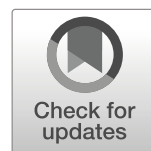



Metacognition and Learning
<https://doi.org/10.1007/s11409-020-09233-9>



Reflection on exam grades to improve calibration of secondary school students: a longitudinal study

Marloes L. Nederhand¹  • Huib K. Tabbers¹ • Joran Jongerling¹ • Remy M. J. P. Rikers^{1,2}

Received: 9 April 2019 / Accepted: 4 June 2020 / Published online: 11 June 2020
© The Author(s) 2020

Abstract

Grades provide students with information about their level of performance. However, grades may also make students more aware of how well they have estimated their performance, their so-called calibration accuracy. This longitudinal quasi-experimental study, set in secondary education, examined how to increase students' awareness of the accuracy of their grade estimates in order to improve their calibration accuracy. During an entire school year, students from year 1, 2, and 3 provided grade estimates after each of their French exams. Subsequently, when students received their grades, the level of reflection support on their earlier estimates was manipulated. The first group of students just received their grade, the second group had to calculate the difference between their estimate and the actual grade, and the third group also had to reflect on reasons for a possible mismatch. We expected that more reflection support would lead to more improvement in calibration accuracy. Results showed that providing grade estimates already improved calibration accuracy over the school year, regardless of level of reflection support. This finding shows that asking for grade estimates is an easy-to-implement way to improve calibration accuracy of students in secondary education.

Keywords Performance feedback · Outcome feedback · Reflection · Self-assessment · Calibration accuracy · Longitudinal design · Overconfidence · Self-regulated learning

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

¹ Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Woudestein, Mandeville Building, T13-54, P.O. Box 1738 3000DR Rotterdam, The Netherlands

² The Roosevelt Center for Excellence in Education, University College Roosevelt, Utrecht University, Utrecht, The Netherlands

Introduction

If asked to estimate their outcome after an exam, many students tend to provide estimates that are far off their actual outcome (Dunlosky and Rawson 2012; Kruger and Dunning 1999). In general, such misjudgment of performance in education can cause problems, because students may not recognize the need to change ineffective learning strategies, or fail to ask for help, leading to underachievement (Dent and Koenka 2016; Dunlosky and Rawson 2012; Nelson and Narens 1990). Especially since students are being asked to become increasingly in charge of their own learning at all levels of education (Trilling and Fadel 2009; Wolters 2010), scholars have urged for more understanding of how to improve calibration accuracy in education (De Bruin and Van Gog 2012; Panadero, Brown and Strijbos 2016). Calibration accuracy is defined as the match between estimated and actual performance (Lichtenstein and 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.

One essential step in helping students to become aware of their miscalibration is to provide performance feedback (Butler and Winne 1995; Finn and Tauber 2015; Stone 2000; Zimmerman 2000). In education, the most common form of performance feedback is grades. From their secondary school years onwards, students take many exams, and hence, receive lots of grades. Potentially, this type of performance feedback allows students to detect their miscalibration, and thus to make better estimates on their next exams. However, recent studies have shown that just providing grades fails to improve calibration accuracy, even after many feedback moments (e.g., Foster, Was, Dunlosky and Isaacson 2017). These findings raise the question whether providing grades is simply ineffective in improving calibration accuracy, or whether students perhaps need more guidance to adequately use and interpret their grades (e.g., Black and Wiliam 2009; Wiliam 2016).

Moreover, studies on how to improve calibration accuracy using performance feedback in the form of grades have predominantly been conducted in university contexts (e.g., Bol Hacker, O'Shea and Allen 2005; Fritzsche et al. 2018; Hacker Bol, Horgan and Rakow 2000; Händel and Fritzsche 2016; Miller and Geraci 2011a; Nietfeld, Cao and Osborne 2006). Calibration accuracy has been examined with younger children as well, even from the age of 6 onwards, showing that children are able to accurately judge their performance on simple tasks (Destan et al., 2017; Finn and Metcalfe, 2008; Labuhn, Zimmerman, and Hasselhorn 2010; Steiner et al. 2020). However, for more complicated tasks and metacognitive skills, research shows an age-trend in metacognitive awareness (De Bruin et al. 2011; Lyons and Ghetti 2011; Roebbers 2017, see also Schneider and Löffler, 2016). In such situations, younger students may stand to gain more from improved calibration accuracy than older ones (Roebbers 2014; Weil et al. 2013). Supporting students at secondary school to think about their performance may thus set the stage for more accurate calibration for the rest of their academic career.

In The Netherlands, but also in many other countries, the first time students are confronted with graded tests at a large scale, is at the beginning of secondary school. Hence, the main aim of the current study is to find out whether and how students in secondary school can be supported to optimally use grades to improve their calibration accuracy.

Measuring calibration accuracy and metacognitive awareness

When investigating how to improve students' performance estimates, one metric that is used frequently is calibration accuracy. Calibration accuracy indicates how well performance estimates match actual performance by calculating their absolute difference (Lichtenstein and Fischhoff, 1977). Although the metric of calibration accuracy provides insight into the

level of mismatch between estimated and actual performance, it does not consider the direction of mismatch. The bias index is used to measure whether students are overconfident or underconfident (Schraw, 2009). The only difference between the bias index and absolute calibration accuracy is that bias score is the non-absolute difference between estimated and actual performance. Therefore, it is possible that negative and positive bias scores cancel each other out when mean bias scores are calculated. Consequently, students may on average show zero bias, but still have poor calibration accuracy. More recently, a third metacognitive measure has been introduced for investigating metacognitive awareness: second-order judgements (SOJ). When students are required to provide an SOJ, they have to rate how confident they are about their performance estimate. A positive association between calibration accuracy and SOJs would be an indication that students are at least aware of the quality of their calibration. Already in 2005, Dunlosky et al. (2005) highlighted the value of SOJs as these can give a more fine-grained image of students' metacognitive awareness. However, it has been relatively recent that SOJs are being measured in metacognitive research (Händel and Fritzsche 2013, 2016; Miller and Geraci, 2011b).

The metrics described above can be examined on a variety of estimates. Studies in classroom settings often ask students to provide a global judgement. When making such an estimate, students indicate how they think they performed on an entire task. A global estimate is made, for example, when students estimate their grade or number of total points obtained on a task or exam (e.g., Bol and Hacker 2001; Callender et al., 2016; Foster et al., 2017; Hacker et al. 2000; Labuhn et al.; Miller and Geraci, 2011b). In contrast to global estimations, it is also possible to ask for local estimates, in which students indicate how well they performed on each individual item. This method is more often used in laboratory research (e.g., Nederhand et al., 2018a, 2019), but has also been applied in educational settings (e.g., Nietfeld et al. 2006).

In addition to asking students local or global estimates, students can be asked to make either predictions or postdictions. The advantage of asking the students to estimate their performance directly after each task or exam ('postdictions', as in Hacker et al. 2000; Nietfeld et al. 2006), rather than before the exam ('predictions', as in Bol et al. 2005; Foster et al. 2017) is that students could incorporate knowledge about test items and the nature of the testing in their estimate. With predictions, students can easily attribute miscalibration to the test rather than to themselves. For example, a well-known reflection by students on inaccuracies in their predictions is that there were too many test items on the exam that did not represent the learning materials (Bol et al. 2005).

The effectiveness of performance feedback to improve calibration accuracy

Koriat (1997) argued that when estimating performance, students use cues such as how much information they can recall (Baker and Dunlosky 2006) or how familiar the test items are to them (Metcalf and Finn 2012). Calibration accuracy thus very much depends on the validity of these cues. Instead of using valid cues, however, students often use unreliable and/or irrelevant cues when estimating their own performance (Baker and Dunlosky 2006; Gutierrez de Blume et al. 2017; Serra and DeMarree 2016; Thiede Griffin, Wiley and Anderson 2010), limiting their ability to improve calibration accuracy over time. This issue was clearly demonstrated by Foster et al. (2017), who asked university students to predict their performance multiple times during an educational psychology course. Throughout the course, students made predictions on thirteen exams and received feedback immediately after each exam. However, the students did not show any improvement in calibration accuracy over time.

