Metacognition Learning (2016) 11:139–170
DOI 10.1007/s11409-015-9139-1

Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks

Trinidad García[1] · Celestino Rodríguez[1] ·
Paloma González-Castro[1] ·
Julio Antonio González-Pienda[1] · Mark Torrance[2]

Received: 18 May 2014 / Accepted: 28 April 2015 / Published online: 8 May 2015

Abstract Calibration, or the correspondence between perceived performance and actual performance, is linked to students' metacognitive and self-regulatory skills. Making students more aware of the quality of their performance is important in elementary school settings, and more so when math problems are involved. However, many students seem to be poorly calibrated, with a tendency towards over-confidence. The present study analyzes the relationship between post-performance calibration accuracy and the metacognitive process shown by 524 fifth- and sixth-grade students while solving two math problems. After calculating a calibration index and establishing the stability of students' judgments and actual performance, differences in the metacognitive process exhibited by students with different calibration accuracy (Accurate vs. Inaccurate groups) were analyzed. The emergence of different calibration patterns and differences in the metacognitive process as a function of mathematics achievement and grade level were also examined. Results indicated that: (a) students in the overall sample were little calibrated and over-confident, showing high stability in their judgments and actual performance across problems; (b) inaccurate students reported using information representation sub-processes (drawing/summarizing) less frequently, but writing and reviewing (and also correcting mistakes) more frequently than their accurate peers; and (c) differences in calibration patterns and the metacognitive process were found when achievement level was considered, whereas grade level did not generate any important effect. These findings suggest the usefulness of process-based measures to examine the metacognitive

✉ Celestino Rodríguez
rodriguezcelestino@uniovi.es

[1] Faculty of Psychology, University of Oviedo, Room 225. Plaza Feijoo s/n, 33003, Oviedo, Asturias, Spain

[2] Nottingham Trent University, Nottingham, UK

processes involved in making post-performance judgments, considering achievement and its possible mediating role in this relationship.

Keywords Calibration · Mathematics · Metacognition · Problem-solving · Process · Self-regulation

Introduction

Solving math problems is a very common activity in elementary school. However, it is also a complex cognitive activity involving multiple processes. In order to successfully perform these tasks, students must integrate cognitive, metacognitive, and self-regulatory mechanisms (Cleary and Chen 2009; Montague et al. 2011) to accurately regulate and monitor their learning processes. An important aspect of monitoring is *calibration*, or the degree to which students' judgments about the correctness or adequacy of their performance corresponds to their actual performance, the latter determined on the basis of an objective measure (Hacker et al. 2008a). These judgments can be expressed in terms of *predictions* or *postdictions*, depending on whether they are made prior to or after completing a task. Postdictions (or post-performance judgments) are indicative of monitoring mechanisms during task performance and have been shown to be more accurate than predictions, presumably due to the additional feedback provided by performing the task (Ackerman and Wolman 2007).

The relevance of calibration mechanisms in mathematics has been substantially demonstrated, suggesting a direct and positive relationship between these metacognitive control processes and mathematics achievement. Within the context of math problem-solving, high-calibrated students tend to perform more successfully than low-calibrated students, presumably because they have more control over their problem-solving processes (Desoete and Roeyers 2006; Hadwin and Webster 2013; Jacobse and Harskamp 2012; Lipko et al. 2009; Özsoy 2012; Rinne and Mazzocco 2014). However, previous literature has shown that students are often poorly calibrated, showing a tendency towards over-confidence. Furthermore, these patterns of inaccuracy tend to persist over time (Bol et al. 2012; Dinsmore and Parkinson 2013; Hacker et al. 2008b; Stolp and Zabrucky 2009). Given the possible implications of different patterns of calibration accuracy on students' learning, it is necessary to determine which factors influence their judgments. In this sense, a review of the current literature shows that: (a) most research has focused on the study of these mechanisms in undergraduate or secondary school students, paying less attention to elementary school ages; (b) most studies have been conducted in laboratory settings, which may be an important constraint for generalization of results; and (c) although both predictions and postdictions have been substantially studied in math problems, the amount of research conducted from the perspective of the analysis of the processes involved in these tasks is almost non-existent to date. This analysis is especially interesting in the case of postdictions, given their relationship with monitoring processes during task performance.

