groups on the post-test, *t*'s < 1.

Limitations and directions for future research

In this study, we conducted a longitudinal intervention study in an ecologically valid setting. Our results give important insight in how outcome feedback can be used to improve calibration accuracy in secondary school. At the same time, however, this study has some limitations. First, as previously mentioned in the discussion, it is possible that all students reflected on their performance, regardless of whether they were supported to do so. From our data, we cannot make inferences about this, and we encourage future research to collect more information about if, and how, this reflection happens. At the same time, we see that students sometimes had a hard time providing a meaningful reflection. Perhaps students would have benefitted from more guidance as to how they should have reflected, for example in a practice session where good and poor reflections are discussed. Future research could further examine whether such a reflection training would help.

Additionally, all intervention groups had the same teacher. This was beneficial because teacher effects were excluded, but we cannot rule out that the teacher did perhaps unintendedly encourage students to provide better estimates of their own performance which may have influenced the results and improved calibration accuracy over time. Roelle, Schmidt, Buchau, and Berthold (2017) showed, for example, that warning students for overconfidence could already prevent these students to give overconfident judgements. Evidence on how teachers influence students' calibration accuracy is scarce and an important avenue for future research. This could perhaps also explain the differences found between the different courses. Although French and German were similar in their exam format and learning assignments, the pattern of calibration accuracy and bias was different. This signals an important avenue for future research to distinguish between different courses and the teacher element in this.

This study was conducted in a specific secondary school context in which the teacher for each course remains the same during the year and knowledge builds upon each other over exams. This contrasts with a university setting, in which lecturers change from one course to another. While we believe our findings to hold in comparable contexts and also in university, more empirical research is needed to test generalizability and further develop the boundaries of when outcome feedback can effectively improve calibration accuracy. To this end, future research could examine calibration accuracy of university students following the same intervention as we did. Moreover, future research in a secondary school setting may further examine the difference between students of different school levels. When finishing primary school, students in the Netherlands are divided over different school levels based on their GPA. The current study included students that were involved in pre-university education which is the highest level. To strengthen the generalizability of our findings, future research could examine the effectiveness of the feedback intervention in lower levels as well, such as vocational education.

A final recommendation for future research entails the cues students use when estimating their performance. In the current study, we only got specific information of the cues used from students in our reflection group. As a first exploratory analysis, we therefore calculated how often each cue was reported during each test, and analyzed whether there was a relation between how often the cue was reported over time. Results showed a significant change in the cue “How well I prepared myself for the exam”. During the year, this cue was reported more often when students reflected on their performance estimate, as shown by the positive regression coefficient $b = 0.05$, $t(234) = 3.03$, $p = .003$. This means that the current study provides reasons to believe students can adjust the cues they use when estimating their performance, but future research is needed to further examine this shift in cue when implementing an intervention to improve calibration accuracy. To this end, future research may also benefit from qualitative methods such as asking students to think aloud when estimating their performance. This can help to attain an even more detailed insight in the (change in) cue use during and after calibration interventions.

Conclusion

Our study shows that estimating own performance and receiving outcome feedback can indeed be beneficial for students. Not only did students improve their calibration accuracy, even when they only estimated their performance, they also improved their exam performance. This indicates that encouraging students to explicitly think about their performance for an extended period of time can help them to improve their calibration and their performance accuracy. Such practice can easily be implemented in school settings and is therefore encouraged.

Chapter 7

Summary and General discussion

Many individuals find it hard to accurately estimate their own performance, leading to inflated performance estimates (Dunlosky & Lipko, 2007; Sheldon, Dunning, & Ames, 2014). Miscalibration can have serious consequences as it negatively influences decisions individuals make following their performance (Dunlosky & Rawson, 2012; Nelson & Narens, 1990). For example, physicians who think they have accurately diagnosed a patient but actually have failed to do so will be unlikely to take action on optimizing the diagnosis. Similarly, students who prematurely think they have mastered the material will stop studying. Therefore, to expand our limited understanding of reducing miscalibration could significantly improve our awareness of our performance and lead to better decision-making in subsequent tasks.

In this dissertation, a series of studies is presented that examine how calibration accuracy can be improved. A central role is given to performance feedback, as it gives insight into prior performance, which can help to improve calibration accuracy (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Lipko et al., 2009; Rawson & Dunlosky, 2007). How this performance feedback affects calibration accuracy on new tasks, however, is still unclear. Hence, the first question of this dissertation was whether providing students with performance feedback could help them to improve their calibration accuracy on new, subsequent tasks.

Another key question concerned differences between performance levels. Previous research has shown that students with low task performance are often poorly calibrated (i.e., overconfident), whereas students with high task performance show more accurate calibration and slight underconfidence (e.g., Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999). These differences in calibration accuracy could potentially affect the use of feedback (Gutierrez de Blume, Wells, Davis, & Parker, 2017b; Stone, 2000; Thiede, Griffin, Wiley, & Anderson, 2010), but the exact relation with feedback remains unclear. The studies in this dissertation therefore included performance level (either as task performance differences or broader experience level differences) to examine its influence on calibration accuracy.

Summary of the main findings

Chapter 2 describes an experimental study aimed at disentangling how performance feedback can increase calibration accuracy on new tasks where such feedback is not immediately present. In this study, students were requested to read texts and learn definitions that they later had to recall during a test. After each recall attempt, students estimated whether their recalled definition was correct, incorrect, or partially correct. Two experimental groups were created. The first group consisted of students who received full definition standards (i.e., the correct definition) while estimating their performance. Hence, students in this group could directly compare their own recall attempt to the actual correct definition. The other half of the students did not receive such standards while estimating their performance. Calibration accuracy was examined on new texts where none of the students had access to standards.

Comparing their own answers to standards provided students with a cue about both their recall performance (did their recall match the required answer?) and calibration accuracy (did their performance estimate match the actual outcome?). More valid insight into own performance and calibration accuracy on prior texts was expected to help students improve their calibration accuracy on a new text. In support of this hypothesis, results showed that students receiving standards showed better calibration accuracy and less overconfidence on subsequent texts where they had no access to standards. Although students with high recall performance calibrated better than students with low recall performance in general, there were no differences in the way they changed their calibration accuracy after having received feedback. Both groups improved their calibration on the subsequent tasks.

That calibration accuracy could indeed be improved over tasks by a simple feedback intervention among students with both low and high recall performance is very promising as this feedback can be easily implemented in educational practice. However, the results of Chapter 2 show that students still have room for improvement in their calibration accuracy. Even after students compared their recall response to the standard, they showed miscalibration; therefore, perhaps, students need more guidance as to how they should compare their recall to the standards. This guidance could help them to obtain better cues of their prior performance, and thus, prior calibration accuracy. With the aim of investigating whether guidance on how to use the performance feedback was indeed effective for calibration accuracy, **Chapter 3** describes a follow-up study using the same materials and design as Chapter 2. Instead of providing full definition standards, as was the case in Chapter 2, students were divided in groups that either received full definition standards or idea-unit standards. In these idea-unit standards, the correct answer was parsed into parts whereby each part had to be present to receive full credit.

Students receiving idea-unit standards were expected to show better calibration accuracy on new texts than students receiving full-definition standards. In support of this hypothesis, results showed that students receiving idea-unit standards calibrated better and were better calibrated (i.e., less overconfident) than students receiving full definition standards. Importantly, providing standards led to a significant improvement in calibration accuracy over subsequent texts. The more standards students received, the better their calibration accuracy became on new texts where standards were absent. The type of standard interacted with this improvement: receiving idea-unit standards led to more improvement. Finally, performance level also affected calibration accuracy over texts. While students with high recall performance calibrated better than students with low recall performance, results showed that especially low performers' calibration accuracy benefitted from the standards received. This group showed the strongest improvement in their calibration over texts.

In Chapters 2 and 3, performance level differences were operationalized as task performance differences. Based on their recall performance, students were divided into high, average, and low performers, and differences in calibration accuracy between these groups were examined (cf. Ehrlinger et al., 2008; Hacker, Bol, Horgan, & Rakow, 2000; Kruger & Dunning, 1999; Nietfeld, Cao, & Osborne, 2006). This commonly used method gives insight into differences between students, but generalizing to groups that differ in terms of their broader experience remains unclear. Does experience or domain expertise influence calibration accuracy? And would calibration accuracy in different experience groups be enhanced after a performance feedback intervention? With the aim of answering these questions, **Chapter 4** describes a study in which broader experience level differences were investigated: a comparison was made between second-year medical students and board-certified medical specialists with many years of clinical experience.

In the study described in Chapter 4, all participants diagnosed three different medical cases. After diagnosing each case, medical students and specialists rated the extent to which they thought the accurate diagnosis had been provided. Subsequently, half of the students and specialists received performance feedback (the correct diagnosis) before continuing with a new medical case about a different disease. The other half of the participants did not see the correct diagnosis. After diagnosing the first three cases, participants diagnosed three new cases about different diseases. It was tested whether (1) receiving feedback on prior cases could lead to better calibration accuracy on the new cases, and (2) whether specialists calibrated better than the students. The results of Chapter 4 showed that specialists indeed calibrated better than students. Furthermore, providing specialists and students with feedback led to better calibration accuracy for both groups on subsequent cases where such feedback was absent. There were no significant differences between participants' experience level and how their calibration accuracy improved after receiving feedback.

Chapters 2 to 4 support the hypothesis that providing performance feedback could enhance calibration accuracy on subsequent tasks. Furthermore, low or inexperienced performers' calibration accuracy improves equally (Chapters 2 and 4) or more (Chapter 3) from the feedback received. In these chapters, functional confidence was measured—how well individuals are able to estimate their own performance. However, according to Miller and Geraci (2011), there are differences between high and low performing students' subjective confidence as well. Subjective confidence indicates how confident individuals are in these performance estimates and is measured by second-order judgements (SOJs) asking how confident are you that your just given performance estimate is indeed correct? Miller and Geraci (2011) argue that although students with low performance generally show low calibration accuracy, they also provide low SOJs. By doing so, it seems that they show some awareness of their poor calibration accuracy. To gain more insight into the role of performance and experience levels on calibration accuracy, SOJs were included as an additional outcome variable in Chapters 5 and 6.

Chapter 5 describes an observational study in which calibration accuracy, bias, and subjective confidence were investigated among university students and secondary school students. Students took an exam, then rated the exam grade they thought they had obtained, and also how confident they were in their estimate. The study investigated whether students who had poor calibration accuracy were aware of this miscalibration by providing low SOJs (c.f., Miller & Geraci, 2011b). Performance level differences were examined as well, both operationalized as task performance differences (high versus low performing students) and more general experience differences (university versus secondary school).

The results of Chapter 5 showed that, in contrast to earlier findings (e.g., Foster, Was, Dunlosky, & Isaacson, 2017; Kruger & Dunning, 1999), students were largely underconfident of their own performance, both in university and secondary education. At the same time, university students showed awareness of their poor calibration accuracy by providing little confidence to miscalibrated estimates. In contrast to Fritzsche et al. (2018), who suggested that SOJs were dependent on students' exam performance, Chapter 5 found that SOJs depended on calibration accuracy instead. Students with low calibration accuracy assigned little confidence to their performance estimates, regardless of whether their exam performance was high or low. However, this alignment between SOJs and calibration accuracy failed to occur among secondary school students: metacognitive awareness of the secondary school students seemed less developed than that of the university students.

With the aim of testing an intervention that targets calibration accuracy and SOJs, the final chapter of this dissertation, **Chapter 6**, described a longitudinal intervention study in a secondary classroom setting. After each exam, students were requested to estimate their exam grade and the confidence they had in this estimate. Students were divided into three groups in which the support to reflect on the match between estimated and actual outcomes was varied. The first group solely estimated their grade, the second group estimated their grade and also had to calculate the difference between their estimated and actual grade, and the third group also reflected on causes for this difference. After their exams, students received feedback in the form of their grades. During an entire school year, students followed one of the three interventions, and their calibration accuracy, bias, SOJs and, and exam performance were examined.