Foster and colleagues noted that students' actual prior performance would have been a valid predictor of future performance. In other words, calibration accuracy may have improved if students had used the outcome feedback about their level of performance on the previous exams as their cue when estimating performance on a subsequent exam (see also Finn and Metcalfe 2008).

In general, grades are not the most effective kind of feedback when the aim is to directly improve performance (Hattie and Timperley 2007; Nicol and Macfarlane-Dick 2006). Grades do not inform students about how well they performed on each individual question or task, what the correct answers should have been, and what students could do to improve their performance on new tasks. At the same time, grades do allow students to become more aware of potential miscalibration, by providing them the opportunity to compare their initial estimate to the actual outcome (Butler and Winne 1995; Labuhn et al. 2010; Nederhand et al., 2018b, 2019; Zimmerman 1990). Thus, although grades may be rather ineffective in improving academic performance at the short term, grades can improve students' monitoring of their own learning, which may improve their self-regulation and thus their performance on the longer term.

Because of the practical advantage of using already available grades in education to improve calibration accuracy, it has been investigated in several studies (Bol and Hacker 2001; Foster et al. 2017; Hacker, Bol and Bahbahani 2008; Hacker et al. 2000; Nietfeld et al. 2006). These studies found that providing grades in itself was insufficient to improve calibration accuracy (Bol et al. 2005; Brown, Andrade and Chen 2015; Foster et al. 2017; Huff and Nietfeld 2009). In general, only when students were asked to actively reflect on the (mis)match between their grade estimate and their actual grade, improvements in calibration accuracy were found (Brown et al. 2015; Hacker et al. 2000; Miller and Geraci 2011a; Nietfeld et al. 2006).

However, it is yet unclear how students should be encouraged to reflect on their calibration accuracy. Previous studies used different kinds of interventions, and manipulated different variables at the same time. For example, Hacker et al. (2000), Huff and Nietfeld (2009), and Nietfeld et al. (2006) provided their students with a training in monitoring skills, in which 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, Callender, Franco-Watkins, and Roberts (2016) and Miller and Geraci (2011a) only asked students to reflect on the mismatch between estimated performance and actual performance, but also provided students with rewards for accurate calibration. Hence, the question remains whether improvements in calibration accuracy were due to the reflection prompts, to the rewards, or even simply to the practice of estimating one's own performance. Furthermore, most studies have been conducted in the context of an educational psychology course (Foster et al. 2017; Nietfeld et al. 2006), and during some of these courses, students were even taught about the risks of overconfidence (e.g., Callender et al. 2016; Hacker et al. 2000; Miller and Geraci 2011a). Consequently, based on these studies, it is hard to predict whether and how grades can be effectively used to improve calibration accuracy in secondary education.

Improving calibration with grades by increasing support

The question how students can best be supported to improve their calibration accuracy after receiving grades thus requires a more systematic approach. It may be fruitful to distinguish

between the different activities students should do when reflecting on their grades, in order to improve their calibration accuracy. First, students need to actually provide a grade estimate before receiving the actual grade. Second, students should check whether there is a mismatch between their initial estimate and their actual grade. Third, when students recognize that their estimate is inaccurate, students should reflect on the cause of their miscalibration, in order to be able to improve their performance estimates on future tasks (De Bruin et al., 2017; Hacker et al. 2000; Nietfeld et al. 2006). An intervention to improve calibration accuracy using grades may thus provide support on all three activities.

The first activity crucial to calibration accuracy, is that students actually estimate their performance. Interventions that successfully enhanced calibration accuracy after providing grades or outcome scores always included practice with making performance estimates (Hacker et al. 2000; Miller and Geraci 2011a; Nietfeld et al. 2006). Sadler (1989) stated that, to improve monitoring, students need ‘evaluative experience’, and Boud, Lawson, and Thompson (2013) argued that students need to engage in monitoring practice, like any other expertise would require (e.g., Ericsson, Krampe and Tesch-Römer 1993). Consequently, it is possible that practising with performance estimates may, in itself, already improve calibration accuracy. However, the two studies that looked at the independent influence of practice with estimating performance before each exam did not find any benefits (Bol et al. 2005; Foster et al. 2017). Note though that these two studies were conducted among university students. Younger students, like in secondary education, may benefit more from practice, because in general, they are less metacognitively aware than adults (Paulus, Tsalas, Proust and Sodian 2014; Weil et al. 2013). Nevertheless, although practising with estimating one’s own performance seems to be required, it does not seem to be sufficient to improve calibration accuracy.

Second, for grades to be effective at improving calibration, the grade 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 (mis)match between their estimated and actual grades or outcomes (Callender et al. 2016; Hacker et al. 2008; Miller and Geraci 2011a; Nietfeld et al. 2006), or have directly provided students with comparison scores (Miller and Geraci 2011a; Nietfeld et al. 2006; Stone and Opel 2000). However, findings have been mixed. For example, Callender et al. (2016) found that students who postdicted their performance and 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 et al. (2008) included comparison scores in their interventions, but found no improvement in students’ calibration accuracy. Given these mixed findings and a lack of research among secondary school students, the question remains whether helping students focus on the mismatch between their estimate and their grade would be an effective way to improve their calibration accuracy.

Third, when students signal a mismatch between their estimated and actual performance, it is important that they 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 change their cues and improve their estimates (Nelson and Narens 1990; Zimmerman 2000). Given their potential, reflection or self-explanation prompts have been included in several interventions (De Bruin et al. 2017; Hacker et al. 2000; Huff and 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 on 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. This means that whereas reflection was beneficial to improve calibration accuracy (De Bruin et al. 2017; Hacker et al. 2000; Huff and Nietfeld 2009; Nietfeld et al. 2006), its effects were not straightforward, were dependent on performance levels, and again, exclusively examined in university student samples. Hence, the question remains whether providing reflection prompts could help secondary school students to improve their calibration accuracy, and also whether performance level may play a role.

Present study

With the current study, we wanted to examine how students in secondary education could effectively be supported in reflecting on their grades to improve their calibration accuracy. During an entire school year, we systematically varied the level of reflection support when students received their actual grades. Our first group of students did not receive any support (Practice-only group). The second group was asked to compare their initial grade estimate to the actual grade (Grade comparison group). The third group was prompted to reflect on the cues used for their estimate and on reasons for a mismatch whenever the estimate differed from their actual grade (Reflection group).

We studied effects of the level of reflection support when students received their actual grades in their course French. During an entire school year, we systematically varied the foreign languages English, French, and German are all mandatory part of the curriculum in the lower grades of Dutch secondary education. The French language is the only Romance language, most dissimilar to Dutch, and other than English, not very often encountered outside the school context in the Netherlands. Furthermore, many studies on metacognitive awareness have been conducted with tasks like word learning, text comprehension and definition recall (e.g., Dunlosky et al. 2018, Nederhand et al., 2018a, b), and the French exams within our study consisted of very similar tasks. Another advantage was that the scoring criteria for the French exams were very objective, like the number of spelling mistakes, or correctness of vocabulary. History classes for instance have exams with assessment criteria that can be more subjective, like essay writing. These points made French a very suitable domain subject for investigating our hypotheses.

Hypotheses

Based on prior research with university students, we formulated several hypotheses. First, we expected to find a relation between the level of support and the improvement of absolute calibration accuracy over time. Based on the limited benefits of estimation practice found in the literature (Bol et al. 2005; Foster et al. 2017), we expected students in the Practice-only group to show the least improvement in calibration accuracy, students in the Grade comparison group to show bigger improvement, and students in the Reflection group to show the strongest improvement. Second, we hypothesized similar effects for improvements in bias. Third, following the reasoning that better calibration accuracy is related to increased metacognitive awareness, we expected more support to also lead to higher SOJs over time.

Besides our main hypotheses, we additionally explored the moderating effect of performance level. Given that performance level may affect calibration accuracy (Carpenter et al. 2016; Kruger and Dunning 1999; Panadero et al. 2016) and its development over time (e.g., Hacker et al. 2008, 2000; Nietfeld et al. 2006), especially in relation to reflection support, we checked whether students' performance level moderated the effect from our interventions. Furthermore, we looked at the improvement in calibration accuracy of the students enrolled in our intervention for two other courses, in which no intervention was implemented. Doing so enabled us to check whether improvements in students' calibration accuracy in the French course could also be due to maturation. Finally, we kept track of student performance, to see if and how our intervention indirectly affected exam performance as well.