According to the above, the main objective of the present study was to analyze the relationship between students' post-performance calibration accuracy and the metacognitive processes they displayed while solving two math problems. A large sample of 524 fifth- and sixth-grade students took part in this study. To address this goal, students were assessed in a regular mathematics class, using the *Triple Task Procedure in Mathematics* (TTPM; García and González-Pienda 2012; García et al. 2015) as a measure of the process.

Importance of calibration for problem-solving

Calibration has been characterized as an important metacognitive process involved in the development of self-regulatory competence (Dinsmore and Parkinson 2013; Hadwin and Webster 2013; Labuhn et al. 2010; Zimmerman and Schunk 2011). Because of its implications in students' motivation, metacognitive control, and self-regulation, it is an essential process affecting academic success and task completion (Alexander 2013; Cleary 2009; Efklides and Misailidi 2010; Schunk and Pajares 2009). In this sense, as Alexander pointed out, students' ability to accurately gauge their progress and performance plays an important role in their subsequent effort and strategy use in problem-solving situations. In this context, mathematical problem-solving is an important issue in current scientific literature. It is characterized as a complex cognitive activity, but students tend to show poor metacognitive skills while performing math problems. Many students (even those without learning disabilities) forge ahead without considering alternative approaches, jumping immediately into calculations, giving impulsive responses, and using trial and error as a strategy to solve these tasks. They commonly get stuck in irrelevant details, or fail to verify solution paths and evaluate the suitability of their responses (Cleary and Chen 2009; Kramarski and Gutman 2006; Montague et al. 2011; Pennequin et al. 2010; Pereis et al. 2009). Literature also indicates that students tend to be inaccurate when asked to predict or judge the results of their performance, either in mathematics and other academic areas, showing a tendency towards over-confidence (Bol and Hacker 2001; Bol et al. 2005, 2010; Hacker et al. 2008a; Lipko et al. 2009; Özsoy 2012). This is an important aspect, as the calibration of students' metacognitions to their performance has important consequences for learning outcomes (Dunlosky and Rawson 2012; Dunlosky and Thiede 2013; Finn and Metcalfe 2014; Hacker et al. 2008a).

Calibration measurement

Whereas the relevance of calibration in learning has been substantially demonstrated, an important issue that must still be addressed is the distinction between *absolute* and *relative calibration*. The term *calibration* used in the present study refers to *absolute accuracy*, or the degree of correspondence between the judged level of performance and the actual performance (Hacker et al. 2008a; Maki et al. 2005). Thus, absolute accuracy (or calibration) differs from the concept of *relative accuracy*, also referred to as *resolution* or *discrimination*. This component has traditionally been defined as the degree to which a person's judgments can predict the likelihood of correct performance of one item relative to another (Hacker et al. 2008a). Both types of calibration are different aspects of metacognitive monitoring, showing low correlations with each other (Maki et al. 2005).

In calibration studies, learners are asked to make confidence judgments, which consist of judging the degree to which they think they will be able to acquire new knowledge or perform a task (prediction), or to judge their performance once the task has been completed (postdiction). This judgment is then compared with an objective measure of that learning (their actual performance), such as the score on an exam or test, after completing the task (Bol et al. 2010; Winne 2004). The more closely a student's judgment of performance matches his/her actual performance, the better calibrated the student is (Hacker et al. 2008a). In the context of the present study, the distinction between prediction and postdiction acquires a special meaning. In this sense, postdictions have been shown to be more accurate than predictions, possibly due to the additional feedback provided by performing the task (Ackerman and

Wolman 2007). This has led many authors to characterize these judgments as informative of monitoring processes during task performance, incorporating this measure into their studies (Bol and Hacker 2001; Bol et al. 2005, 2010; Hacker et al. 2008b; Nietfeld et al. 2006; Sheldrake et al. 2014).