Support was expected to be related to increasing calibration accuracy over time. However, the results of Chapter 6 showed that any one of the intervention led to improved calibration accuracy, and even an increase of exam performance over time. Simply estimating exam grades already helped students to become better calibrated over the year. Besides improving their calibration accuracy, students in the reflection group also changed their SOJs. During the school year, they became more confident in their performance estimates. Refuting the hypothesis that performance level could influence how students change their calibration accuracy after performance feedback is given, the results showed no differences in the improvement of calibration accuracy between students of different performance level groups. In line with the study presented in Chapters 2 and 4, students with low and high exam grades both improved their calibration accuracy over exams after estimating their grade and receiving feedback. Such an improvement was not found in subjects where no intervention had taken place.

General discussion

The findings of the studies in this dissertation show that calibration accuracy can be improved by providing performance feedback. Students learn from this feedback and show improved calibration accuracy on current and subsequent tasks. Importantly, this effect is shown in a variety of settings and for different tasks: in a laboratory setting with a text learning task (Chapter 2 and 3), with medical diagnosis (Chapter 4), and in a classroom setting (Chapter 6). Furthermore, when compared to high performers, low performers benefitted just as much from the performance feedback to improve their calibration accuracy (Chapters 2, 4, and 6), or even benefitted more (Chapter 3). In this section, the results of the studies are discussed per research question.

Research question I: Does performance feedback improve calibration accuracy?

Feedback literature argues that to improve task performance, feedback should ideally be detailed—informing students about their performance, what the correct answers should have been, and how students can improve their performance on future tasks (Nicol & Macfarlane-Dick, 2006). This dissertation indicates that simpler forms of feedback, such as performance feedback, can already be effective to improve calibration accuracy. Students who could compare their answer to a standard had better calibration accuracy than students who could not. The underlying explanation for our findings is that by receiving performance feedback, individuals obtain a valid cue of their performance, which can then be used when making estimates (Koriat, 1997; Rawson & Dunlosky, 2007). Accordingly, because students base their estimate on this valid cue, instead of an invalid one (Gutierrez de Blume, Wells, Davis, & Parker, 2017; Thiede et al., 2010), their estimate improves.

As experimental evidence was lacking, it could only be speculated that providing performance feedback could improve calibration accuracy on subsequent tasks. If students received feedback on prior tasks, would they be better calibrated on new ones? The studies in this dissertation that incorporated a feedback intervention (Chapters 2-4 and 6) uniformly support the premise that individuals can learn to improve their calibration accuracy by receiving performance feedback. This means that these students do not only show improved calibration accuracy on the current task when they receive feedback, but also provide better performance estimates on subsequent tasks. The reason for this effect is again expected to involve the cue-utilization of students. When receiving a standard, students examine whether this standard matches with their given answer and hence receive a cue about the quality of their performance (cf. Rawson & Dunlosky, 2007). Importantly, however, students also receive a cue about the quality of their initial performance estimate. Did the initial estimate match the actual outcome? Both this performance cue and calibration accuracy cue could be used by students when estimating their performance on a new task. For example, when students were too confident about their performance on similar prior tasks, they could lower their estimates on a subsequent one. Acquiring cues of own performance and calibration accuracy by receiving outcome performance (Chapter 6), or being in the position to generate this (Chapters 2-4), can therefore help students become better calibrated.

The positive effects of performance feedback on calibration accuracy add to the literature that, to date, describes a rather mixed pattern (Bol, Hacker, O'Shea, & Allen, 2005; Callender, Franco-Watkins, & Roberts, 2015; Foster et al., 2017; Hacker et al., 2000; Miller

& Geraci, 2011a). For example, in line with studies showing that performance feedback could help to improve calibration accuracy (Callender et al., 2015; Hacker et al., 2000; Miller & Geraci, 2011a; Nietfeld et al., 2006), the results show an improvement in calibration accuracy in all studies in which performance feedback was incorporated. This is in contrast to findings by Bol et al. (2005) and Foster et al. (2017). Foster et al. (2017) even showed that after 13 exams and thus 13 feedback moments, students remained unaware of their own performance. As discussed in Chapter 6, one difficulty with the studies that examined whether calibration accuracy could be improved over subsequent tasks is that many variables were manipulated simultaneously, such as the type of metacognitive training, whether students were actively engaged in generating the feedback and their reflection, the time frame, type of course, and whether students were provided with incentives. This makes it difficult to define critical elements that caused the (absence of) effects (was it feedback or something else?), which in turn makes it difficult to compare studies. This dissertation therefore provides a valuable addition to the existing literature by examining the exclusive influence of performance feedback on calibration accuracy in a highly controlled environment (Chapters 2 and 3), and also by testing the occurrence of the effect with other types of tasks and participants (Chapters 4 and 6).

In conclusion, the answer to the first research question set out in the Introduction, “Does providing performance feedback help students to enhance their calibration accuracy in such a way that they will also show better calibration accuracy on new, subsequent tasks?”, would be: Yes, it does. Asking students or professionals to estimate their performance and giving them the opportunity to compare it to their final outcome and initial performance estimates can indeed help them to become better calibrated.

Research question II: What is the effect of performance level on calibration accuracy after feedback is given?

Previous research has highlighted that, in addition to feedback, performance levels can influence the quality of performance estimates (Kruger & Dunning, 1999; Sanchez & Dunning, 2017; Sheldon et al., 2014) and potentially affect the use of feedback (Hacker et al., 2000; Nietfeld et al., 2006). We were mostly interested in whether performance level would influence the use of feedback. As a first step, however, we examined whether we indeed found differences in calibration accuracy between performance levels. Did students with low task performance show poorer calibration accuracy than students with high task performance?

The studies in this dissertation showed that performance level was related to calibration accuracy. In line with the Dunning-Kruger effect (Kruger & Dunning, 1999), high performers were better calibrated than low performers, and whereas low performers were overconfident, high performers were underconfident in Chapters 2 and 3. The same pattern emerges when looking at expertise differences (Chapters 4 and 5). Medical specialists showed better calibration accuracy than medical students, and contrary to secondary school students, university students' confidence in their performance estimates aligned with their calibration accuracy. However, while the predominant focus of the existing literature has been on overconfidence (e.g., Foster et al., 2017; Magnus, Peresetsky, & Roy, 2018; Sanchez & Dunning, 2017), this dissertation showed general underconfidence in Chapters 4 to 6. Furthermore, this underconfidence even caused high performers to calibrate less well than low performers in Chapter 5, which is in contrast to the Dunning-Kruger effect and the findings of Chapters 2 and 3. Factors that could have affected these conflicting results include the experimental versus classroom setting, and the stakes of the tests: low stakes in the case of the experimental studies, high stakes in the case of the course exams in Chapter 5. More research is needed to examine factors that elicit overconfidence or underconfidence. Furthermore, given that underconfidence was a larger problem than overconfidence in Chapters 4-6, the results underscore the importance of including underconfidence in metacognitive research. For example, underconfidence may affect a student's decision to disengage from learning and to timely continue on to a new task. Research on this topic is scarce, however, and we encourage future research to examine when and why students underestimate their performance and what consequences follow.

With a view to obtaining better insight in metacognitive awareness, Chapters 5 and 6 included second-order judgements (i.e., how confident students are in the accuracy of their performance estimates). In line with the reasoning of Miller and Geraci (2011b) that poor calibration accuracy should be accompanied with lower confidence judgements, Chapter 5 showed that students with high miscalibration were less confident in their estimates. Interestingly, and in contrast to recent findings by Fritzsche, Händel, and Kröner (2018), Chapter 5 showed that the confidence judgements given by students were independent of performance level—low performers did not give low confidence ratings by default. Instead of being solely related to performance level (Fritzsche et al., 2018; Miller & Geraci, 2011b), results showed that confidence judgements were related to calibration accuracy instead. Students who showed miscalibration—regardless of whether these students had high or low exam grades—tended to be aware of this by providing lower confidence judgements.

Although the pattern between calibration accuracy and subjective confidence emerged among adult university students in which metacognitive awareness has reached a plateau (Weil et al., 2013), secondary school students did not show an alignment between their calibration accuracy and subjective confidence. Chapter 6 showed, however, that the SOJs of secondary school students could be improved: students who, after receiving feedback, reflected on their calibration accuracy improved their SOJs over the year.

Taken together, the findings of this dissertation showed that the estimates students make of their own performance, and the confidence they assign to these estimates, can be influenced by both performance level and experience. In the literature, there is debate over how these performance differences may affect the use of feedback. On the one hand, students with low task performance are thought to benefit less from feedback because they may have difficulties to adequately understand it (Stone, 2000). On the other hand, such students have considerable room for improvement because they naturally tend to use inadequate cues more when estimating their performance (Gutierrez de Blume et al., 2017; Thiede et al., 2010). Even though performance level influenced calibration accuracy, the findings in this dissertation show that students with both high and low task performance improve their calibration after receiving feedback. No differences between high or low performers and experienced or unexperienced participants were found in Chapters 2, 4, and 6. Furthermore, favouring the explanation that low performers could benefit more from feedback because they tend to use invalid cues (Gutierrez de Blume et al., 2017; Thiede et al., 2010), Chapter 3 showed that especially students with low task performance showed improved calibration accuracy over time.

To conclude, this thesis contributes to the debate on whether low and high performance on a task influence the use of feedback to improve calibration accuracy. To answer our second research question, “Does the effectiveness of performance feedback to improve calibration accuracy depend on performance level?” the answer would be generally not. Our findings have shown that, after receiving performance feedback, low or inexperienced performers improve their calibration accuracy equally (Chapters 2, 4 and 6) or even more (Chapter 3) than high, or more experienced performers.

Theoretical and research implications

This dissertation supports the notion that individuals—from medical specialists to secondary school students—can improve their calibration accuracy by receiving performance feedback. The findings on performance feedback are consistent throughout our research: we examined its effects in different settings, with variations in the type of performance feedback, and with different groups of participants. Furthermore, our studies showed that low or inexperienced performers' calibration accuracy improves just as much, or even more, from the feedback. This demonstrates that stronger hypotheses can be made of how performance feedback enhances calibration accuracy on subsequent tasks. More specifically, the results of this thesis are in line with the premise that providing performance feedback can give students a cue of their performance which can be used when making new performance estimates (Koriat, 1997; Thiede et al., 2010). The more cues of prior performance and calibration accuracy are available, the better calibration accuracy becomes, as shown by the learning curves in calibration accuracy in Chapters 3 and 6. Hence, the studies of this dissertation fill the gap in the literature of whether students can learn how to better estimate their performance by receiving performance feedback.

A theoretical limitation is that the studies in this dissertation mainly give procedural insights. Students can improve their calibration accuracy on a subsequent task, and they are expected to do this by using better cues (Koriat, 1997; Thiede et al., 2010). However, how exactly they change their cue-utilization remains unclear. Did students indeed learn how to estimate performance by becoming more aware of how they should score their answer? Or did they adjust their estimate based on prior calibration accuracy? (i.e., "On the previous task I was overconfident so I probably will be now as well, maybe I should lower my estimate"). Or, perhaps, did they anchor their performance estimates on their prior performance?