Method

Participants

The initial sample consisted of 261 students (49.8% female, 50.2% male) from 9 classes across first, second and third year (pre-university level) at a Dutch suburban secondary school.¹ The mean age of these students was 14 years (from 12 to 17 years old). Informed consent was requested from both students and their parents or caretakers. 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 of parents or caretakers, and their data were hence excluded. The group of students that abstained from participation did not differ from the group participating in the study in gender and age characteristics. The final sample of students consisted of 219 students (83.90% of original sample): 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). Table 1 shows the distribution of the students over grades and classes.

Design & Procedure

The design and procedure of the current study were approved by the institutional ethical research committee of the Erasmus University Rotterdam.

Our study was conducted in an actual classroom environment with minimal intrusion of the researchers. The first author introduced herself, briefed the students from each classroom that participated, and handed out consent forms. None of the researchers were present during the students' classes and exams. Instead, the teacher was provided with forms to hand out to the students. Before each data collection started, the main researcher checked for each class whether the correct forms were ready. After each exam, we collected the data from the teacher. Furthermore, we regularly met with the teacher to discuss if there were any irregularities.

Figure 1 depicts the set-up of the current study. All participants from the study had the same French teacher, and attended two French classes per week. During these classes, all participants within a grade year were taught exactly the same content. The school year was further divided into four periods, and during each period students took two French exams. The first exam of each period was administered by the teacher during a regular class meeting. All

¹ The students involved in this study were of upper secondary school to lower high school. Compared to the USA and UK system, the Dutch classes 1–3 match the classes 7 to 9 and 8 to 10 respectively.

Table 1 Distribution of students over grades, classes, and interventions

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

students within one class took the test at the same time, and the teacher aimed to test all classes 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. All classes were allocated to different classrooms and the classes within each year took their French exam at the exact same time. Over the school year, students took eight French exams, except for students in year 1, who took seven exams and only participated in the official testing weeks from the second period onwards. In all years, the intervention started during the first testing week in which students participated. Any teacher-administered exam before this testing week was excluded from our intervention, so students could get adjusted to the way of testing.

At the end of each French exam, students in all groups received an estimate form from the teacher to provide a grade estimate and a confidence score. After filling out the estimate form, students put the form in an envelope and sealed it to ensure confidentiality of their estimates. Students then handed both their exam and their sealed envelopes to their teacher. During the next class meeting, approximately one week later, the graded exams were returned to the students (grades were also presented online).

Within each grade year, classes were randomly assigned to one of the three intervention conditions. Students in the Practice condition did not receive any reflection support, but just received their graded exams. The students in the Grade comparison condition received their graded exam together with their estimate form in the sealed envelope. After opening the envelope, they were asked to fill out a comparison form to compute the difference between their estimated grade and their actual grade. The students in the Reflection condition also received their estimate form with the graded exam, but were asked to fill out a reflection form that not only asked to compute a difference score, but also to reflect on their estimate. After they filled out their form, students from the Grade comparison condition and Reflection condition put the form in a new envelope, together with the estimate form, sealed it, and handed it to their teacher. In all classes, the meeting continued with a plenary discussion of how the teacher had graded the exam, as this was a required part of the French lessons.

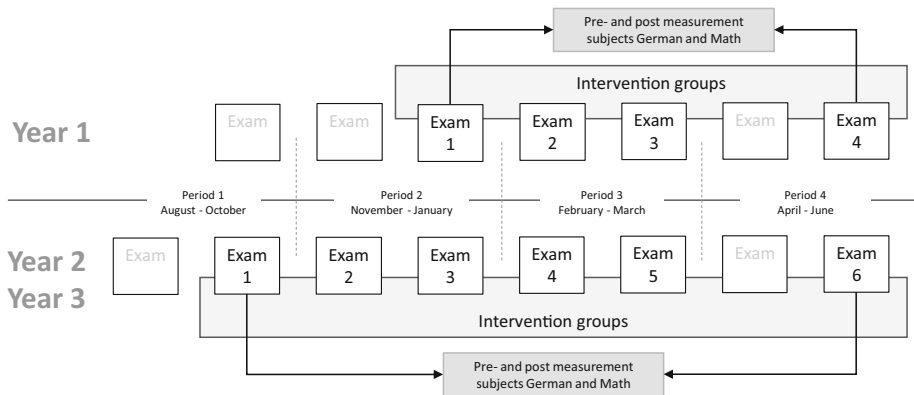


Fig. 1 The design over time of the current study. In Year 1, 2, and 3 of a secondary school, we implemented an intervention at several exams. The exams printed in grey were not included in the intervention

To check whether any improvement in calibration accuracy on the French exams could be attributed to the intervention itself, and not to a general maturation effect, we also measured calibration accuracy of the same students in our intervention in two other courses: German (i.e., a very similar course on foreign language) and Math (i.e., very different course from an unrelated domain). Estimate forms were administered for German and Math at the beginning (pre-test) and end (post-test) of the intervention period for students from all intervention groups (see Fig. 1).

After the final exam period of the year, we provided all students with an evaluation questionnaire, in which we asked them how they had experienced the study and whether they felt they had learned something about their calibration accuracy.

Materials

Exams

Each French exam consisted of questions on vocabulary, sentence construction, and grammar. Students had to translate single words, fill in blanks (e.g., “Chaque année, il y a un spectacle avec des _____ à la plage”), change the verb tense in sentences, and provide common expressions in French (e.g., “How can you say that you will arrive around five o’clock?”). Furthermore, each exam also included a test of either listening or reading comprehension, which required students to answer specific questions about a French written text or spoken fragment (e.g., “What was said about the History teacher?”), and also to compare the content of different texts or fragments (e.g., “Describe what text 1 and text 2 have in common”).

Each exam was scored by the teacher on a scale from 1 to 10. The assessment norms for each exam were based on an absolute instead of relative norm. Furthermore, the assessment norm was based on the guidelines laid down in the French educational method that was used in the secondary school. The scoring norms were not adjusted during the school year and were the same for all students within the same year. This means that there were strict rules about how many mistakes led to which grade. These rules were also clear to the students, as the number of points that could be obtained were outlined before each set of questions in the exam.

In each year, the first exam in period 4 (printed in grey in Fig. 1) did not contain a listening or reading comprehension part, and therefore deviated from the other exams. We excluded the data from this exam from our analyses and excluded it from the continuous numbering of the exams.

Forms

Estimate form On the estimate form, students first had to give an estimate of the grade they thought they would receive for the exam on a scale from 1 to 10 (one decimal allowed; 10 representing the highest score possible). Second, students had to provide an SOJ by rating the confidence they had in their estimate on a five-point scale (1 = very unconfident; 5 = very confident, cf. Händel and Fritzsche 2016; Miller and Geraci 2011b).

Grade comparison form On the grade comparison form, students first had to fill in their initial grade estimate. Second, they had to fill in the grade they actually obtained. The third and last step required students to calculate the difference between their estimate and actual grade.

Reflection form 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 had provided students with statements to choose from (Bol et al. 2005; Niefeld et al. 2006), because open reflection question would elicit “do not know” responses too easily (Krosnick and Presser 2010). At the same time, however, open questions allow for responses students would not provide otherwise. We therefore decided to use a combination of open and closed questions. First, to gain more insight into the cues used by when estimating their grade, we posed the question: ‘How did you come up with the estimate of the grade of your exam?’ All students could then choose one or more of eight possible reasons for miscalibration that we found in the literature (Bol and Hacker 2001; Bol et al. 2005). The statements were: (1) ‘The grades I normally get for this school course’; (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 felt during the exam’; (6) ‘How relaxed/stressed I felt 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 your answer below)’. Second, we wanted students whose estimate deviated from their actual grade to reflect on this mismatch. In a pilot study with students from three different Dutch secondary schools ($n = 407$), we found that the 25% best-calibrated students had a maximum deviation of less than 0.5 points. Therefore, all students who had a deviation score of at least 0.5 had to answer the following open question: ‘How do you explain the mismatch between your estimated grade and actual obtained grade?’ Finally, the reflection form 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. During these weekly support hours, the teacher provided some extra explanation about the course content. They were mandatory for students who had performed poorly on their exams, but other students were allowed to participate on a voluntarily basis.