Among the methods to calculate calibration, Parkinson et al. (2010) note dichotomous ratings, categorical ratings (such as Likert-type scales), data count, and 100-mm scales as the most commonly used measures. In studies using dichotomous ratings, students are asked to complete a multiple-choice recall measure of whether they feel "confident" or "not confident" about each answered item (Schraw et al. 2012). In the case of Likert-type scales, they are requested to rate the degree to which they feel confident, ranging from "not confident" to "very confident" (Hattie 2013). Within the numerous indexes used to express judgment accuracy, different measures are feasible, from Pearson correlations to more complex measures, such as rho or the Gamma coefficient (Dinsmore and Parkinson 2013; Schraw 2009; Winne and Muis 2011). In the present study, a dichotomous measure of calibration was used. Specifically, Phi coefficient was calculated (Schraw et al. 2012). As these authors point out, although continuous measures may be more sensitive, dichotomous judgments are the most common in the literature. This may be due to the fact that a dichotomous scoring scheme is easier for participants to understand than a more complex scheme. This is especially relevant when studies are conducted with young samples. In this sense, although there are several dichotomous measures of calibration (e.g., the G index, the Odds ratio, Gamma, Kappa, Phi, or the Sokal distance measure), Schraw et al. (2012) showed in their study that all of them present high and similar sensitivity and specificity.

As mentioned above, an important issue in calibration studies is the fact that students tend to be inaccurate in their performance judgments, with a tendency to over-confidence (Bouffard et al. 2011; Butler 2011; Dinsmore and Parkinson 2013; Hadwin and Webster 2013). This tendency has been found in studies at different educational stages, across different subject areas, and in mathematics in particular (Bol and Hacker 2001; Bol et al. 2005, 2010; Rinne and Mazzocco 2014; Sheldrake et al. 2014). Furthermore, there is a good deal of evidence suggesting that calibration judgments tend to be stable over time (Bouffard et al. 2011; Hacker et al. 2008a). These studies also indicate that performance judgments (and related biases) are resistant to improvement. This has led to an important line of research on the determining factors underlying students' judgments and how to increase their accuracy. In this context, numerous studies have focused on exploring many different components. Aspects such as students' general cognitive ability or prior knowledge, academic achievement or their level of expertise in making such judgments, task or item characteristics, affective-motivational components such as goal-setting orientation or self-efficacy beliefs, or more stable and persistent traits such as attribution styles or personality, have been proposed as possible explanatory factors (Alexander 2013; Bol et al. 2012; Dinsmore and Parkinson 2013; Hacker et al. 2008b; Hadwin and Webster 2013; Stolp and Zabrucky 2009).

Metacognition, self-regulated learning, and calibration in mathematics

Within this perspective, another important determining factor of students' judgment accuracy is the link between calibration and metacognitive and self-regulatory mechanisms. The two terms (metacognition and self-regulation) have occasionally been used interchangeably (Dinsmore et al. 2008). However, it is important to note that there is a large difference between them. On the one hand, self-regulation is defined in the context of learning as the control students exert over their cognition, behaviour, emotion, and motivation to achieve established goals. On the other hand, metacognition (or the ability to think about thinking) refers to the cognitive component of self-regulation, in charge of processes such as self-awareness of problem-solving, monitoring, and controlling one's mental processing (Dinsmore et al. 2008; Panadero and Alonso-Tapia 2014). Nevertheless, both metacognitive and self-regulatory components are articulated under the paradigm of Self-regulated Learning (SRL: Zimmerman 2000, 2008).