The purpose of this dissertation was to test whether performance feedback could improve calibration accuracy. Now that the findings show that such improvement indeed can take place, there is a distinct need for future research to examine the mechanism behind this improvement. This holds especially true given the many inconsistent findings on the role of feedback to improve calibration (Bol et al., 2005; Callender et al., 2015; Foster et al., 2017; Hacker, Bol, & Bahbahani, 2008; Hacker et al., 2000; Huff & Nietfeld, 2009; Miller & Geraci, 2011a; Nietfeld et al., 2006). One way to obtain more information of how students change the cues they use is by simply asking them how they came up with their performance estimates (cf. Bol et al., 2005). For example, in Chapter 6, students had to score the cues used when estimating their exam grades. In addition, student interviews or think-aloud sessions in which students explain how they estimate their performance can provide more detailed knowledge on cue use (Schraw, 2010).

When we further understand the causes for improved calibration accuracy, we can obtain more insight into the durability of the feedback effect on calibration accuracy as well. In this dissertation, students received performance feedback over tasks and the improvement in calibration was examined on a new task or exam. However, it remains unclear whether the interventions would have lasted over longer time-frames. What would have happened when the participants in Chapters 2 to 4 returned to the labs one week later and received a similar task? Would they still be better calibrated? Chapter 6 gives reason to believe this is so, as the exams did not immediately follow upon one another, but sometimes featured time intervals of more than a month. To strengthen the empirical evidence, however, more research is needed on this topic.

Furthermore, it is unclear whether the effects on calibration accuracy transfers to different tasks. Spontaneous transfer is unlikely to occur because the standards and assessment-criteria used vary across tasks (Van Gog, Kostons, & Paas, 2010). Hence, while students improved their calibration accuracy in the subject of French in Chapter 6, they did not do so in the subject of math. However, on a course that had similar assessment-criteria (German), we also failed to find any improvements. This seems to indicate that being able to estimate own performance is task specific, instead of a general ability as is sometimes suggested (Van der Stel & Veenman, 2010). More research is needed to establish whether and, if so, how calibration accuracy can be improved across tasks.

Another important notion for future research concerns the large influence of the type of performance estimates used to measure calibration accuracy. A difficulty in the field of metacognition is that many measurement types can be used to grasp metacognitive skills. For example, students can be asked to predict their performance, to judge their performance during a task, or to postdict their performance (Nelson & Narens, 1990). The studies in this dissertation almost all focused merely on postdiction accuracy. When students postdict instead of predict their performance, they benefit from privileged knowledge about the task. This can prevent them from making up excuses about the test content. For example, the statement that their miscalibration was because exam questions did not properly represent the learning materials (Bol et al., 2005) does not hold for postdictions. Students knew the content and questions when judging their performance, and should have incorporated this in their estimate. In support of the hypothesis that improvement in calibration accuracy after feedback is given can be influenced by estimates given, Chapter 2 shows that whereas postdiction accuracy improved, prediction accuracy could become even worse due to underconfidence for average performance. It is therefore important for future research to acknowledge that results on postdictions do not directly generalize to other types of estimates. This may also be the reason for non-significant results in such well-cited meta-analyses as Sitzmann and Ely (2011), who generalized among studies using predictions, postdictions, and metacognitive questionnaires. To examine the effects of monitoring, it is essential that future research distinguishes between different types of judgements to prevent premature generalizations.

Research on calibration accuracy has the ultimate goal to improve students' performance. By becoming more aware of their current performance and their calibration accuracy, students can select better learning strategies and control decisions (Butler & Winne, 1995; Fernandez & Jamet, 2017; Koriat, 2012; Metcalfe, 2009; Nelson & Narens, 1990; Tuysuzoglu & Greene, 2015). The results of Chapter 2 and 6 support the hypothesis that an improvement in calibration accuracy is related to better task performance. Students in both chapters improved their calibration accuracy as well as their task performance on new tasks or exams. Now that this dissertation shows that performance feedback can be effectively used to enhance calibration accuracy, more research is required to further examine how this improved calibration accuracy influences control decisions and, consequently, performance.

Practical implications

The studies in this dissertation have wide-ranging practical implications. Self-regulation and the ability to adapt oneself to new situations or tasks through monitoring have become increasingly important in today's educational contexts (Trilling & Fadel, 2009; Wolters, 2010). By examining feasible performance feedback interventions, this dissertation gives guidelines as to how monitoring can be improved. Importantly, the effects of performance feedback on calibration accuracy apply to students of both different performance levels and of different experience groups. Performance feedback interventions appear successful for many different students, both in university and secondary school, and even for medical specialists. The current dissertation therefore underscores the potential of (1) having students estimate their performance, and (2) providing them with self-testing opportunities in which they can generate feedback on the accuracy of their performance and calibration accuracy, or in which they receive this feedback directly. The practical value of both activities lie in the ease in which they can be implemented in both educational and organizational practice. Furthermore, results are obtained quickly. Already after three feedback moments, medical students and specialists in our study increased their calibration accuracy, and in over six cases, students already showed a learning curve. Hence, to help students improve their calibration accuracy in a simple, efficient, and time-saving manner, encouraging them to estimate their performance and to compare this estimate to the final outcome is recommended.

Conclusion

The main purpose of this dissertation was to provide more insight into whether it is possible to enhance calibration accuracy on new, subsequent tasks when students are provided with performance feedback. The series of studies uniformly support the idea that performance feedback is indeed effective for improving calibration accuracy, for both high and low performing students. With these findings, we hope to have set the stage for new research venues that further unravel the theoretical underpinnings of how calibration accuracy improves after performance feedback. Furthermore, our findings contribute to educational practice. To improve students' calibration accuracy, teachers should advice their students to provide performance estimates after they have conducted a task, and to attend to outcome feedback, or to generate this feedback themselves by using standards. Doing so helps students to become more aware of both their final performance and their calibration accuracy, and this, in turn, helps them to provide better performance estimates on subsequent tasks, even when feedback is absent.

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. https://doi.org/10.1016/j.learninstruc.2012.10.003

Archer, J. C. (2010). State of the Science in Health Professional Education: Effective Feedback. *Medical Education*, *44*(1), 101–108. https://doi.org/10.1111/j.1365-2923.2009.03546.x

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

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine*, *121*(5), 2–23. https://doi.org/10.1016/j.amjmed.2008.01.001

Blendon, R. J., DesRoches, C. M., Brodie, M., Benson, J. M., Rosen, A. B., Schneider, E., … Steffenson, A. E. (2002). Views of practicing physicians and the public on medical errors. *New England Journal of Medicine*, *347*(24), 1933–1940. https://doi.org/10.1056/NEJMsa022151

Blom, S., & Severiens, S. (2008). Engagement in self-regulated deep learning of successful immigrant and non-immigrant students in inner city schools. *European Journal of Psychology of Education*, *23*(1), 41-58. https://doi.org/10.1007/BF03173139

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education*, *69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology*, *3*, 1–6. https://doi.org/10.3389/fpsyg.2012.00229

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *Journal of Experimental Education*, *73*(4), 269–290. https://doi.org/10.3200/JEXE.73.4.269-290

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education, 38*(8), 941-956. https://doi.org/: 10.1080/02602938.2013.769198

Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 444–457. https://doi.org/10.1080/0969594X.2014.996523

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

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2015). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*. https://doi.org/10.1007/s11409-015-9142-6

Chevalier, A., Gibbons, S., Thorpe, A., Snell, M., & Hoskins, S. (2009). Students' academic self-perception. *Economics of Education Review*, *28*(6), 716–727. https://doi.org/10.1016/j.econedurev.2009.06.007

Custers, E. J. F. M. (1995). *The development and function of illness scripts: studies on the structure of medical diagnostic knowledge*. Maastricht University.

Custers, E. J., Boshuizen, H. P., & Schmidt, H. G. (1996). The influence of medical expertise, case typicality, and illness script component on case processing and disease probability estimates. *Memory & Cognition*, *24*(3), 384–399. https://doi.org/10.3758/BF03213301

Davis, D. A., Mazmanian, P. E., Fordis, M., Harrison, R. Van, Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. *Journal of the American Medical Association*, *296*(9), 1094–1102. https://doi.org/10.1001/jama.296.9.1094

De Bruin, A. B. H., & Van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, *22*(4), 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003

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., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

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., & Thiede, K. W. (2013a). 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

Dunlosky, J., & Thiede, K. W. (2013b). Metamemory. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 283–298). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0019

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., & 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., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *Journal of General Psychology*, *132*(4), 335–346. https://doi.org/10.3200/GENP.132.4.335-346

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*(1), 5–17. https://doi.org/10.1037/0022-3514.84.1.5

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

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: a perspective from the expert-performance approach with deliberate practice. *Academic Medicine*, *90*(11), 1471–86. https://doi.org/10.1097/ACM.0000000000000939

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Psychological Review*, *100*(3), 363–406.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: a reformulation and research agenda. *Academic Medicine*, *80*(10), S46–S54. https://doi.org/10.1097/00001888-200510001-00015

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

Finn, B., & Metcalfe, J. (2014). Overconfidence in children’s multi-trial judgments of learning. *Learning and Instruction*, *32*, 1–9. https://doi.org/10.1016/J.LEARNINSTRUC.2014.01.001

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students’ experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*(4), 567–586. https://doi.org/10.1007/s10648-015-9313-7

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

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., … Elstein, A. S. (2005). Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. *Journal of General Internal Medicine*, *20*(4), 334–9. https://doi.org/10.1111/j.1525-1497.2005.30145.x

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

García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students’ metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning*, *11*(2), 139–170. https://doi.org/10.1007/s11409-015-9139-1

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. https://doi.org/10.1080/20445911.2016.1229669

Glaser, M., & Weber, M. (2007). Why inexperienced investors do not learn: They do not know their past portfolio performance. *Finance Research Letters*, *4*(4), 203–216. https://doi.org/10.1016/J.FRL.2007.10.001

Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology*, *106*(1), 58–68. https://doi.org/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., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101–121. https://doi.org/10.1007/s11409-008-9021-5

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

Händel, M., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students. *Memory and Cognition*, *44*(2), 229–241. https://doi.org/10.3758/s13421-015-0552-0

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

Hayes, A. F. (2013). Introduction to mediation, moderation, and conditional process analysis. New York, NY: Guilford Press.

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning*, *4*(2), 161–176. https://doi.org/10.1007/s11409-009-9042-8

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

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory*, *16*(2), 125–136. https://doi.org/10.1080/09658210701763899

Krosnick, J. A., & Pesser S. (2009). Question and Questionnaire Design. In J. Wright, & P. Marsden (Eds), *Handbook of Survey Research* (pp. 263-314) San Diego, CA: Elsevier.

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

Landy, F. J., & Conte, J. M. (2009). *Work in the 21st Century: An Introduction to Industrial and Organizational Psychology* (3rd ed.). Hoboken, New Jersey: John Wiley & Sons.

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.

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

Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgements. *International Journal of Behavioral Development*, *26*(4), 327–333. https://doi.org/10.1080/01650250143000210

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

Martin, D., Regehr, G., Hodges, B., & McNaughton, N. (1998). Using videotaped benchmarks to improve the self-assessment ability of family practice residents. *Academic Medicine*, *73*, 1201–1206.

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

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review*, *15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. *Learning and Instruction*, *22*(4), 253–261. https://doi.org/10.1016/j.learninstruc.2011.10.004

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, *173*(21), 1952–8. https://doi.org/10.1001/jamainternmed.2013.10081

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

Miller, T. M., & Geraci, L. (2011b). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 502–506. https://doi.org/10.1037/a0021802

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. J. P. (Submitted). Learning to Calibrate: Providing Standards to Improve Calibration Accuracy for Different Performance Levels.