Dependent variables and analyses

The actual grades on the exams were taken as exam performance score, ranging from 1 to 10. Calibration accuracy was defined as the absolute difference between students’ estimated grade

and their actual grade (both on a scale from 1 to 10), and could range from 0 to 9. Bias scores were calculated as the signed difference between estimated grade and actual grade (Dunlosky and Thiede 2013; Schraw 2009), and could range from -9 to $+9$. The confidence rating on the estimate form was used as the second-order judgement-score (SOJ), ranging from 1 to 5.

To answer our research questions and hypotheses, we examined the change in scores on our dependent variables over time. An alpha level of .05 was used for statistical tests. 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). Because our data set had a nested structure of (1) different measurements over time (2) for each student (3) within classes, we applied a multilevel analysis using HLM7 software. Before running any of the analyses, we checked whether the assumptions for multilevel analysis were met, which was the case. To test whether a multilevel analysis was necessary, we first ran intercept-only models for our outcome variables calibration accuracy, bias scores, SOJs, and exam performance. Doing so allowed us to determine whether the outcome variables had a statistically detectable amount of variance on the second (student) and third (class) level (Hox, 2010).

Results showed a nested structure on the Student Level for calibration accuracy, $\chi^2(207) = 419.06$, $p < .001$; bias scores, $\chi^2(207) = 688.58$, $p < .001$; SOJs, $\chi^2(207) = 739.87$, $p < .001$; and exam performance, $\chi^2(207) = 1924.27$, $p < .001$. No statistically detectable effect of the level Class was found for calibration accuracy, $\chi^2(8) = 12.71$, $p = .122$; SOJs, $\chi^2(8) = 12.56$, $p = .127$; and exam performance $\chi^2(8) = 14.27$, $p = .074$. However bias scores showed a statistically detectable effect on Class Level, $\chi^2(8) = 23.88$, $p = .003$, indicating that students in different classrooms differed more from each other in terms of their bias scores, than students from the same class. Note that the Class level we included, consisted of 9 classes: 1a, 1b, 1c; 2a, 2b, 2c; 3a, 3b, 3c. It is therefore possible that the differences between classes would be caused by differences between years. For example, the 3 classes from year 3 may behave differently from the 3 classes in year 1 and year 2. When testing the 3 level model for Bias, we therefore included the variable Year as a predictor.

In sum, the tests indicated that we should use a two-level multilevel design for analyzing our hypotheses on calibration accuracy, exam performance and SOJs, and a three-level multilevel design for analyzing our hypotheses on bias scores. Our main predictors were Time and Intervention group. Time was coded as the number of the exam students took. For example, the first exam, number 1, was the first time they received the intervention. The intervention groups were dummy coded against the Practice group, which served as a reference group in our analyses. Finally, the predictor Year was included in the model testing bias scores. Year was dummy coded as well, in which Year 1 served as a reference group for Year 2 and Year 3.

When finding an improvement in calibration accuracy, it is hopefully the result of students changing their estimates based on a better metacognitive awareness of their performance. However, an improvement in calibration accuracy may also be simply the result of changes in students' performance leading to a better fit with their estimates, without any change in their metacognitive awareness. In our study, whenever we found changes on our metacognitive variables—calibration accuracy, bias, and SOJs—we always checked whether this was accompanied by changes in estimates, and not just by changes in performance.

Results

Before running any of the analyses, we did a fidelity check to see whether students in each intervention group had done what they were supposed to do. In total, 99.8% of the students

had provided an estimate on each French exam. When looking at the Comparison and Reflection group, we saw that 99.9% of the students had correctly followed the instructions to calculate the difference between their performance and actual performance. Finally, our data from the Reflection group showed that all students had indicated which cues they had used when estimating their performance, and that 74.5% of the students whose estimate deviated more than 0.5 points had provided a written reflection on their mismatch. We do not know whether the remaining 25.5% who had an inaccurate estimate had actually reflected on this mismatch (at least they did not write it down). It should be noted that this was an open question, in contrast to the closed question about the cues, which may explain part of the lower response rate (e.g., Griffith et al. 1999).

In sum, our fidelity check showed that our intervention mostly worked as intended, except for a small subgroup of students in the reflection condition who did reflect on the cues used for their estimate, but not on reasons for a mismatch. We therefore continued with the analyses as planned, taking into account the students with missing reflections in our conclusions.

Preliminary analyses

Table 2 shows the descriptives of our main dependent variables on the first exam (pre-test) of French, as well as from German and Math. Before testing our main hypotheses, we first examined possible pre-test differences of the intervention groups.

First, exam performance (i.e., grades) on the French course was investigated to check whether the intervention groups' exam performance was comparable at the beginning of the intervention. Findings showed an overall mean of 6.68 (median = 6.60, $SD = 1.36$), which is largely sufficient (5.5 is the cut-off score for a sufficient grade). An analysis with exam performance on the pre-test French as dependent variable and intervention group as independent variable showed no statistically detectable differences $t's < 2$.

The average calibration accuracy of the students on the pre-test French was 0.97 (median = 0.80, $SD = 0.68$). This means that at the pre-test, students' estimates were on average about one point off their actual grade. An analysis with calibration accuracy on the pre-test French as dependent variable and intervention group as independent variable (Practice vs. Grade comparison vs. Reflection) showed no statistically detectable differences, all $t's < 2$.

When looking at the bias scores on the pre-test French, results showed that, in contrast to previous studies (e.g., Kruger and Dunning 1999), students from our sample did not show a specific tendency to be overconfident. Instead, the average bias over all groups was $M = -0.20$ (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 statistically detectable differences, all $t's < 2$.

Second-order judgements (SOJs) on the pre-test French 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 French as dependent variable and intervention group as independent variable showed a statistically detectable difference between the Reflection group ($M = 2.99$, $SD = 0.81$) and the other two groups $b = -0.35$, $t(193) = -2.47$, $p = .014$. The difference between the Practice group ($M = 3.32$, $SD = 0.83$), and Grade comparison group ($M = 3.27$, $SD = 0.74$) was not statistically detectable, $b = -0.06$, $t(193) = -0.41$, $p = .682$. This indicates that at the pre-test, students in the Reflection group generally provided lower confidence scores than students from the other two groups.

Table 2 Means and standard deviations on the dependent measures for the different intervention groups on the first exam (pretest), for French, as well as for German and Math

	Practice		Grade comparison		Reflection	
	<i>M</i> [min, max]	<i>SD</i>	<i>M</i> [min, max]	<i>SD</i>	<i>M</i> [min, max]	<i>SD</i>
French						
Performance	6.85 [3.50, 9.60]	1.33	6.68 [3.00, 9.00]	1.39	6.52 [3.30, 9.50]	1.37
Bias	-0.28 [-2.80, 2.40]	1.15	-0.04 [-2.40, 3.00]	1.13	-0.29 [-2.70, 1.90]	1.22
Calibration	0.93 [0.00, 2.80]	0.72	0.91 [0.00, 3.00]	0.66	1.05 [0.00, 2.70]	0.66
SOJ	3.32 [1.00, 5.00]	0.83	3.27 [1.00, 5.00]	0.74	2.97 [1.00, 5.00]	0.82
Math						
Performance	6.91 [3.40, 9.40]	1.37	6.77 [3.00, 9.40]	1.46	6.97 [3.00, 10.00]	1.55
Bias	-0.44 [-3.70, 2.70]	1.42	-0.13 [-5.10, 2.30]	1.35	-0.51 [-4.20, 3.00]	1.81
Calibration	1.19 [0.00, 3.70]	0.87	1.07 [0.00, 5.10]	0.83	1.51 [0.00, 4.20]	1.11
SOJ	3.30 [2.00, 5.00]	0.76	3.35 [1.00, 5.00]	1.03	3.25 [1.00, 5.00]	0.93
German						
Performance	6.38 [3.00, 8.90]	1.41	6.22 [3.00, 9.40]	1.28	6.17 [4.30, 9.20]	1.23
Bias	0.53 [-2.10, 3.10]	1.28	0.59 [-1.80, 3.00]	1.06	0.70 [-1.70, 3.20]	1.23
Calibration	1.13 [0.00, 3.10]	0.79	0.96 [0.00, 3.00]	0.74	1.17 [0.00, 3.20]	0.79
SOJ	3.32 [1.00, 5.00]	0.78	3.28 [1.00, 5.00]	0.94	3.23 [1.00, 4.00]	0.59

In sum, these results showed that at the start of the intervention, the intervention groups did not seem to differ from one another in calibration accuracy, bias, and exam performance. The only difference was found on SOJs: at the beginning of the intervention, students in the Reflection group were a bit less confident about their estimates than students in the Bias and Practice group.