SRL explores the acquisition, evaluation, and regulation of knowledge. It is rooted in the conception of learners as being capable of monitoring their own learning, setting goals, evaluating, and regulating their own learning progress (Stolp and Zabrucky 2009). This perspective is intimately linked to the concept of calibration, or the ability to accurately judge one's performance on a task. Calibration is relevant in this context, as it consists of a metacognitive control process that provides information about the status of one's knowledge and strategies at a cognitive level. It also sets the stage for effective self-regulation by generating the internal feedback that students use to control their learning and performance (Bol et al. 2012; Dunlosky and Rawson 2012). In this sense, given that Zimmerman's (2000) model establishes that self-regulated learners are more aware of what they do and do not know, they are also expected to exhibit more accurate perceptions of their performance. This statement is supported by studies showing that self-regulated learners are commonly high-achieving students who are more accurately calibrated than their peers with lower levels of self-regulation. These studies also reveal that high-achieving students tend to be somewhat under-confident, whereas low-achieving students are commonly inaccurate and over-confident when asked to predict or postdict their performance (Bol and Hacker 2001; Bol et al. 2005; Hacker et al. 2008b; Özsoy 2012). These results support the role of calibration as a metacognitive and self-regulatory process, as well as its relationship with students' achievement.

Regarding the impact of calibration mechanisms on mathematics (and problem-solving in particular), the relationship between calibration mechanisms and mathematics has been substantially demonstrated (Desoete and Roeyers 2006; Jacobse and Harskamp 2012; Özsoy 2012; Rinne and Mazzocco 2014). First, calibration accuracy has been shown to explain between 16 and 36 % of the variance of mathematics achievement in different studies (see Jacobse and Harskamp 2012). Specifically, a recent study by Rinne and Mazzocco (2014) supported this association, suggesting that calibration accuracy of mental arithmetic judgments represents an important developmental predictor of future mathematics performance in elementary school. Second, highly calibrated students have been observed to solve math problems more successfully. Specifically, previous studies show that these students can solve more complex problems using more flexible and effective strategies, such as separating the tasks into simpler parts, or designing different ways to represent information, concepts, and relationships. They also formulate self-queries to clarify their thoughts and they tend to evaluate the result of their performance more frequently than their less accurate peers (Özsoy 2012). Hence, highly calibrated or accurate students tend to show more self-regulatory and metacognitive skills during problem-solving (Boekaerts and Rozendaal 2010; Dinsmore and Parkinson 2013; Hadwin and Webster 2013; Stolp and Zabrucky 2009; Winne and Muis 2011). However, previous research has also revealed that many students fail to use self-regulatory and metacognitive mechanisms spontaneously when solving math problems, leading to poor performance and inaccurate self-evaluations (Kramarski and Gutman 2006; Pennequin et al. 2010; Pereis et al. 2009). In this context, one of the greatest challenges for researchers is to obtain evidence of the cognitive, metacognitive, and strategic processes used by students during these tasks, and of how these processes relate to their ability to make accuracy judgments. This information can be gathered by means of on-line measures, such as Think-Aloud (Jacobse and Harskamp 2012; Montague et al. 2011) or Triple Task protocols (Olive and Piolat 2002; Piolat et al. 2001, 2005).

Process-based measures: the triple task procedure in mathematics (TTPM) and its predecessor

Self-regulatory and metacognitive mechanisms have been traditionally assessed by means of questionnaires or structured interviews. These kinds of measures are based on students' self-reports about how they tackle a problem, or the extent to which they use different strategies. Students' responses are then scored depending on their quality, and a total score is calculated. However, this kind of assessment—although widely applied—may yield inconsistent or incorrect information due to inaccurate recall or response biases, such as social desirability. In this sense, these tools have been considered useful to assess students' declarative and situational knowledge, but they do not provide information about how learners transfer such knowledge to regulate their problem-solving process (Cleary and Chen 2009; Veenman 2011). In order to address this issue, the use of on-line methods (i.e., process-based measures) is an alternative. These measures are taken concurrently with task performance and have been shown to be especially useful to provide evidence of students' cognitive processes during different activities (Azevedo and Aleven 2013; Throndsen 2011; Tillema et al. 2011).