Nederhand, M. L., Tabbers, H. K., Abrahimi, H., & Rikers, R. M. J. P. (Accepted). Improving calibration over texts by providing standards both with and without idea-units. J*ournal of Cognitive Psychology*.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*(26), 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

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

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, *24*(2), 124–139. https://doi.org/10.1006/ceps.1998.0991

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *Journal of Experimental Education*, *65*(3), 213–228. https://doi.org/10.1080/00220973.1997.9943455

Paulus, M., Tsalas, N., Proust, J., & Sodian, B. (2014). Metacognitive monitoring of oneself and others: Developmental changes during childhood and adolescence. *Journal of Experimental Child Psychology*, *122*(1), 153–165. https://doi.org/10.1016/j.jecp.2013.12.011

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin and Review*, *24*, 1774–1784. https://doi.org/10.3758/s13423-017-1242-7

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40.

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

Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology*, *109*(1), 99–117. https://doi.org/10.1037/edu0000132

Sadler, D. R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science, 18*(1), 119–144.

Sanchez, C., & Dunning, D. (2017). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10–28. https://doi.org/10.1037/pspa0000102

Sargeant, J., Armson, H., Chesluk, B., Dornan, T., Eva, K., Holmboe, E., … van der Vleuten, C. (2010). The processes and dimensions of informed self-assessment: A conceptual model. *Academic Medicine*, *85*(7), 1212–1220. https://doi.org/10.1097/ACM.0b013e3181d85a4e

Schmidt, H. G., Mamede, S., Berge, K. Van Den, Gog, T. Van, Saase, J. L. C. M. Van, & Rikers, R. M. J. P. (2014). Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine*, *89*(2), 285–291. https://doi.org/10.1097/ACM.0000000000000107

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*(3), 114–121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

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

Schraw, G. (2010). Measuring self-regulation in computer-based learning environments. *Educational Psychologist*, *45*(4), 258–266. https://doi.org/10.1080/00461520.2010.515936

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology*, *18*(4), 455–463. https://doi.org/10.1006/ceps.1993.1034

Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. *Memory and Cognition*, *44*(7), 1127–1137. https://doi.org/10.3758/s13421-016-0624-9

Sheldon, O. J., Dunning, D., & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: reactions to feedback about deficits in emotional intelligence. *The Journal of Applied Psychology*, *99*(1), 125–37. https://doi.org/10.1037/a0034138

Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*. https://doi.org/10.1037/a0022777

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

Thorpe, A., Snell, M., Hoskins, S., & Bryant, J. (2007). False uniqueness: The self-perception of new entrants to higher education in the UK and its implications for access – a pilot study. *Higher Education Quarterly*, *61*(1), 3–22. https://doi.org/10.1111/j.1468-2273.2006.00335.x

Trilling, B., & Fadel, C. (2009). *21st Century skills: Learning for life in our times*. San Francisco, CA: John Wiley & Sons.

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 der Stel, M., & Veenman, M. V. J. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences*, *20*(3), 220–224. https://doi.org/10.1016/j.lindif.2009.11.005

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 Gog, T., Kostons, D., & Paas, F. (2010). Teaching students self-assessment and task-selection skills with video-based modeling examples. In *S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 296–301). Austin, TX: Cognitive Science Society*.

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

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., … Blakemore, S. J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271. https://doi.org/10.1016/j.concog.2013.01.004

Wolters, C. A. (2010). Self-regulated learning and the 21st century competencies. https://doi.org/Availabe online at: http://www.hewlett.org/uploads/Self_Regulated_Learning__21st_Century_Competencies.pdf.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3–17.

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.

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

Nederlandse samenvatting

Summary and discussion in Dutch

Mensen vinden het vaak erg moeilijk een juiste schatting te maken van hun eigen prestatie (Dunlosky & Lipko, 2007; Sheldon et al., 2014). Het verkeerd inschatten van de eigen prestatie (ook wel miskalibratie genoemd) kan echter negatieve gevolgen hebben voor de beslissingen die mensen maken nadat ze een bepaalde taak hebben gedaan (Dunlosky & Rawson, 2012; Nelson & Narens, 1990). Zo zal een arts die ten onrechte overtuigd is van de juistheid van de gestelde diagnose, geen actie ondernemen deze diagnose te corrigeren. Op eenzelfde manier zal een student die denkt de leerstof onder de knie te hebben geen verdere aandacht besteden aan zijn schoolboeken en in plaats daarvan stoppen met leren.

Omdat miskalibratie belangrijke gevolgen kan hebben voor de prestatie en leerprocessen van mensen, is het opmerkelijk dat we weinig weten over hoe deze kalibratie verbeterd kan worden. In dit proefschrift werden daarom interventies getoetst met als doel om kalibratieaccuratesse te verbeteren. Prestatiefeedback speelt hierbij een centrale rol. Zulke feedback geeft mensen meer inzicht in hun prestatie en dit inzicht kan hen vervolgens helpen hun kalibratieaccuratesse te verbeteren (Dunlosky et al., 2011; Lipko et al., 2009; Rawson & Dunlosky, 2007). Hoe prestatiefeedback de kalibratie vervolgens op nieuwe en toekomstige taken verbetert, was echter nog onduidelijk. De eerste vraag die in dit proefschrift werd beantwoord is daarom of het aanbieden van prestatiefeedback studenten kan helpen om betere schattingen van hun prestatie te maken op nieuwe taken.

Een tweede doel van dit proefschrift was om beter in kaart te brengen hoe prestatieniveau invloed heeft op kalibratieaccuratesse nadat studenten feedback hebben gekregen. Eerder onderzoek toont aan dat studenten met een lage taakprestatie vaak moeite hebben deze prestatie accuraat in te schatten. Ze laten hierbij met name overschatting zien. Dit in tegenstelling tot studenten met een hoge taakprestatie, die vaak juist accurate schattingen maken maar soms de neiging hebben tot onderschatting (e.g., Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999). Deze verschillen in kalibratieaccuratesse kunnen van invloed zijn op het effect van feedback (Gutierrez de Blume et al., 2017; Stone, 2000; Thiede et al., 2010), maar hoe precies bleef wederom onduidelijk. In dit proefschrift werd daarom eveneens onderzocht of prestatieniveau (geoperationaliseerd als verschillen in taakprestatie en verschillen in ervaring) van invloed is op kalibratieaccuratesse en het effect van feedback. Om deze twee hoofdvragen te onderzoeken zijn vijf studies verricht die hieronder worden samengevat.

Samenvatting van de hoofdbevindingen

Hoofdstuk 2 beschrijft een experiment waarin onderzocht werd of het geven van het juiste antwoord (een standaard), studenten kan helpen om hun kalibratie te verbeteren op nieuwe taken waar zulke standaarden niet langer beschikbaar zijn. Studenten moesten verschillende teksten lezen en per tekst vier definities leren. Na het lezen van elke tekst en het leren van de bijbehorende vier definities, kregen studenten een toets waarin ze de definities uit hun geheugen moesten ophalen. Vervolgens maakten studenten een schatting van hun prestatie; was hun opgehaalde definitie correct, deels correct, of incorrect? Studenten waren verdeeld in twee experimentele groepen. De eerste groep kreeg tijdens het inschatten van de eigen prestatie de correcte definitie in beeld, terwijl ze ook hun eigen opgehaalde definitie zagen. Dit betekent dat studenten in deze standaardgroep hun eigen opgehaalde definitie direct konden vergelijken met de beoogde definitie. De andere helft van de studenten kreeg geen standaard te zien.

Kalibratieaccuratesse van alle studenten werd getest op nieuwe teksten, waarbij standaarden niet direct toegankelijk waren. Er werd verwacht dat doordat studenten extra inzicht kregen in zowel hun geheugenprestatie (leek hun opgehaalde definitie op de standaard?) als hun kalibratieaccuratesse (hadden ze een goede schatting gemaakt van de juistheid van hun antwoord?), standaarden zouden leiden tot een verbeterde kalibratie en verminderde overschatting. De resultaten ondersteunden deze hypothese. Belangrijker, het geven van standaarden zorgde ook voor een verbetering in de kalibratieaccuratesse op een nieuwe taak. Daarnaast werd gevonden dat hoewel studenten met een hoge taakprestatie beter gekalibreerd waren dan studenten met een lage taakprestatie, er geen verschil was tussen deze twee groepen in de mate van verbetering in hun kalibratieaccuratesse.

Dat kalibratieaccuratesse van zowel studenten met een hoge als lage taakprestatie verbeterd kan worden met een simpele feedbackinterventie is veelbelovend, met name omdat het geven van prestatiefeedback en standaarden makkelijk geïmplementeerd kan worden in de onderwijspraktijk. Echter, uit Hoofdstuk 2 bleek dat er in de kalibratie-accuratesse van de studenten nog steeds ruimte voor verbetering was, ook na het krijgen van prestatiestandaarden. Mogelijkerwijs zouden studenten profijt hebben van preciezere standaarden. **Hoofdstuk 3** beschrijft daarom een follow-up studie met dezelfde materialen en grotendeels hetzelfde design als in Hoofdstuk 2. Nu werd echter bekeken of het aanbieden van extra richtlijnen over hoe de standaarden gebruikt konden worden zou leiden tot een verdere verbetering in kalibratieaccuratesse. Opnieuw moesten studenten teksten lezen en definities leren die ze later tijdens een test moesten ophalen. Een groep kreeg standaarden die gelijk waren aan de standaarden in Hoofdstuk 2. De andere groep kreeg standaarden die meer gedetailleerd waren, *idea-unit* standaarden. Bij deze idea-unit standaarden was de definitie in verschillende delen gesplitst. Wanneer elk van de delen was benoemd bij het ophalen van de definitie, kreeg de student het maximaal aantal punten.

Had de student echter een of meerdere delen gemist, dan was de definitie slechts deels correct of zelfs incorrect. De verwachting was dat het geven van idea-unit standaarden studenten verder zou helpen om hun kalibratieaccuratesse op nieuwe teksten te verbeteren.

De resultaten van Hoofdstuk 3 lieten zien dat studenten die idea-unit standaarden kregen, beter kalibreerden en minder overschatting lieten zien dan studenten die normale standaarden kregen. Daarnaast werd er een significante leercurve in kalibratieaccuratesse gevonden. Hoe meer standaarden studenten hadden ontvangen, hoe beter hun kalibratieaccuratesse werd op nieuwe teksten waar geen standaarden meer aanwezig waren. Hierbij gold dat het type standaard van invloed was op de verbetering in kalibratieaccuratesse over teksten heen: idea-unit standaarden leidden tot een grotere verbetering dan normale standaarden. Ook het prestatieniveau van de studenten was van invloed. Vooral studenten die laag scoorden op de ophaaltaak hadden profijt van de standaarden: zij lieten een sterkere verbetering in hun kalibratie zien dan studenten met een hoge taakprestatie.

In de Hoofdstukken 2 en 3 werd prestatieniveau geoperationaliseerd als verschillen in prestatie op de taak (cf. Ehrlinger et al., 2008; Hacker, Bol, Horgan, & Rakow, 2000; Kruger & Dunning, 1999; Nietfeld, Cao, & Osborne, 2006). Dit is een veelgebruikte methode die weliswaar inzicht geeft in verschillen tussen studenten, maar minder duidelijkheid biedt over verschillen in groepen die in grotere mate variëren in termen van ervaring. Zouden dezelfde resultaten gevonden worden wanneer bijvoorbeeld zeer ervaren medisch specialisten vergeleken werden met beginnende medische studenten? En zou deze kalibratieaccuratesse in verschillende ervaringsgroepen eveneens verbeterd kunnen worden met prestatiefeedback?