Improvement of absolute calibration accuracy over time

Our first hypothesis was that the level of support in using the grades was related to the improvement of calibration accuracy in the course French over time. To examine calibration accuracy over time and among our intervention groups, we ran a two-level model in HLM7. Table 3 presents the results from our analysis with all variables and interactions included. Results overall showed a statistically detectable change in calibration accuracy over time $b = -0.03$, $t(215) = -2.95$, $p = .004$. Note that the negative regression coefficient indicates that, over time, calibration accuracy improved. We also examined whether the effect of time differed between students or groups. Results showed a non-statistically detectable random effect of time: $\chi^2(213) = 217.16$, $p = .408$. This means that the improvement did not differ for students from the different intervention groups. Consequently, when adding the intervention groups to the model, there were no statistically detectable interaction effects between any of the groups and time, all t 's < 2 (see Table 3). This means that our first hypothesis—the level 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 affected by a change in performance estimates, but also by a change in performance level without any change in estimates (i.e., students'

Table 3 Regression variables calibration accuracy, bias, SOJs, and performance

	Calibration accuracy		Bias		SOJ		Performance	
	<i>b</i>	<i>p</i>	<i>b</i>	<i>p</i>	<i>b</i>	<i>p</i>	<i>b</i>	<i>p</i>
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	-0.01	.829	0.05	.002
Level 2 (Student)								
Intervention groups*								
Grade comparison	0.12	.292	0.01	.962	-0.16	.194	-0.25	.281
Reflection	0.10	.410	0.02	.912	-0.50	<.001	-0.44	.069
Performance level	-0.07	.170	-0.39	< .001	0.01	.898	-	-
Interactions with Time								
Intervention groups*								
Grade comparison	-0.05	.054	0.04	.528	0.02	.596	<0.01	.914
Reflection	-0.03	.365	-0.03	.666	0.06	.035	0.03	.366
Performance level	0.01	.377	0.01	.563	-0.01	.740	-	-
Level 3 (Classes)								
Year*								
Year 2	-	-	-0.95	.002	-	-	-	-
Year 3	-	-	-1.64	<.001	-	-	-	-
Interactions with Time								
Year*								
Year 2	-	-	0.17	.031	-	-	-	-
Year 3	-	-	0.39	<.001	-	-	-	-
Interaction with Time and Intervention								
Grade comparison x Year 2	-	-	<0.01	.947	-	-	-	-
Grade comparison x Year 3	-	-	-0.07	.291	-	-	-	-
Reflection x Year 2	-	-	-0.11	.121	-	-	-	-
Reflection x Year 3	-	-	-0.05	.437	-	-	-	-

*The intervention groups and years students could be in were both dummy-coded. For the intervention groups, the Practice condition served as a reference group. For the different Years students could be in, Year 1 served as a reference group

grades become more in line with their estimate). To check whether the improved calibration accuracy was caused by adjustments of performance estimates (as hypothesized), rather than by changes in performance only, we conducted a follow-up analysis. We used the same multilevel analysis, but this time with grade estimate as the dependent variable, instead of calibration accuracy. Results showed that there was still a statistically detectable effect, $b = .05$, $t(215) = 2.89$, $p = .004$, indicating that students' grade estimates changed over time. Hence, the improvement in calibration accuracy was accompanied by a change in performance estimates, and not just by a change in exam performance.

Performance level differences

We also examined the effect of performance level on calibration accuracy and the possible improvement over time because prior studies found that performance level could moderate this improvement (e.g., Hacker et al. 2000; Nietfeld et al. 2006). First, 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 calibration accuracy. Second, as already indicated by the non-statistically detectable random effect of

calibration accuracy over time, performance level did not influence the improvement of calibration over time, $b = 0.01$, $t(212) = 0.89$, $p = .377$.

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

Finally, to exclude a maturation effect, we tested whether students from our intervention groups had also improved in calibration accuracy on two courses that were not included in our intervention: German (similar in instructional approach) and Math (different in instructional approach). For these courses, we only had two calibration accuracy scores (before and after our intervention period), so we analyzed the differences between pre-test and post-test. The intercept-only models showed no level 3 (Class-level) variance for Math and German. Consequently, we ran 2-level models. Results showed that calibration accuracy on the German and Math exams did not change statistically detectable from pre-test to post-test (pretest: $M_{Ger} = 1.15$, $SD_{Ger} = 0.77$; $M_{Math} = 1.25$, $SD_{Math} = 0.96$; post-test: $M_{Ger} = 1.03$, $SD_{Ger} = 0.80$; $M_{Math} = 1.19$, $SD_{Math} = 0.89$), all t 's < 2 . This indicates that the improvement in calibration accuracy we found on French exams, was not paired with a similar improvement in calibration accuracy on German and Math exams. Furthermore, our results showed that the intervention group students were in did not impact these results, all t 's < 2 .

Improvement of bias scores over time

To test the change in bias scores in the French course over time, a three-level model was used as explained in the method section (see also Table 3). We expected that students' bias would decrease when students were supported in reflecting on the grades they received. However, results showed no statistically detectable change in bias scores over time, $b = -0.03$, $t(8) = -0.60$, $p = .567$, nor a statistically detectable random effect of bias scores over time, $\chi^2(205) = 232.17$, $p = .094$. This means that bias scores did not improve, and that it also did not matter whether students received support in reflecting on the mismatch between estimated and actual grades.

Performance level differences

As with calibration accuracy, we examined the role of performance level. When examining the main effect, we found a statistically detectable effect on bias scores, $b = -0.34$, $t(200) = -4.43$, $p < .001$. High performance was associated with more underconfidence as indicated by the negative regression coefficient. As already indicated by the non-statistically detectable random effect, performance level did not influence how bias scores changed over time in a statistically detectable way $b = -0.01$, $t(200) = -0.23$, $p = .821$.

Effects of school year

Our results showed Year was a statistically detectable predictor of Bias (see Table 3). Students in year 1 showed more overconfidence than students in year 2 and year 3. Furthermore, there was an interaction between Time and Year. Depending on the year students were in, they improved their overconfidence or underconfidence. Students in Year 1 became less

overconfident, while students in year 3 improved their underconfidence. There was no statistically detectable interaction between Time, Intervention and Year, however. The way students improved after receiving Practice, Comparison or Reflection support did not depend on the year students were in, all t 's < 2.

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

We also tested the change in bias scores on German and Math. The intercept-only models that were ran showed that bias measured at Math exams showed level 3 variance, in contrast to bias measured at German exams, in which no level 3 variance was shown. Consequently, we ran a 3-level model for Math bias, and a 2-level model for German bias.

Results showed no statistically detectable main effect of Time for bias measured on Math exams and no random level-2 variance, t 's < 2, indicating that students did not differ from each other in how they changed their bias from pre-test to posttest. Results did show level-3 variance, however. When adding Year as a predictor, results showed that only students in year 3 improved their bias $t = 9.92$, $p < .001$, whereas students in year 2 and 1 did not show such an improvement. Results showed a statistically detectable change for bias scores on German exams, $t = -2.53$, $p = .012$. Students were less overconfident on the post-test ($M = 0.32$, $SD = 1.26$) than on the pre-test ($M = 0.69$, $SD = 1.20$), and as with calibration accuracy, this finding was the same for students in the different intervention groups, t 's < 2.