An example of these measures is the *Triple Task* technique (Piolat et al. 2001). This procedure has a long tradition in the study of the processes involved in composition writing (Olive and Piolat 2002; Piolat et al. 2005). In traditional Triple Task studies, participants are required to perform three tasks simultaneously: a primary task (e.g., the composition of a text); a secondary probe task (based on response time [RT]); and a third task in which they are asked to verbalize or label the actions or thoughts that are interrupted by the probe. This method uses directed introspection. In this sense, probes are presented at a certain time interval, and after each reaction to a probe, participants are asked to categorize their actions or thoughts according to a given category system that includes different writing sub-processes. This technique is aimed at providing information about the activation of different sub-processes and cognitive demands as the written task progresses. For this purpose, this protocol is based on two measures: (a) introspective data, which provide information about how the writing sub-processes are activated over time; and (b) RT to the probes, which allows studying the cognitive effort involved in these sub-processes. Specifically, RTs are taken before (i.e., baseline condition) as well as during Triple Task administration (i.e., secondary probe task). RTs taken during Triple Task are called "secondary RTs" and are linked to a reported action or sub-process. Differences between RTs in the two conditions are interpreted as an estimation of the cognitive effort associated with each writing sub-process or category. Before starting the tasks, participants are trained to recognize each of the sub-processes of the given category system. As a measure of the process, the Triple Task procedure differs from Think-aloud protocols (Montague et al. 2011) in two main aspects: (a) the Triple Task procedure uses directed introspection, as opposed to the undirected introspection used in Think-aloud. In this sense, providing a system of categories facilitates students' categorization of their thoughts or actions. It also allows the experimenter to limit the number of target sub-processes, exploring those of special interest; and (b) response transcription and coding are not required. This aspect makes the Triple Task technique less costly than Think-aloud, and therefore, also more suitable to conduct studies with larger sample sizes.

Within this context, an adaptation of the Triple Task method has been recently proposed for application in mathematical problem-solving, the *Triple Task Procedure in Mathematics* (TTPM; García and González-Pienda 2012; García et al. 2015). This is a modification of the traditional Triple Task technique, which grants less prominence to cognitive effort (RTs) than the original version. Instead, it has been designed to examine the process underlying solving math problems from the viewpoint of the main SRL phases of Planning, Execution, and Evaluation (Rosário et al. 2008; Zimmerman 2000). The authors combined the SRL perspective with Bransford and Stein's (1993) IDEAL model of problem-solving to design a new category system, which serves as the basis for directed introspection. García et al. (2015) administered this protocol to 510 fifth- and sixth-grade students from Northern Spain, who carried out mathematical tasks of varying difficulty. Students in this study showed ineffective planning strategies and a lack of evaluation mechanisms. However, students who successfully solved the problems showed significantly better metacognitive skills while performing the problems, mainly related to the SRL phase of planning. These differences gained prominence as task difficulty increased, suggesting a relationship among problem-solving performance, metacognitive processes, and task difficulty. A brief review of the SRL and problem-solving models used in TTPM is presented below, before the proposed assessment protocol more extensively described in this section.