De studie in **Hoofdstuk 4** werd uitgevoerd onder medisch specialisten en geneeskundestudenten. De specialisten en studenten kregen zes verschillende klinische casus waarvoor zij een diagnose moesten opstellen. Na het diagnosticeren van elke casus gaven de specialisten en studenten een schatting van de juistheid van hun diagnose. Daarna kreeg de helft van de studenten en specialisten feedback: zij kregen de juiste diagnose te zien voordat zij verder gingen met een nieuwe casus. De andere helft van de specialisten en studenten kreeg deze feedback niet. Dit herhaalde zich drie keer, waarna alle specialisten en studenten drie nieuwe casus kregen. Er werd getest of (1) het krijgen van feedback op drie voorgaande casus de kalibratieaccuratesse op de drie nieuwe casus had verbeterd, en (2) of medisch specialisten beter kalibreerden dan de geneeskundestudenten. De resultaten van Hoofdstuk 4 lieten zien dat de medisch specialisten betere schattingen maakten dan de geneeskundestudenten. Ook zorgde het aanbieden van prestatiefeedback voor een verbetering in de kalibratieaccuratesse op de drie nieuwe casus voor zowel specialisten als studenten.

De Hoofdstukken 2 tot 4 ondersteunen de hypothese dat het geven van prestatiefeedback kan helpen om kalibratieaccuratesse te verbeteren op nieuwe taken. Daarnaast lieten de studies zien dat, in vergelijking met mensen met een hoge taakprestatie, de kalibratieaccuratesse van mensen die een lage taakprestatie hebben evenveel (Hoofdstuk 2 en 4) of zelfs meer (Hoofdstuk 3) verbetert na de feedback. In deze hoofdstukken werd telkens gemeten in welke mate individuen in staat zijn hun prestatie te schatten (Miller & Geraci, 2011). Volgens Miller en Geraci (2011) zijn er echter ook verschillen in hoe zeker mensen zijn over hun prestatieschattingen. Dit wordt gemeten met zogeheten *second-order judgements* (SOJs). Miller en Geraci (2011) toonden aan dat studenten met een lage prestatie vaak niet alleen lage kalibratieaccuratesse laten zien, maar ook onzeker bleken te zijn over hun schatting. Ze leken zich dus bewust te zijn dat hun gegeven prestatieschatting niet juist is. Om meer inzicht te krijgen in hoeverre mensen zich bewust zijn van een lage kalibratieaccuratesse werden ook SOJs meegenomen als uitkomstmaat in de hoofdstukken 5 en 6.

Hoofdstuk 5 beschijft een observationele studie waarin kalibratieaccuratesse, bias (de mate van over- of onderschatting) en SOJs werden onderzocht bij universiteitsstudenten en middelbare scholieren. Nadat de studenten en leerlingen een toets hadden gemaakt, maakten zij een schatting van het cijfer dat zij behaald dachten te hebben. Daarnaast gaven zij aan hoe zeker zij waren van hun schatting. De mate van miskalibratie en bias werd bekeken, en er werd onderzocht of studenten die veel miskalibratie lieten zien ook onzekerder waren over hun schattingen (cf. Miller & Geraci, 2011). Ook verschillen tussen prestatieniveaus werden bekeken, waarbij prestatieniveau geoperationaliseerd werd als zowel prestatieverschillen op de toets (hoog versus laag presterende studenten) als ervaringsverschillen (universiteit versus middelbare school).

In tegenstelling tot eerdere bevindingen (e.g., Foster et al., 2017; Kruger & Dunning, 1999) lieten de resultaten zien dat studenten op zowel de universiteit als middelbare school zichzelf met name onderschatten. Universiteitsstudenten bleken zich echter meer bewust van hun miskalibratie, blijkend uit een significante relatie tussen de kalibratieaccuratesse en SOJs. Hoe slechter de kalibratie, des te minder vertrouwen studenten aan gaven te hebben in deze schatting. Dit patroon was onafhankelijk van de toetsprestatie van de studenten: zowel studenten met een lage als hoge toetsprestatie toonden zich bewust van hun miskalibratie. De relatie tussen kalibratieaccuratesse en SOJs werd echter niet gevonden bij de middelbare scholieren. De kwaliteit van de schatting was niet gerelateerd in het vertrouwen dat de scholieren hadden in die schatting—hun metacognitieve bewustzijn leek minder ontwikkeld.

De laatste studie van dit proefschrift wordt beschreven in **Hoofdstuk 6**. Gedurende een volledig schooljaar deden middelbare scholieren (atheneum) mee aan een interventiestudie met als doel om hun kalibratieaccuratesse te verbeteren. Na het maken van elke toets schatten de scholieren het cijfer dat zij dachten te gaan halen, en gaven zij aan hoeveel vertrouwen zij hadden in die schatting. Tijdens de volgende les kregen de studenten hun daadwerkelijke cijfer te horen. De scholieren werden ingedeeld in drie verschillende groepen waarin de ondersteuning om te reflecteren op de match tussen de geschatte en daadwerkelijke prestatie werd gevarieerd. De eerste groep hoefde enkel hun cijfer in te schatten. De tweede groep moest hun cijfer schatten, maar ook het verschil berekenen tussen hun daadwerkelijk behaalde cijfer en hun geschatte cijfer. De derde en laatste groep bestond uit studenten die, bovenop de eerdergenoemde stappen, reflecteerden over hoe hun schatting tot stand was gekomen en (indien van toepassing) waarom deze afweek van hun daadwerkelijke prestatie. De effectiviteit van de interventie op kalibratieaccuratesse, bias, SOJs en toetsprestatie werd onderzocht.

Er werd verwacht dat de hoeveelheid ondersteuning om te reflecteren samen zou hangen met de verbetering in kalibratieaccuratesse over het jaar. De resultaten van de studie lieten echter zien dat de kalibratieaccuratesse van de leerlingen verbeterde bij elke interventiegroep. Het simpelweg inschatten van de eigen prestatie hielp studenten om hun kalibratie te verbeteren op volgende toetsen. Naast dat zij hun kalibratieaccuratesse verbeterden, pasten leerlingen in de reflectiegroep daarbij ook hun SOJs aan. Gedurende het jaar kregen zij meer vertrouwen in hun schattingen. Er waren geen verschillen tussen leerlingen met een goede en minder goede toetsprestatie voor wat betreft de verbetering in hun kalibratieaccuratesse. Dit ondersteunt wederom dat prestatieniveau geen of minimale invloed heeft op hoe feedback gebruikt wordt om de kalibratieaccuratesse te verbeteren. Op vakken waar geen interventie was geïmplementeerd werd geen verbetering in kalibratieaccuratesse gevonden.

Algemene discussie

De studies in dit proefschrift laten zien dat het geven van prestatiefeedback helpt om de kalibratieaccuratesse op zowel een huidige als een volgende taak te verbeteren. Dit effect is aangetoond in een verscheidenheid aan settingen en bij verschillende taken: in een laboratoriumsetting met een geheugentaak (Hoofdstukken 2 en 3), bij het stellen van medische diagnoses (Hoofdstuk 4) en in een schoolsetting (Hoofdstuk 6). Daarnaast lijken studenten die goed en minder goed presteren vaak evenveel profijt te hebben van de feedbackinterventie. Naar aanleiding van de bevindingen kunnen de onderzoeksvragen worden beantwoord, en zullen de implicaties worden bediscussieerd.

Onderzoeksvraag I: Kan prestatiefeedback helpen om de kalibratieaccuratesse te verbeteren?

In de literatuur wordt beschreven dat effectieve feedback aan verschillende eisen moet voldoen. Studenten moeten ten eerste informatie krijgen over hun uiteindelijke prestatie, ze moeten inzicht krijgen in wat de correcte antwoorden hadden moeten zijn, en ook in hoe ze hun prestatie op een nieuwe taak kunnen verbeteren (Nicol & Macfarlane-Dick, 2006). De studies in dit proefschrift laten echter zien dat simpelere vormen van feedback, zoals prestatiefeedback, al effectief kunnen zijn om kalibratieaccuratesse te verbeteren. Studenten die hun antwoorden konden vergelijken met een standaard hadden een betere kalibratieaccuratesse dan studenten die geen standaarden tot hun beschikking hadden. Dit kan verklaard worden doordat studenten door het krijgen van feedback een valide cue krijgen van hun prestatie (Koriat, 1997; Rawson & Dunlosky, 2007). Het gebruiken van die cue wanneer studenten hun prestatie inschatten, leidt vervolgens tot een verbetering in de kalibratieaccuratesse (Gutierrez de Blume, Wells, Davis, & Parker, 2017; Thiede et al., 2010).

Door ontbrekend empirisch bewijs kon slechts gespeculeerd worden of prestatiefeedback ook effect zou hebben op de kalibratieaccuratesse op volgende, nieuwe taken. Als studenten feedback hadden ontvangen op een vorige taak, zouden zij dan hun kalibratie op nieuwe taken verbeteren? De studies in dit proefschrift waarin een feedbackinterventie werd getest (Hoofdstukken 2-4 en 6) laten uniform zien dat mensen hun kalibratieaccuratesse kunnen leren verbeteren door het krijgen van prestatiefeedback. Dit betekent dat studenten niet alleen een verbetering in hun kalibratieaccuratesse laten zien op de huidige taak waar ze direct feedback ontvangen, maar dat deze verbetering zich ook voordoet op een nieuwe taak. De reden voor dit effect is eveneens gelegen in het gebruik van cues. Wanneer studenten een standaard krijgen, kunnen zij evalueren of deze standaard overeenkomt met hun gegeven antwoord. Hierdoor krijgen zij een aanwijzing over de kwaliteit van hun prestatie (cf. Rawson & Dunlosky, 2007). Ze krijgen echter ook een aanwijzing van de kwaliteit van hun eerdere prestatieschatting: kwam hun gegeven schatting overeen met de daadwerkelijke uitkomst? Zowel deze prestatie- als kalibratieaanwijzing kunnen vervolgens door de studenten gebruikt worden wanneer zij een schatting maken over een nieuwe taak. Wanneer studenten bijvoorbeeld inzien dat zij zichzelf hebben overschat op vorige taken, kunnen zij hun schattingen iets naar beneden bijstellen op daaropvolgende taken. Het krijgen van uitkomstfeedback (Hoofdstuk 6), of het zelf genereren van prestatiefeedback (Hoofdstuk 2-4), kan daarom helpen om de kalibratieaccuratesse van studenten te verbeteren.

De gevonden positieve effecten van feedback dragen bij aan de literatuur die vooralsnog een gemixt beeld gaf over de effectiviteit van performance feedback (Bol, Hacker, O'Shea, & Allen, 2005; Callender, Franco-Watkins, & Roberts, 2015; Foster et al., 2017; Hacker et al., 2000; Miller & Geraci, 2011a). Zo zijn de bevindingen in lijn met onderzoek waarin kalibratieaccuratesse verbeterde wanneer studenten inzicht kregen in, en aangemoedigd werden om te reflecteren op, de match tussen geschatte en daadwerkelijke prestatie (Callender, Franco-Watkins, & Roberts, 2015; Hacker, Bol, Horgan, & Rakow, 2000; Miller & Geraci, 2011a; Nietfeld, Cao, & Osborne, 2006). Het onderzoek is echter niet in overeenstemming met Bol et al. (2005) en Foster et al. (2017). In de studie van Foster hebben studenten zelfs na dertien examens nog steeds geen verbetering laten zien in hun kalibratieaccuratesse. Een reden voor de conflicterende bevindingen in de literatuur is dat er vaak verschillende variabelen tegelijk gemanipuleerd worden, zoals het type feedback, de inhoud van een metacognitieve training die wordt gegeven, of er een metacognitieve training wordt gegeven, de tijdspanne waarover de interventie loopt en of studenten beloond werden voor een goede kalibratie. Het is daardoor niet alleen lastig om te bekijken welke variabelen van invloed zijn, maar ook vergelijkingen tussen studies wordt hierdoor bemoeilijkt. Het huidige proefschrift draagt daarom in bij aan de huidige literatuur door de effecten van feedback te onderzoeken in een sterk gecontroleerde setting (Hoofdstuk 2 en 3), en vervolgens de houdbaarheid van het effect te toetsen bij andere taken en proefpersonen (Hoofdstuk 4 en 6).