Improvement of second-order judgements over time

Following the reasoning that students who calibrate better also have more confidence in their performance estimates, we expected that the improvement in calibration accuracy would be accompanied by an increase in SOJs over time.

Results showed no overall change in SOJs over time, $b < -0.01$, $t(251) = -0.22$, $p = .829$. The random effect of SOJs over time was however statistically detectable, $\chi^2(213) = 254.04$, $p = .028$, indicating that there were differences between (groups of) students in how their confidence judgements changed over time. To examine whether this difference was due to the intervention group that students were in, we examined the interaction between the intervention groups and time. Results showed that there were no statistically detectable differences between the Practice group and the Grade comparison group, $t < 2$, see also Table 3. However, students in the Reflection group differed statistically detectable 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

We tested whether students of different performance levels changed their confidence scores differently over time. Results showed that performance level did not statistically detectable 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 statistically detectable 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 also tested the improvement in SOJs for German and Math on pre-test and posttest. The intercept-only models showed no level 3 (Class-level) variance for Math and German. Consequently, we ran 2-level models.

Results showed a statistically detectable main effect of Time for German SOJs, $t = -2.17$, $p = .031$. Students were a little less confident about their estimates (i.e., showed lower SOJs) on the posttest ($M = 3.12$, $SD = 0.87$) compared to the pre-test ($M = 3.27$, $SD = 0.78$). This effect was not random $\chi^2(170) = 185.86$, $p = .192$, indicating that this decrease in SOJs occurred for all students. For SOJs on Math exams, results showed no statistically detectable change from pre-test ($M = 3.31$, $SD = 0.92$) to posttest ($M = 3.24$, $SD = 0.87$), nor were there random variations between students in how they changed their Math SOJs, all t 's < 2 .

Improvement of exam performance (grades) over time

The mean and standard deviations for exam performance for the course French are reported in Table 2. To test a change in exam performance scores over time, we used a two-level model as explained in the method section. We tested whether performance improved over time and differed between different intervention groups. Our results showed that performance changed statistically detectable over exams, $b = 0.05$, $t(215) = 3.19$, $p = .002$. During the school year, students obtained better exam grades. Results also showed that the random effect of time on performance level was statistically detectable, $\chi^2(213) = 259.17$, $p = .017$, 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 however showed that differences in improvement over time were not due to the intervention groups, all t 's < 2 (see also Table 3). Furthermore, there were no main effects of our intervention groups, all 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. The intercept-only models that were ran showed both German and Math performance had level 3 variance. Therefore, we ran three-level models for both variables.

Results showed no statistically detectable main effect of Time, neither for German nor for Math, both t 's < 2 . We examined whether performance from pretest to posttest varied between persons or classes. To do so, we added a random effect for Time in level 2 (students) and level 3 (classes). After adding these effects the models did not converge, indicating that models with a random effect for Time did not fit the data. Consequently, we can conclude that our data showed no differences in Math performance from pre-test ($M = 6.88$, $SD = 1.46$) to post-test ($M = 6.96$, $SD = 1.71$) and in German performance from pre-test ($M = 6.12$, $SD = 1.24$) to post-test ($M = 6.09$, $SD = 1.40$).

Exploratory analyses

Reflections

As an exploration, we looked into the reflections provided by the students. Whereas all students compared their estimate to their grade, and reflected on the cues used when estimating their performance, 25.5% did not provide a reflection on a mismatch when asked to do so. When looking at the reflections written down, we see that 61.4% of these students provided an in-depth reflection such as “I noticed that I did not take enough time to read the instructions properly which caused me to misunderstand some questions”. From the remaining students, 25.2% just described the situation, for example, “I rated myself too low”; and another 13.4% indicated that they did not know why their estimate deviated from their actual grade.

Besides looking at the exact reflections, we also looked at which cues students in our reflection group used more often over the year. To do so, we included each of the cues as outcome variable in our model and included Time as a predictor. Over the year, the cue “How well I prepared myself for the exam” was reported more often, $b = 0.05$, $t(234) = 3.03$, $p = .003$. The other cues did not show significant increases or decreases.

Gender effects

In our analyses, we also checked for possible gender effects. There was a statistically detectable main effect of gender on Exam performance $b = -0.04$, $t(212) = -2.24$, $p = .026$, in which girls performed better than boys. In addition, there was a main effect of gender on students' Bias, $b = -0.39$, $t(190) = -2.89$, $p = .004$. Boys overestimated their performance to a larger extent than girls. However, gender did not affect how exam performance, nor bias changed over time, t 's < 2 . Furthermore, our results showed no statistically detectable main effect of gender on calibration accuracy and SOJs, nor on how calibration accuracy and SOJs changed over time, all t 's < 2 .

Discussion

Although the potential of using grades as performance feedback to aid students in more accurate monitoring of their own learning has been widely acknowledged (Brown et al. 2015; Carpenter et al. 2016; Clark 2012; Finn and Tauber 2015; Foster et al. 2017; Hacker et al. 2000; Miller and Geraci 2011a; Nietfeld et al. 2006), research has not yet been able to provide clear guidelines on how this can be achieved with especially younger students. The central aim of this article was therefore to examine whether and how providing support to students in secondary education in reflecting on their grades would improve their calibration accuracy over a full school year.

The results show that calibration accuracy already improves when students estimate their grade after each exam, even without any reflection support when they receive their actual grade. The results also show that this improvement was not just due to changes in performance, but that students actually changed their performance estimates. This indicates that the students became more aware how to accurately estimate their performance. This improvement of calibration accuracy is in line with studies showing that performance feedback in the form of grades or outcome scores can help students become more aware of their performance

(Callender et al. 2016; Hacker et al. 2000; Miller and Geraci 2011a; Nietfeld et al. 2006). Furthermore, whereas prior research found that performance level influenced whether or not students improved their calibration accuracy (e.g. Hacker et al. 2008), the current study did not find any such differences. Our intervention thus effectively helped students of all levels to improve their calibration accuracy.

Interestingly, and in contrast to our hypothesis, we did not find any differences in the improvement of calibration accuracy between our three intervention groups. Students who only estimated their performance improved their calibration as much as students who at each exam had to calculate the difference between their estimate and the actual grade, and also as much as students who received additional reflection prompts on each of their estimates. This is in contrast to findings by Bol et al. (2005) and Foster et al. (2017), who showed that university students who only estimated their performance did not improve their calibration accuracy over time. It may be that estimation practice only works for younger students, as they have more room for improvement in their calibration accuracy. But it should also be noted that the current study focused on postdictions instead of predictions, which were used by Bol et al. (2005) and Foster et al. (2017). When postdicting their performance students have more knowledge about the test. It is therefore possible that estimation practice only improves postdictions but not predictions. Furthermore, other than in the studies by Bol et al. (2005) and Foster et al. (2017), after receiving their actual grade, students received feedback on the scoring of the exam from the teacher. Increased knowledge about the scoring method may be especially helpful when making postdictions, as was the case in this study.

How best to explain the benefit of estimation practice in the current study? According to Koriat (1997), improvements in metacognitive awareness can be explained by a change in the cues used by the students. When improving calibration accuracy, students shift from using less valid cues (e.g., performance estimates on previous exams, Foster et al. 2017) to cues that are more informative of actual performance. An important question is, therefore, whether students in the current study also changed the cues to judge their performance. In the current study, we only obtained information of the cues used from students in our reflection group. Results show that these students indeed adjusted the cues used when estimating their grade. The cue 'How well I prepared myself for the exam' was reported more often. Hence, the results of this study show that students from secondary school can learn to better estimate their grade, and that they seem to do so by changing the cues they use when judging their performance.

In our study, extra reflection support did not further improve calibration performance. One reason may be that the intervention was not strong enough. While all students in our intervention followed the instructions to compare their estimate to their grade, and reflect on the cues used when estimating their performance, a quarter did not provide a reflection on a mismatch when asked to do so. Furthermore, our exploratory qualitative analysis showed that only 61.4% of these students provided an in-depth reflection. Consequently, it is possible that our reflection support did not lead to extra improvements because too many students did not use the reflection prompt, or did not reflect properly. However, it is difficult to make any solid conclusions. For example, stating that you do not know why your estimate deviated or describing the situation can signify a lack of reflection, but can also set the stage for deeper reflection—students may explore the reasons for their mismatch at a later stage with their parents or friends. Nevertheless, our intervention was very short, and may have been insufficient to stimulate deep reflection for these younger students. We therefore encourage future research to examine whether and how stronger interventions can be used to enhance calibration

accuracy, and how we can further help students adequately reflect on their performance to improve their calibration accuracy.