Self-regulated learning model

There are different theories and models that attempt to explain SRL mechanisms. Most of them share the common ground that self-regulation involves different processes (e.g., planning, monitoring, etc.) and is cyclical (i.e., previous performance provides feedback for future tasks). Authors such as Panadero and Alonso-Tapia (2014) suggest that the models proposed by Boekaerts (1999), Winne (2001), and Zimmerman (2000, 2008) are the most representative of the last few years. Within this context, Zimmerman's model has been the most frequently cited in the literature, and the most extensively used to guide interventions (Moos and Ringdal 2012; Stoeger and Ziegler 2008). Due to its practical implications, it was selected as a framework for the TTPM design. Briefly, this model comprises three cyclical phases (forethought, performance, and self-reflection), which correspond to the processes that occur before, during, and after SRL takes place. In the forethought phase, students analyze the task and assess their capacity to perform it, establishing goals and plans to complete it. The second phase is performance, in which two types of processes are carried out: self-control and self-observation. This phase is aimed at keeping track of progress during task performance, and involves using the strategies identified during the forethought phase. These strategies can be classified as metacognitive (e.g., self-instruction, imagery, time management, etc.), and motivational strategies (e.g., self-given messages to remind one of the goal, self-rewards, etc.). The last phase is self-reflection, in which students judge their performance and formulate the reasons for their results (Panadero and Alonso-Tapia 2014).

Problem-solving model

Several problem-solving models have been proposed from different conceptual perspectives. One of these perspectives is rooted in cognitive psychology. Focused on the mental processes underlying problem-solving, some models conceive the application of general problem-solving strategies, called "heuristics", as the basis for problem-solving expertise and performance. Heuristics involve different processes (e.g., identifying the problem, drawing a graph or diagram, thinking about structurally similar problems, simplifying the problem, reviewing, or generalizing), which are commonly expressed as a series of stages through which problem-solving progresses. The models of Polya (1954), Gick (1986), and the IDEAL Model of Bransford and Stein (1993) have been the most studied models within this perspective. The IDEAL Model describes five stages: Identifying potential problems, Defining and representing the problem, Exploring possible strategies, Acting on those strategies, and Looking back and evaluating the effects of those activities. It is currently one of the most extended models (Newton et al. 2009; 2012). In contrast to the other mentioned models, it establishes the problem-solving process as cyclical, in accordance with Zimmerman's SRL Model (2000, 2008). As it also provides a comprehensive system to examine the metacognitive and self-regulatory processes involved in problem-solving during performance, this model makes up the other basis for the TTPM.

From the combination of Zimmerman's (2000, 2008) SRL model and the IDEAL model (Bransford and Stein 1993), a system emerged with eight categories or sub-processes, organized in three higher level categories, corresponding to the main SRL phases (Rosário et al. 2008; Zimmerman 2000): Planning, Execution, and Evaluation. Table 1 shows the category system used. In accordance with previous studies, an additional category ("other") has been included to reflect all the thoughts or activities unrelated to problem-solving performance. Some examples of processes within this category are day-dreaming (e.g., "I'm thinking about what I'm going to do this afternoon") or possible distracters (e.g., "I'm looking out the window"). This additional category was also included in the categorization training phase, which is previous to the TTPM (described in Procedure).

An important issue in this category system is the allocation of the drawing/summarizing sub-process in the SRL planning phase. In this sense, imagery, or the construction of mental images to organize the information, is similar in essence to drawing/summarizing, and falls under the performance phase in Zimmerman's (2000, 2008) SRL model. Specifically, it consists of a metacognitive strategy under self-control, as mentioned above. The location of this strategy in the planning phase in TTPM is due to the problem-solving model adopted (i.e., IDEAL model). Specifically, in this model, the Definition and representation stage is prior to execution (or Action based on the strategy). Previous studies using TTPM (García and González-Pienda 2012; García et al. 2015) confirmed the allocation of this sub-process under

Table 1 Category system. Based on the SRL phases (Rosário et al. 2008; Zimmerman 2000, 2008) and the IDEAL model (Bransford and Stein 1993)

SRL Model	IDEAL Model	Process categories (I am …)
Planning	Identification of the problem	Reading
	Definition and representation	Drawing or summarizing
		Recalling similar problems
	Exploration of possible strategies	Thinking about a solution
Execution	Action based on the strategy	Calculating
		Writing a response
Evaluation	Look at effects of solutions	Reviewing
		Correcting mistakes
"Other"		Doing something unrelated

the SRL planning phase. This latter study analyzed the temporal sequence of the process in a sample of 350 elementary students during TTPM. This sequence was based on the percentage of students who reported each of the eight TTPM categories or sub-processes across 20 different moments while solving several math problems. Results indicated that: (a) although the drawing/summarizing sub-process tends to be present both in planning and execution phases, it is mainly activated during the first phase; and (b) the frequency of drawing/ summarizing decreases dramatically as the frequency of the calculation sub-process (i.e., execution phase) increases.