Het antwoord op de onderzoeksvraag "Kan prestatiefeedback helpen om kalibratieaccuratesse te verbeteren zodat studenten ook een betere kalibratieaccuratesse laten zien op volgende, nieuwe taken?" luidt: Ja, prestatiefeedback kan inderdaad ingezet worden om kalibratieaccuratesse te verbeteren, ook op nieuwe taken. Studenten of professionals vragen om hun prestatie in te schatten waarna ze de mogelijkheid krijgen om deze schatting te vergelijken met hun uiteindelijke uitkomst kan hen inderdaad helpen om beter te kalibreren.

Onderzoeksvraag II: Wat is het effect van prestatieniveau op kalibratieaccuratesse nadat feedback is gegeven?

Eerder onderzoek heeft aangetoond dat prestatieniveau van invloed is op de kalibratieaccuratesse (Kruger & Dunning, 1999; Sanchez & Dunning, 2017; Sheldon et al., 2014). Daarbij is eveneens gevonden dat het van invloed kan zijn op de effectiviteit van feedback om kalibratie te verbeteren (Hacker et al., 2000; Nietfeld et al., 2006). De hoofdvraag in dit proefschrift was daarom of het prestatieniveau van de studenten de effecten van feedback kon beïnvloeden. Als een eerste stap onderzochten we echter of er inderdaad verschillen waren in de kalibratieaccuratesse tussen de verschillende prestatieniveaus. Hadden studenten met een lagere taakprestatie inderdaad een slechtere kalibratieaccuratesse dan studenten met een hoge taakprestatie?

De resultaten van de studies in dit proefschrift bevestigen dat prestatieniveau van invloed is op de kalibratie. Overeenkomstig met het Dunning-Kruger effect (Kruger & Dunning, 1999) tonen de studies in Hoofdstuk 2 en 3 aan dat studenten die goed presteren een betere kalibratieaccuratesse hebben dan studenten die minder goed presteren. Deze laatste groep overschat zichzelf daarnaast, terwijl studenten die goed presteren zichzelf onderschatten. Hetzelfde patroon doet zich voor wanneer we kijken naar expertiseverschillen (Hoofdstuk 4 en 5). Medisch specialisten hebben een betere kalibratieaccuratesse dan geneeskundestudenten (Hoofdstuk 4), en in tegenstelling tot middelbare scholieren, lijken universiteitsstudenten wel een metacognitief bewustzijn te laten zien (Hoofdstuk 5). Hoewel in eerder onderzoek echter met name overschatting is beschreven (e.g., Foster et al., 2017; Magnus, Peresetsky, & Roy, 2018; Sanchez & Dunning, 2017), tonen Hoofdstuk 4 tot 6 aan dat de proefpersonen zichzelf over het algemeen onder- in plaats van overschatten. Deze onderschatting zorgde er tevens voor dat in Hoofdstuk 5 studenten die goed presteerden slechter kalibreerden dan studenten met een lage taakprestatie. Factoren die kunnen verklaren waar de onderschatting precies door veroorzaakt werd omvatten onder andere de schoolsetting in plaats van experimentele labsetting, en of er veel van de test afhangt. In de experimentele studies hangt er weinig af van de testen die worden gemaakt, maar in de schoolsetting hangt er een stuk meer van af. Additioneel onderzoek is nodig om te identificeren welke factoren ofwel overschatting ofwel onderschatting veroorzaken. Daarnaast moet het meenemen van onderschatting in toekomstig metacognitief onderzoek niet worden vergeten, gegeven dat onderschatting een groter probleem leek dan overschatting in Hoofdstuk 4 tot 6. Het onderschatten van de eigen prestatie kan ertoe leiden dat studenten moeite hebben om te stoppen met leren en tijdig door te gaan naar een nieuwe taak. Onderzoek op dit gebied is echter schaars en we bemoedigen toekomstig onderzoek daarom te onderzoeken wanneer en hoe studenten hun prestatie onderschatten en welke consequenties dit heeft.

Om beter inzicht te krijgen in metacognitief bewustzijn zijn in Hoofdstuk 5 en 6 "second-order judgements" geïncludeerd als uitkomstmaat. Overeenkomstig met de beredenering van Miller en Geraci (2011b) dat een slechte kalibratieaccuratesse gepaard zou moeten gaan met weinig vertrouwen in de gegeven schatting, liet Hoofdstuk 5 zien dat studenten met een hoge miskalibratie minder vertrouwen hadden in hun gegeven schatting. In tegenstelling tot de bevindingen van Fritzsche, Händel, and Kröner (2018), liet Hoofdstuk 5 daarnaast zien dat dit effect niet afhankelijk was van het prestatieniveau van de studenten. Studenten met een lage prestatie gaven niet per definitie aan minder vertrouwen te hebben in hun schatting. In plaats dat de schattingen enkel gerelateerd waren aan prestatieniveau (Fritzsche et al., 2018; Miller & Geraci, 2011b), was het vertrouwen van de studenten gerelateerd aan hun kalibratieaccuratesse. Studenten die miskalibreerden—ongeacht of deze studenten een goede of slechte prestatie hadden op de examens—toonden zich bewust van de miskalibratie door hun lage vertrouwen in die schatting. De samenhang tussen kalibratieaccuratesse en subjectief vertrouwen werd gevonden bij de universiteitsstudenten, waarbij aangenomen wordt dat hun metacognitief bewustzijn een plateau heeft bereikt (Weil et al., 2013). Middelbare scholieren lieten echter geen dergelijke samenhang zien tussen hun kalibratieaccuratesse en vertrouwen. Er was geen relatie in de mate van kalibratieaccuratesse en hoeveel vertrouwen de scholieren hadden in hun schatting. Hoofdstuk 6 toonde echter aan dat er wel verandering mogelijk was in het vertrouwen dat middelbare scholieren hebben in hun schatting. Leerlingen die na het krijgen van feedback reflecteerden op hun kalibratieaccuratesse, verbeterden niet alleen deze kalibratieaccuratesse, maar kregen ook meer vertrouwen in hun prestatie-schatting.

Naast de bevindingen dat prestatieniveau en expertise van invloed zijn op de kalibratieaccuratesse en het metacognitief bewustzijn, dragen de studies in dit proefschrift bij aan het debat over of prestatieniveau wel of niet invloed heeft op hoe de kalibratieaccuratesse verbetert na het krijgen van feedback. Aan de ene kant werd aangenomen dat studenten met een lage taakprestatie wellicht minder profijt hebben van feedback omdat ze moeilijkheden kunnen ervaren de feedback goed te begrijpen (Stone, 2000). Aan de andere kant zouden deze studenten misschien juist meer profijt hebben van de feedback omdat ze geneigd zijn minder valide cues te gebruiken wanneer ze hun prestatie inschatten(Gutierrez de Blume et al., 2017b; Thiede et al., 2010). Ondanks dat prestatieniveau van invloed was op de kalibratieaccuratesse, laten de bevindingen in dit proefschrift zien dat studenten met zowel een hoge als lage taakprestatie, en veel of weinig expertise, hun kalibratie verbeteren na het krijgen van feedback (Hoofdstuk 2, 4 en 6). Daarnaast werd gevonden dat, ondersteunend voor de verklaring dat laagpresteerders meer profijt hebben van de feedback door hun neiging onjuiste aanwijzingen te gebruiken (Gutierrez de Blume et al., 2017; Thiede et al., 2010), de kalibratieaccuratesse van laagpresteerders meer vooruitgaat na het krijgen van feedback in Hoofdstuk 3.

De tweede onderzoeksvraag "Hangt de effectiviteit van prestatiefeedback om kalibratieaccuratesse te verbeteren af van het prestatieniveau van de student?" kan hierop als volgt beantwoord worden: over het algemeen niet. Studenten met een lage taak-prestatie of weinig expertise verbeteren hun kalibratieaccuratesse gelijk aan studenten met een hoge taakprestatie of veel expertise (Hoofdstuk 2, 4 en 6). Alleen in Hoofdstuk 3 werd gevonden dat studenten met een lage taakprestatie een sterkere verbetering zien in hun kalibratieaccuratesse.

Theoretische- en onderzoeksimplicaties

De studies in dit proefschrift laten zien dat mensen—van medisch specialisten tot middelbare scholieren—hun kalibratieaccuratesse kunnen verbeteren door het krijgen van prestatiefeedback. Het effect van prestatiefeedback is daarbij consistent, ondanks dat het effect getest is in verschillende settingen, met verschillende vormen prestatiefeedback, en met een variërende groep participanten. De studies laten eveneens zien dat de kalibratieaccuratesse van laag presterende studenten, of groepen met minder expertise even veel, of juist meer, verbeteren na het krijgen van de feedback. Dit betekent dat er sterkere hypotheses gemaakt kunnen worden over of prestatiefeedback effectief ingezet kan worden om de kalibratieaccuratesse te verbeteren op nieuwe taken. De resultaten van dit proefschrift zijn daarbij in overeenstemming met de veronderstelling dat het geven van feedback informatie geeft over de prestatie en daarmee ook de kalibratie, en dat studenten die aanwijzingen vervolgens effectief kunnen inzetten wanneer zij nieuwe prestatie-schattingen maken (Koriat, 1997; Thiede et al., 2010). Hoe meer aanwijzingen studenten hebben ontvangen over eerdere taken, hoe beter hun kalibratieaccuratesse wordt op nieuwe taken. Dit wordt aangetoond door de leercurve in kalibratieaccuratesse in zowel Hoofdstuk 3 als Hoofdstuk 6. De studies in dit proefschrift beantwoorden daarmee de open vraag in de literatuur of studenten kunnen leren beter te kalibreren door het krijgen van prestatiefeedback: dat kunnen ze inderdaad.

Een theoretische limitatie van de studies in dit proefschrift is dat deze voornamelijk procedureel inzicht bieden. Er treedt een verbetering op in de kalibratieaccuratesse nadat studenten feedback krijgen, en er wordt verwacht dat dit effect wordt veroorzaakt door het gebruik van betere aanwijzingen (Koriat, 1997; Thiede et al., 2010), maar de manier waarop studenten het gebruik van aanwijzingen veranderen blijft onduidelijk. Hebben de studenten geleerd hoe ze hun prestatie moeten inschatten doordat ze zich meer bewust zijn geworden van hoe ze hun antwoord moeten scoren? Of veranderden ze hun schattingen op basis van de kalibratieaccuratesse op vorige taken? ("Op de vorige taken heb ik mijzelf overschat, misschien is het beter als ik mijn verwachting nu iets naar beneden bijstel"). Of misschien hebben de studenten juist hun eerdere prestatie als anker gebruikt voor hun nieuwe schattingen? Het doel van het huidige proefschrift was om te testen of prestatiefeedback effectief ingezet kon worden om kalibratieaccuratesse te verbeteren. Nu blijkt dat dit inderdaad kan, onderstreept dit het belang van toekomstig onderzoek naar het mechanisme achter deze verbetering. Met name gezien de vele inconsistente bevindingen die tot nu toe in de literatuur zijn gevonden (Bol et al., 2005; Callender et al., 2015; Foster et al., 2017; Hacker, Bol, & Bahbahani, 2008; Hacker et al., 2000; Huff & Nietfeld, 2009; Miller & Geraci, 2011a; Nietfeld et al., 2006). Een manier om de aanwijzingen die studenten gebruiken tijdens het schatten van hun prestatie inzichtelijker te maken, is door hen te vragen hoe zij zijn gekomen tot hun prestatieschattingen (cf. Bol et al., 2005). In Hoofdstuk 6 zijn de studenten bijvoorbeeld gevraagd om de aanwijzingen die zij bij hun schatting hadden gebruikt te scoren. Ook kan studenten door middel van hardop-denk sessies gevraagd worden meer inzicht te geven in hoe zij hun prestatie schatten (Schraw, 2010).