In the current study, we also included other measures of metacognitive awareness. Although we did not find any differences between our intervention groups on calibration accuracy and bias, we did find that reflection prompts influenced the second-order judgements students made. Over the school year, students in the reflection group became more confident in their performance estimates. This finding indicates that metacognitive awareness in terms of confidence judgements may develop separately from calibration accuracy. Given that the current study is the first intervention study that examined the effects on SOJs, more research is required to gain deeper understanding of how this metacognitive awareness is affected by the reflection prompts.

Limitations and directions for future research

This study has some limitations. First, we did not randomly assign students to intervention groups on an individual basis. Instead, whole classes were assigned to different intervention groups. This quasi-experimental design was chosen because we could not control spill-over within classes when students had to do different intervention exercises during the same class. A drawback of assigning classes to conditions is that there could have been class-level effects. We tried to control for class-level effects in the current study by including only classes that had the same teacher. Moreover, we controlled for class-level effects by applying multilevel analysis. Given our corrections, we consider it unlikely that our effects have been influenced by class-level effects, but we cannot rule them out completely. A replication of this study would provide important insight in the robustness of our findings.

Second, we did not include a separate group that did not participate in any intervention. Therefore, one could argue that the improvement in calibration accuracy we found was not due to any aspect of our intervention, but to some other factor, like maturation or the teacher. First, to control for maturation, we collected data on two other courses in which no intervention was implemented: German and Math. The German course was very similar to the French course in terms of instructional approach. Similarities included the type of exam given, as the German exam also tested vocabulary, sentence construction, grammar, listening, and reading comprehension, and in the fixed grading criteria that were used. Results indicate that both on German and Math, students did not show any improved calibration accuracy. Hence, maturation alone cannot explain our findings in the course French. A second alternative explanation for the improvement in calibration accuracy is that there might have been a specific teacher effect, as all intervention groups had the same French teacher. Interestingly, however, this teacher teaches all French classes during the first three years of high school, meaning that students have the same French teacher for three years. If the improved calibration was due to a teacher effect, students in year 3 should have been better calibrated than students in year 1. In our study, no such difference was found. As a consequence, we deem it unlikely that the increase in calibration accuracy is due to the teacher. We do, however, encourage future research to strengthen our findings and further examine the role of simple practice on calibration accuracy by including a group that participates in a pretest and posttest only. Furthermore, future research should determine whether our intervention also works for different teachers or whether additional training for teachers is necessary.

Third, we chose to conduct our study in the specific context of foreign language learning. In Dutch L2 learning, the assessment criteria are clear and easily understood by students.

Furthermore, the tasks students have to do for the subject French match the tasks from many prior studies. At the same time, although this context has characteristics that makes it very suitable to investigate our hypotheses, it may also limit the generalizability of our conclusions. For example, given that learner anxiety plays an important role in L2 learning (Teimouri, Goetze, & Plonsky 2019), it is unclear whether our interventions would have similar effects in different domains. Also, the improvements we found in achievement may partly be the result of an increase in self-efficacy, as these are strongly related in L2 learning (Lamb, 2017). In further research, it would not only be interesting to apply our intervention in different domains, but also investigate the mediating role of both learner anxiety and self-efficacy.

The current study is based on the premise that students can, to a certain degree, recall the cues used when judging their performance. If students are unable to do so, reflecting has little additional value. Prior research has shown that students may have difficulty recalling how they learned study materials (Winne and Jamieson-Noel, 2002). Note, however, that in the study by Winne and Jamieson-Noel, students had to recall their learning tactics once. In our study, we repeatedly encouraged students to focus on the cues they had used while estimating their performance. When students had not paid attention or could not fully recall which cues they had used, repeatedly having to reflect may at least have primed them to think of cues when estimating their performance on a new exam. Furthermore, in our intervention, students received their grades and reflected on their estimates shortly after taking the exam, as the grades were always available within a week. However, we cannot rule out the possibility that students were unable to accurately recall the correct cues used when estimating their performance, so this may be one reason why the additional support did not have the expected benefits. It is therefore important that future research makes an effort to test whether and how students can recall cues, or to ask students to provide cues when giving the estimate instead of when receiving feedback.

In the school taking part in the intervention, students took tests at the same time, and there were strict guidelines on how to score the exams. This allowed us to do our intervention study in a rather controlled setting. However, other schools may be less organized, and the grading procedure may not be implemented as strictly. Furthermore, there are differences between school subjects in how reliable the grading procedure is. For example, grading criteria for a history essay are often much less clear than for a test of French words. Thus in order to investigate the generalizability of our findings, more research is needed with different schools and with different school subjects.

Studies should also focus on whether the results could generalize to other types of metacognitive judgements. As mentioned in the introduction, there are many different ways in which students can be asked to estimate their performance, and it is possible that results found with one type of estimate do not immediately generalize to another type. For example, while students may improve their postdiction accuracy, they may remain poorly calibrated when predicting their performance. Also, it is possible that students may provide incorrect local estimates while being accurately calibrated on a more global level. To determine whether interventions generalize among different types of judgements, future research is required.

Another direction for future research is to examine the role of motivation. In the current study, students' motivation was not specifically targeted. It is important to note, however, that students also need to be motivated to invest time and effort in improving their calibration accuracy, as also mentioned in the self-regulated learning model of Zimmerman (2000). In addition to strengthening the intervention, future research should therefore test whether targeting students' motivation would benefit their metacognitive awareness.

Finally, we encourage future research to examine the shift in cue use when implementing an intervention to improve calibration accuracy. To this end, research may benefit from qualitative methods, such as asking students to think aloud when estimating their performance (Lajoie 2008; Winne and Perry 2000). This approach could help to attain an even more detailed insight into the (change in) cue use during and after calibration interventions.

Implications

Our results have important implications for practice. Student miscalibration is a widely demonstrated phenomenon in a variety of settings (Finn and Tauber 2015; Kruger and Dunning 1999; Sheldon et al. 2014), significantly contributing to students' inability to properly regulate their own learning, and thus leading to underachievement (Dunlosky and Rawson 2012; Finn and Tauber 2015; Metcalfe and Finn 2008; Nelson and Narens 1990). Subsequently, studies have focused on how to improve calibration accuracy, but results have remained mixed. Furthermore, most of these results are hard to generalize to a secondary school context, as studies have been conducted in a university context (e.g. Callender et al., 2016; Dunlosky and Rawson, 2012; Hacker et al., 2000; Miller and Geraci, 2011b), or even in a laboratory setting (e.g., Finn and Tauber 2015; Metcalfe and Finn 2008; Nederhand et al. 2019). Thus, the current study is one of the first to be able to provide specific guidelines as to how secondary schools may help students to become better calibrated. Asking students to estimate their grade after finishing their exam, a minimal exercise that can be readily implemented in class, and providing students with feedback regarding their outcomes, can already improve calibration accuracy.

Although not the main focus of this study, our results also underscore the importance of taking (formative) tests. When students estimate their performance after each test, such practice can help to become better calibrated. In turn, this improved monitoring can assist students in effectively regulating their own learning (Dent and Koenka 2016; Dunlosky and Rawson 2012; Nelson and Narens 1990). For example, students can gain more understanding of when they require help and when they should continue or stop studying. In line with this assumption, we already find better exam results for French at the end of our intervention. It should be noted that the current study was focused primarily on improving calibration accuracy. Thus, the improvement in exam performance could be further strengthened if we would also provide students with strategies on how to improve their regulation and learning if they signaled a mismatch between their estimated and actual performance. In the current study, students tried to find solutions for such mismatches themselves by, for example, watching Youtube videos about the topic or practising more with their parents. More evidence-based and clear guidelines would make the difference when not only aiming to improve their calibration accuracy but also their performance. Yet, the focus should not fully shift to providing students with instruction how to improve their performance only, as this can still leave students unaware of how they performed and how they improved this. This does not adequately prepare them for self-regulation which is problematic given that students become increasingly in charge of their own learning at all levels of education (Trilling and Fadel 2009; Wolters 2010). The findings of our study are therefore promising and can imply that schools can help students to develop their monitoring skills effectively with an easily implemented intervention.

Conclusion

To conclude, this study has shown that students in secondary education can learn to provide more accurate estimates of their own exam performance by a simple and easily implemented intervention. Asking students to estimate their own performance and providing them with performance feedback in the form of grades helps them to better calibrate their estimates on subsequent exams. This study is unique for a number of reasons. Through our year-long design in an ecologically valid secondary education environment, we were able to improve students' calibration accuracy in their actual day-to-day practice on their actual high-stake exams. Furthermore, the intervention tested in this study can be implemented very easily: teachers only have to hand the students the estimation forms and envelopes. Given that the effects of our study are shown in an actual classroom setting and did not require specific training for the teacher, the potential for practical application of our intervention is high. Our easy-to-implement intervention in which students explicitly have to think about their performance is a promising approach to improve students' calibration accuracy and an interesting steppingstone for future research. An important avenue for future research could be to test the boundaries of the effects (e.g., does the intervention work for different subject domains?) and the effects of additional teacher training (e.g., are teachers with different pedagogical styles able to use the intervention effectively?).

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 would like to 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 Abrahami for helping us with the data input.

Compliance with ethical standards

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- 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>.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5–31. <https://doi.org/10.1007/s11092-008-9068-5>.
- 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., 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 and 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. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, 11(2), 215–235. <https://doi.org/10.1007/s11409-015-9142-6>.
- Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, 28(2), 353–375. <https://doi.org/10.1007/s10648-015-9311-9>.
- Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24, 205–249. <https://doi.org/10.1007/s10648-011-9191-6>.
- 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>.
- De Bruin, A. B., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, 109(3), 294–310. <https://doi.org/10.1016/j.jecp.2011.02.005>.
- 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>.
- Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, 28(3), 425–474. <https://doi.org/10.1007/s10648-015-9320-8>.
- Destan, N., Spiess, M. A., de Bruin, A., van Loon, M., & Roebbers, C. M. (2017). 6- and 8-year-olds' performance evaluations: Do they differ between self and unknown others? *Metacognition and Learning*, 12(3), 315–336. <https://doi.org/10.1007/s11409-017-9170-5>.
- 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., 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>.
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, 24, 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>.
- Dunlosky, J., Hartwig, M. K., Rawson, K. A., Lipko, A. R. (2018) Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology* 64(3), 467–484. <https://doi.org/10.1080/17470218.2010.502239>
- Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Psychological Review*, 100(3), 363–406.
- Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, 58(1), 19–34. <https://doi.org/10.1016/j.jml.2007.03.006>.
- 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>.
- Fritzsche, E. S., Händel, M., & Kröner, S. (2018). What do second-order judgments tell us about low-performing students' metacognitive awareness? *Metacognition and Learning*, 13, 159–177. <https://doi.org/10.1007/s11409-018-9182-9>.

- Griffith, L. E., Cook, D. J., Guyatt, G. H., & Charles, C. A. (1999). Comparison of open and closed questionnaire formats in obtaining demographic information from Canadian general internists. *Journal of Clinical Epidemiology*, 52(10), 997–1005. [https://doi.org/10.1016/S0895-4356\(99\)00106-7](https://doi.org/10.1016/S0895-4356(99)00106-7).
- Gutiérrez 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. (2013) Students’ confidence in their performance judgements: a comparison of different response scales. *Educational Psychology* 35(3), 377–395. <https://doi.org/10.1080/01443410.2014.895295>
- 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>.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>.
- Hox, J. J. (2010). Multilevel Analysis. In *Multilevel analysis: Techniques and applications. International encyclopedia of statistical science* (2nd ed.). New York, NY: Routledge. https://doi.org/10.1007/978-3-642-04898-2_387.
- 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>.
- 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>.
- Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. Van Marsden & J. D. Wright (Eds.), *Handbook of survey research* (Vol. 62, 2nd ed., pp. 263–314). Bingley, UK: Emerald Group Publishing Limited. <https://doi.org/10.1111/j.1432-1033.1976.tb10115.x>.
- 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>.
- Lajoie, S. P. (2008). Metacognition, self regulation, and self-regulated learning: A rose by any other name? *Educational Psychology Review*, 20(4), 469–475. <https://doi.org/10.1007/s10648-008-9088-1>.
- Lamb, M. (2017). The motivational dimension of language teaching. *Language Teaching*, 50(3), 301–346. <https://doi.org/10.1017/S0261444817000088>.
- 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.
- Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development*, 82(6), 1778–1787. <https://doi.org/10.1111/j.1467-8624.2011.01649.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>.
- Miller, T. M., & Geraci, L. (2011a). 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., Abrahami, H., & Rikers, R. M. J. P. (2018a). Improving calibration over texts by providing standards both with and without idea-units. *Journal of Cognitive Psychology*, 30(7), 689–700. <https://doi.org/10.1080/20445911.2018.1513005>.

- Nederhand, M. L., Tabbers, H. K., Splinter, T. A. W., & Rikers, R. M. J. P. (2018b). The effect of performance standards and medical experience on diagnostic calibration accuracy. *Health Professions Education*, 4, 300–307. <https://doi.org/10.1016/j.hpe.2017.12.008>.
- Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. J. P. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology*, 33, 1068–1079. <https://doi.org/10.1002/acp.3548>.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, 26(26), 125–141. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](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>.
- Panadero, E., Brown, G. T. L., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, 28(4), 803–830. <https://doi.org/10.1007/s10648-015-9350-2>.
- 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>.
- Roebers, C. M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes. In P. J. Bauer & R. Fivush (Eds.), *The Wiley handbook on the development of children's memory* (Vol. 2, pp. 865–894). West Sussex: Wiley-Blackwell.
- Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, 45, 31–51. <https://doi.org/10.1016/j.dr.2017.04.001>.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. <https://doi.org/10.1007/BF00117714>.
- Schneider, W., & Löffler, E. (2016). The development of metacognitive knowledge in children and adolescents. In J. Dunlosky & S. K. Tauber (Eds.), *Oxford Handbook of Metamemory* (pp. 491–518). Oxford University Press.
- Schraw, G. (2009). Measuring metacognitive Judgements. In *Handbook of Metacognition in Education* (pp. 439–462). doi:<https://doi.org/10.4324/9780203876428>.
- 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), 125–137. <https://doi.org/10.1037/a0034138>.
- Steiner, M., van Loon, M. H., Bayard, N. S., & Roebers, C. M. (2020). Development of Children's monitoring and control when learning from texts: Effects of age and test format. *Metacognition and Learning*, 15(1), 3–27. <https://doi.org/10.1007/s11409-019-09208-5>.
- Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, 83(2), 282–309. <https://doi.org/10.1006/obhd.2000.2910>.
- 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>.
- Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41(2), 363–387. <https://doi.org/10.1017/S0272263118000311>.
- 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>.
- Trilling, B., & Fadel, C. (2009). 21st century skills. *Jossey-Bass*, 256. <https://doi.org/10.1145/1719292.1730970>.
- Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & 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>.
- William, D. (2016). The secret of effective feedback. *Educational Leadership*, 73(7), 10–15 Retrieved from <https://www.scopus.com/record/display.uri?eid=2-s2.0-84978969364&origin=resultslist&sort=plf-f&src=s&sid=54c5ee97b5760a54b634802811095048&sot=autdocs&sdt=autdocs&sl=17&s=AU-ID%286602672352%29&relpos=5&citeCnt=4&searchTerm>.

- Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology*, 27, 551–572. [https://doi.org/10.1016/S0361-476X\(02\)00006-1](https://doi.org/10.1016/S0361-476X(02)00006-1).
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). San Diego, CA: Academic Press.
- Wolters, C. A. (2010). Self-regulated learning and the 21st century competencies. USA University of Houston, (May), 27. Retrieved from http://www.hewlett.org/uploads/Self_Regulated_Learning__21st_Century_Compencies.pdf
- 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.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.