Wanneer het mechanisme achter de verbeterde kalibratieaccuratesse verduidelijkt wordt, kan meer inzicht verworven worden in de retentie van het feedback-effect. In dit proefschrift kregen studenten prestatiefeedback na elke taak, en werd hun kalibratieaccuratesse gemeten op een direct daaropvolgende taak. Zouden de participanten in Hoofdstuk 2 tot 4 echter nog steeds een betere kalibratie laten zien als zij na een week terug zouden komen in het lab? Uit Hoofdstuk 6 blijkt dat er in ieder geval sprake is van retentie, omdat de examens waarop de studenten hun prestatie moesten schatten ruim een maand na elkaar worden aangeboden. Ondanks dit grote tijdsinterval, werd er een leercurve in de kalibratieaccuratesse gevonden. Om het bewijs voor retentie echter verder te versterken is meer onderzoek nodig.

Het is daarnaast eveneens onduidelijk of de effecten van kalibratieaccuratesse niet alleen binnen een taak, maar ook tussen taken zichtbaar kan worden. Spontane transfer van de ene naar de andere taak lijkt onwaarschijnlijk omdat standaarden en beoordelings-criteria verschillen tussen typen taken (Van Gog, Kostons, & Paas, 2010). Dit wordt onderschreven door Hoofdstuk 6. Hoewel leerlingen hun kalibratieaccuratesse verbeteren op het vak Frans, laten zij geen verbetering zien op het vak wiskunde, waar geen interventie plaatsvond. Opvallend is hierbij wel dat ook op het vak Duits, waarbij de beoordelingscriteria grotendeels overeenkomen met die van het vak Frans, de leerlingen eveneens geen verbetering lieten zien. Dit betekent dat het juist inschatten van de eigen prestatie taakspecifiek lijkt, in plaats van een algemene vaardigheid zoals in het verleden soms gesuggereerd (Van der Stel & Veenman, 2010). Er is echter meer onderzoek nodig om de verbetering in kalibratieaccuratesse tussen taken verder in kaart te brengen.

Een andere belangrijke aanbeveling voor toekomstig onderzoek gaat over de invloed van het type prestatieschattingen die gebruikt worden wanneer kalibratieaccuratesse wordt gemeten. Er zijn namelijk vele meetvormen zijn die gebruikt kunnen worden om metacognitieve vaardigheden in kaart te brengen. Zo kunnen studenten gevraagd worden om hun prestatie te schatten voor, tijdens, of na een taak (Nelson & Narens, 1990). De studies in dit proefschrift hebben zich bijna alle gefocust op het maken van een schatting na een taak (postdicties). Wanneer studenten hun prestatie schatten na, versus voor, een taak, profiteren ze van extra kennis over de inhoud van deze taak. Die kennis kan hen ervan weerhouden excuses te verzinnen over de inhoud van de taak wanneer zij feedback krijgen. Daar waar bij schattingen voor een taak (predicties) gezegd kan worden dat de inhoud van de toets simpelweg te veel afweek van de leermaterialen en dat daarom de vooropgestelde verwachting niet overeenkomt met het daadwerkelijk behaalde resultaat (Bol et al., 2005), houdt dit excuus minder stand in het geval van postdicties. Studenten wisten de inhoud van de taak, en als dit anders was dan zij verwacht hadden dan hadden ze dit mee moeten nemen in hun oordeel. Ter ondersteuning van de hypothese dat het effect van feedback kan afhangen van het type schatting dat wordt gevraagd, laat Hoofdstuk 2 zien dat daar waar postdicties verbeteren na de feedback, predicties van studenten met een gemiddelde prestatie juist slechter worden. Het is daarom van belang dat toekomstig onderzoek naar kalibratieaccuratesse duidelijk onderscheid maakt tussen verschillende prestatieschattingen en voorzichtig is met het generaliseren van de resultaten naar andere typen schattingen. Een onvoorzichtige generalisatie tussen de verschillende typen schattingen kan bijvoorbeeld bijdragen aan non-significante bevindingen in goed geciteerde artikelen als die van Sitzmann and Ely (2011). Wanneer gekeken wordt naar de effecten van monitoring is het dus essentieel dat toekomstig onderzoek duidelijker onderscheid maakt tussen de verschillende typen schattingen.

Onderzoek naar kalibratieaccuratesse heeft als uiteindelijke doel om de prestatie van studenten te verbeteren. Er wordt verondersteld dat doordat studenten bewust worden van hun prestatie, zij beter in staat zijn beslissingen te nemen die hun leren ten goede komt

(Butler & Winne, 1995; Fernandez & Jamet, 2017; Koriat, 2012; Metcalfe, 2009; Nelson & Narens, 1990; Tuysuzoglu & Greene, 2015). De resultaten van Hoofdstuk 2 en 6 ondersteunen de hypothese dat een verbetering in kalibratieaccuratesse samenhangt met een verbetering in de taakprestatie. Studenten in beide hoofdstukken verbeterden niet alleen hun schattingen, maar ook hun prestatie op de geheugentaak (Hoofdstuk 2) en de examens (Hoofdstuk 6). Nu blijkt dat prestatiefeedback effectief ingezet kan worden om kalibratieaccuratesse te verbeteren, is echter meer onderzoek nodig naar hoe deze verbetering vervolgens doorspeelt in de taakprestatie, bijvoorbeeld doordat studenten betere beslissingen maken tijdens het leren of na het doen van een taak.

Praktische implicaties

De studies in dit proefschrift hebben belangrijke praktische implicaties. Zelfregulatie en aanpassingsvermogen aan nieuwe situaties of aan nieuwe taken wordt steeds belangrijker in het hedendaagse onderwijs (Trilling & Fadel, 2009; Wolters, 2010). Door een gemakkelijk in te voeren prestatiefeedbackinterventie te testen, geeft dit proefschrift richtlijnen hoe de monitoring van studenten in verschillende situaties verbeterd kan worden. De effecten van prestatiefeedback houden daarbij stand voor studenten van verschillende prestatieniveaus en expertisegroepen. Zowel op de universiteit als op de middelbare school is de interventie effectief, en zowel studenten als experts kunnen ervan profiteren. Het huidige onderzoek onderstreept daarmee het belang om (1) studenten aan te sporen prestatieschattingen te maken, en (2) hen de mogelijkheid te geven om inzicht te krijgen in hun daadwerkelijke prestatie. De praktische waarde van deze twee activiteiten ligt in het feit dat ze makkelijk toe te passen zijn in zowel een onderwijs- als organisatiecontext, en dat zowel mensen die goed als minder goed presteren er profijt van hebben. Ten slotte vindt een verbetering al na korte tijd plaats: na drie feedbackmomenten laten medisch specialisten en geneeskunde-studenten al een betere kalibratieaccuratesse zien en er wordt al een leercurve gevonden over zes teksten in Hoofdstuk 3. Studenten kunnen dus op een makkelijke en snelle manier geholpen worden hun kalibratieaccuratesse te verbeteren door hun prestatie te schatten en deze schatting te vergelijken met de uiteindelijke uitkomst of oplossing.

Conclusie

Het hoofddoel van dit proefschrift was om meer inzicht te krijgen of het mogelijk is kalibratieaccuratesse te verbeteren door studenten prestatiefeedback te geven. De studies bevestigen uniform de hypothese dat prestatiefeedback effectief ingezet kan worden, zowel voor studenten die goed presteren als studenten die minder goed presteren. Deze bevindingen geven een eerste aanzet voor verder onderzoek naar hoe precies die verbetering tot stand komt. Daarnaast dragen de bevindingen bij aan de onderwijspraktijk. Leraren zouden hun studenten en leerlingen kunnen adviseren hun prestatie in te schatten na het doen van een taak of toets en hen vervolgens aan te sporen aandacht te geven aan prestatiefeedback, ofwel deze feedback zelf te genereren door hun antwoorden te vergelijken met standaarden. Hierdoor worden studenten zich meer bewust van zowel hun prestatie als kalibratie, en dit helpt hen om een betere inschattingen te maken van hun prestatie op nieuwe taken, ook wanneer feedback daar niet meer beschikbaar is.

Curriculum Vitae and Publications

Curriculum Vitae

Marloes Nederhand was born in Rotterdam, The Netherlands, in June 1992. After high school, she studied Psychology at Erasmus University Rotterdam. As a student, she participated in a research traineeship and in the multidisciplinary Erasmus Honours Program, and successfully obtained her Bachelor's degree in Organizational Psychology in 2013. Subsequently, she started her Master Industrial and Organizational Psychology, also at Erasmus University Rotterdam. During her Master internship, Marloes was involved in the development and execution of trainings provided to medical specialists working at the Erasmus Medical Center, following her interest in life-long learning. In 2014, she obtained her Master's degree (cum laude) and started working as a PhD-Candidate at the Institute of Psychology, Education and Child Studies at Erasmus University Rotterdam. Her project was funded by a Research Excellence Initiative grant and focused on how people can learn to better estimate their own performance. As a PhD-Candidate, Marloes presented her work at various international and national conferences, taught several courses, and supervised multiple research projects of both Bachelor and Master students. She has occasionally served as an ad-hoc reviewer for a number of journals, and her work has won multiple prizes, including the "Best paper award" from the Interuniversity Center for Educational Research (ICO) in 2017 and the "Graduate School Award for PhD Excellence" by the Erasmus Graduate School of Social Science and the Humanities in both 2015 and 2017. Currently, Marloes is employed as a postdoctoral researcher at the Institute of Psychology, Education and Child Studies at the Erasmus University Rotterdam. She also serves as a member of the supervisory board of the Erasmus Honours Program.

Publications

Nederhand, M.L., Tabbers, H.K., Homaira, A., & Rikers, R.M.J.P. (2018). Improving calibration over texts by providing standards both with and without idea-units. *Journal of Cognitive Psychology*. doi:10.1080/20445911.2018.1513005

Nederhand, M.L., Tabbers, H.K., Splinter, T.A.W., & Rikers, R.M.J.P. (2018). The effect of performance standards and medical experience on diagnostic calibration accuracy. *Health Professions Education*. doi:10.1016/j.hpe.2017.12.008

Submitted manuscripts

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (Under revision). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels.

Nederhand, M.L., Tabbers, H.K., De Bruin, A.B., & Rikers, R.M.J.P. (Submitted). Metacognitive awareness as measured by second-order judgements among university and secondary school students.

Nederhand, M.L., Tabbers, H.K., Jongerling, J., & Rikers, R.M.J.P. (Submitted). Outcome feedback and reflection to improve calibration of secondary school students: A longitudinal study.

Presentations

Nederhand, M.L. (2018). Improving calibration accuracy with performance feedback: An overview. Research Meeting Utrecht University.

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (2017). Providing Standards both with and without Idea-Units to improve Calibration Accuracy. Part of symposium “How am I doing? Improving students’ self-assessments”. 17th Biennial EARLI Conference for Research on Learning and Instruction: Tampere: Finland (2017, August 27 – 2017 September 2).

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (2017). Outcome Feedback and Reflection to Improve Calibration of High School Students: A Longitudinal Study. 17th Biennial EARLI Conference for Research on Learning and Instruction: Tampere: Finland (2017, August 27 – 2017 September 2). Graduate School Award for PhD Excellence, Erasmus Graduate School of Social Science and the Humanities.

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (2017). Improving calibration accuracy by providing standards: high and low performers do not benefit equally. American Educational Research Association: San Antonio, Texas: USA (2017, April 27 – 2017, May 1).

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (2017). Improving Calibration Accuracy on New Tasks for Different Performance Levels. National Spring School, ICO: Utrecht University, Utrecht: The Netherlands (2017, April 20 - 2017, April 21). Best paper award, Interuniversity Center for Educational Research (ICO).

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (2016). Students living up to their standards? Effects of external standards and performance level on calibration accuracy. Special Interest Group of EARLI: Metacognition: Nijmegen University, Nijmegen: The Netherlands (2016, August 23 - 2016, August 26).

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (2016). Leren weten wat je weet: Het effect van externe standaarden en prestatieniveau op kalibratie-accuratesse. Onderzoeks Research Dagen 2016: Erasmus Universiteit Rotterdam, Rotterdam: The Netherlands (2016, May 25 - 2016, May 27).

Nederhand, M.L., Tabbers, H.K., & Rikers, R.M.J.P. (2016). Diagnostic feedback to improve calibration accuracy of medical specialists and students. American Educational Research Association: Washington DC: USA (2016, April 08 - 2016, April 12).

Nederhand, M.L., Tabbers, H.K., Splinter, T.A.W. & Rikers, R.M.J.P. (2015). Improving calibration accuracy of medical specialists and medical students. 16th Biennial EARLI Conference for Research on Learning and Instruction: Cyprus (2015, August 25 - 2015, August 29). Graduate School Award for PhD Excellence, Erasmus Graduate School of Social Science and the Humanities.

Nederhand, M.L., Tabbers, H.K., Splinter, T.A.W. & Rikers, R.M.J.P. (2015). Diagnostic feedback to improve calibration accuracy of medical specialists and students. ICO National Fall School: Utrecht University, Utrecht: The Netherlands (2015, November 05 - 2015, November 06).

Nederhand, M.L., Tabbers, H.K., Splinter, T.A.W. & Rikers, R.M.J.P. (2014). The Effects of Feedback and Expertise on Diagnostic Calibration Accuracy. Graduate Research Day: Erasmus University Rotterdam, Rotterdam: The Netherlands (2014, October 01).

Dankwoord

Graag spreek ik mijn dank uit aan verschillende personen. Allereerst Remy en Huib. Heel hartelijk dank voor het vertrouwen, jullie kritische blik en humor. Ik heb de afgelopen jaren veel geleerd. Remy, ondanks dat je Rotterdam hebt verlaten kon ik je altijd bereiken en voelde ik me hartelijk welkom in Middelburg, dankjewel daarvoor. Huib, dank voor de fijne dagelijkse begeleiding en je onbegrensde vrolijkheid en relativeringsvermogen.

Graag bedank ik ook de leden van de promotiecommissie, Guus, Katharina, Sílvia, Peter, Fred en Henk. Fred en Guus, dank voor jullie geloof in mij, ik heb dat zeer gewaardeerd. Katharina, thank you very much for traveling to Holland for the ceremony, I sincerely appreciate that you are part of the committee. Sílvia, you made me feel very welcome in iMERR during the medical study. I am happy that you take part in the committee. Peter, jouw bevlogenheid heeft mijn interesse in statistiek als student aangewakkerd. Dankjewel dat je bij mijn promotie aanwezig bent. Henk, ik vind het heel leuk dat onze paden na ons interview tijdens mijn minor elkaar nog zo vaak hebben gekruist. Dank dat je er ook bij de promotie als commissielid bij bent.

Ook noem ik graag Tamara, dankjewel voor je betrokkenheid. Heel fijn dat ik ondanks je drukke schema met je kon sparren. Anique, dankjewel voor het meedenken en jouw introductie bij John. Joran, dank dat je mij hebt geholpen met de statistiek van de veldstudie. Christiaan en Marcel, veel dank voor de hulp bij de labstudies en de excel-macro's. Ook bedank ik graag Awee – voor je betrokkenheid en mogelijkheden die je mij naast mijn PhD bood binnen de honours begeleidingscommissie. Jacqueline, dank dat ik na mijn PhD meteen bij jou als postdoc aan de slag kon.

Ik bedank graag alle studenten, specialisten, docenten, ouders en leerlingen die hebben meegedaan of meegeholpen aan de onderzoeken. Lennart Rem en Liesbeth Dirksen, hartelijk dank voor de ondersteuning van en betrokkenheid bij de artsenstudie. Floris Yperlaan en Peter van Wijk dank ik graag voor de mogelijkheid om onderzoek te doen op SCALA. Jullie betrokkenheid en interesse hebben mij geïnspireerd en gemotiveerd. Speciale dank aan Wilma Lambregts. Mede door jouw enthousiasme en meedenken is de veldstudie een succes geworden.

Collega's binnen het instituut, bij ICO en binnen de pubgroup, bedankt voor de fijne samenwerking, jullie kritische blik en de gezelligheid op conferenties. Homaira, veel dank voor de hulp bij de data-invoer en de labstudie.

Familie en vrienden, opa en oma, hartelijk dank voor de betrokkenheid en gezelligheid. Speciaal noem ik Annet, Marjolein, Pilar en Renata. Vincent and Hsin-ya, thank you for being our friends and allowing me to cuddle with Isabella.

Graag bedank ik ook mijn twee – bij toeval zwarte-band houdende – paranimfen. Jason, het was een feest jouw kamergenoot te zijn. De statistiek- en pizzasessies zal ik nooit vergeten. Je bent een vriend voor het leven. José, bedankt voor alle discussies en vooral ook je humor. Jouw doorzettingsvermogen vormt een bron van inspiratie voor mij. Ik ben heel blij een helft te zijn van 'de zusjes'.

Pap en mam, ik prijs me gelukkig met jullie als ouders. Dankjulliewel voor de steun, het vertrouwen en jullie geloof in mij. Titiaan, dankjewel dat je er altijd voor mij bent, vaak bij mijn conferentiepresentaties aanwezig was en de tijd hebt genomen om mijn stukken te voorzien van feedback. Door onze vele gedachtewisselingen ben ik inmiddels ook goed op de hoogte van alle ins en outs over winst- en stemrechtloze aandelen. We zijn een goed team.

ICO dissertation series

329. Kock, Z.D.Q.P. (23-06-2016). *Toward physics education in agreement with the nature of science: Grade 9 electricity as a case.* Eindhoven: Eindhoven University of Technology.
330. Trinh Ba, T. (28-6-2016) *Development of a course on integrating ICT into inquiry-based science education.* Amsterdam: Vrije Universiteit Amsterdam.
331. Gerken, M. (29-06-2016). *How do employees learn at work? Understanding informal learning from others in different workplaces.* Maastricht: Maastricht University.
332. Louws, M.L. (06-07-2016) *Professional learning: what teachers want to learn.* Leiden: Leiden University.
333. Geel, M.J.M. van, & Keuning T. (08-07-2016). *Implementation and Effects of a Schoolwide Data-Based Decision Making Intervention: A Large-Scale Study.* Enschede: University of Twente.
334. Bouwer, I.R., & Koster, M.P. (02-09-2016) *Bringing writing research into the classroom: The effectiveness of Tekster, a newly developed writing program for elementary students.* Utrecht: Utrecht University.
335. Reijners, P.B.G. (02-09-2016.) *Retrieval as a Cognitive and Metacognitive Study Technique to Learn from Expository Text.* Heerlen: Open University of the Netherlands.
336. Hubers, M.D. (08-09-2016). *Capacity building by data team members to sustain schools' data use.* Enschede: University of Twente.
337. Hsiao, Y.P. (23-09-2016). *Peer Support to Facilitate Knowledge Sharing on Complex Tasks.* Heerlen: Open University of the Netherlands.
338. Scheer, E.A. (23-09-2016). *Data-based decision making put to the test.* Enschede: University of Twente.
339. Bohle Carbonell, K. (28-9-2016). *May I ask you...? The influence of Individual, Dyadic, and Network Factors on the Emergence of Information in Exchange Teams.* Maastricht: Maastricht University.
340. Claessens, L.C.A. (30-09-2016). *Be on my side, I'll be on your side: Teachers' perceptions of teacher– student relationships.* Utrecht: Utrecht university.
341. Jansen in de Wal, J. (18-11-2016). *Secondary school teachers' motivation for professional learning.* Heerlen: Open University of the Netherlands.
342. Kock, W.D. de. (24-11-2016). *The effectiveness of hints during computer supported word problem solving.* Groningen: University of Groningen.
343. Oonk, C. (07-12-2016). *Learning and Teaching in the Regional Learning Environment: Enabling Students and Teachers to Cross Boundaries in Multi-Stakeholder Practices'.* Wageningen: Wageningen University.
344. Beckers, J. (09-12-2016). *With a little help from my e-portfolio; supporting students' self directed learning in senior vocational education.* Maastricht: Maastricht University.
345. Osagie, E.R. (14-12-2016) *Learning and Corporate Social Responsibility. A study on the role of the learning organization, individual competencies, goal orientation and the learning climate in the CSR adaptation process.* Wageningen: Wageningen University.
346. Baggen, Y. (13-01-2017). *LLLIGHT 'in' Europe - Lifelong Learning, Innovation, Growth and Human Capital Tracks in Europe*. Wageningen: Wageningen University.
347. Wouters, A. (09-02-2017). *Effects of medical school selection. On the motivation of the student population and applicant pool.* Amsterdam: VU Medisch Centrum.

348. Baas, D.M. (01-05-2017). *Assessment for Learning: more than a tool*. Maastricht: Maastricht University.

349. Pennings, J.M. (04-05-2017). *Interpersonal dynamics in teacher-student interactions and relationships.* Utrecht: Utrecht University.

350. Lans, R.M. (18-05-2017). *Teacher evaluation through observation.* Groningen: University of Groningen.

351. Grohnert, T. (18-05-2017). *Judge/Fail/Learn; enabling auditors to make high-quality judgments by designing effective learning environments*. Maastricht: Maastricht University.

352. Brouwer, J. (22-05-2017). *Connecting, interacting and supporting. Social capital, peer network and cognitive perspectives on small group teaching*. Groningen: University of Groningen.

353. Van Lankveld, T.A.M. (20-06-2017). *Strengthening medical teachers' professional identity. Understanding identity development and the role of teacher communities and teaching courses*. Amsterdam: Vrije Universiteit Amsterdam.

354. Janssen, N. (23-06-2017). *Supporting teachers' technology integration in lesson plans.* Enschede: University of Twente.

355. Tuithof, J.I.G.M. (23-06-2017). *The characteristics of Dutch experienced history teachers' PCK in the context of a curriculum innovation.* Utrecht: Utrecht University.

356. Van Waes, S. (23-06-2017). *The ties that teach: Teaching networks in higher education.* Antwerp: University of Antwerp*.*

363. *Evens, M. (30-06-2017). Pedagogical content knowledge of French as a foreign language: Unraveling its development.* Leuven: KU Leuven.

364. *Moses, I. (07-09-2017). Student-teachers' commitment to teaching.* Leiden: Leiden University.

365. Wansink, B.G.J. (15-09-2017). *Between fact and interpretation. Teachers' beliefs and practices in interpretational history teaching.* Utrecht: Utrecht University.

367. Binkhorst, F. (20-10-2017). *Connecting the dots. Supporting the implementation of Teacher Design Teams*. Enschede: University of Twente.

368. Stoel, G.L. (14-11-2017). *Teaching towards historical expertise. Developing students' ability to reason causally in history*. Amsterdam: University of Amsterdam.

369. Van der Veen, M. (28-11-2017). *Dialogic classroom talk in early childhood education*. Amsterdam: Vrije Universiteit Amsterdam.

370. Frèrejean, J. (08-12-2017). *Instruction for information problem solving.* Heerlen: Open University of the Netherlands.

371. Rezende Da Cunha Junior, F. (19-12-2017). *Online groups in secondary education*. Amsterdam: Vrije Universiteit Amsterdam.

372. Van Dijk, A.M. (22-12-2017). *Learning together in mixed-ability elementary classrooms.* Enschede: University of Twente.