Some of the potentialities of the TTPM are its design and implementation features (see Procedure section in the present study), which facilitate its applicability to broad samples, and more important, in classroom settings (García et al. 2015). In this sense, one of the most discussed aspects in calibration studies is the fact that they are commonly conducted in laboratory settings (Dinsmore and Parkinson 2013; Hadwin and Webster 2013; Winne and Muis 2011). As Hacker et al. (2008a) pointed out, although these contexts often provide important information on calibration accuracy mechanisms, generalizing these findings to different contexts, especially to classroom contexts, can be difficult. Laboratory study components, material and procedures—such as highly structured contexts, tasks with little meaning for students, or tasks presented over short periods of time—differ considerably from those used in classrooms settings, where students must deal with multiple requirements from the environment and the task. Additionally, although a large amount of research has focused on the study of the relationship between calibration, metacognitive and self-regulatory mechanisms in mathematics (Bol et al. 2010; Hacker et al. 2008a, 2008b; Rinne & Mazzocco 2014; Sheldrake et al. 2014), and in mathematical problem-solving in particular (Jacobse and Harskamp 2012; Zimmerman et al. 2011), none of these studies has been conducted from the perspective of the analysis of the relationship between students' accuracy judgments and the metacognitive process involved in solving math problems. The study of this relationship is especially interesting in the case of postdictions, which have been suggested to be informative of monitoring mechanisms during task performance.

The present study

This study is aimed at examining the metacognitive process shown by a sample of 524 fifth- and sixth-grade students while solving two math problems, and how this process explains differences in students' post-performance accuracy judgments. For this purpose, the TTPM technique (García and González-Pienda 2012; García et al. 2015) will be used as a measure of the process. Additionally, the possible influence of variables such as mathematics achievement and grade level both on metacognitive processes and calibration accuracy will be analyzed. In this sense, whereas the relationship between students' achievement in calibration has been substantially demonstrated (Bol et al. 2005, 2010; Hacker et al. 2008b; Özsoy 2012), it is not clear whether grade level may lead to the same pattern of results. However, as different grade levels may indicate different cognitive or developmental stages (Boston and Smith 2009), this variable is expected to exert some degree of influence on the analysed variables.

Overall, this study aimed to answer the following questions: (a) Do students with different accuracy in their post-performance judgments show differences in their problem-solving processes? (b) Which SRL phases or sub-processes explain these differences? (c) Will different patterns of calibration accuracy emerge according to students' achievement in mathematics and grade level, and will these variables account for differences in the metacognitive processes? Prior to addressing these questions, the correspondence between students' post-performance judgments and actual performance was estimated to establish the degree to which the students were calibrated and the tendency of their judgments (i.e., over- or under-confidence) in both problems. In addition, post-performance judgment and actual performance stability across problems were analyzed. Previous laboratory and classroom studies have shown that calibration tends to be relatively stable over time and across tasks. Students also tend to be more stable in their performance judgments than in their actual performance (Bol et al. 2005; Bouffard et al. 2011; Hacker and Bol 2004; Hacker et al. 2008b).

Given these assumptions and according to previous research results, it is expected that:

1) Students will be imprecise in their judgments, showing a tendency towards over-confidence.
2) Students will be stable in their judgments and actual performance across problems. This stability will be higher in judgments than in actual performance.
3) Differences in the metacognitive process (TTPM phases and sub-processes) are expected to be found between students with different accuracy in their post-performance judgments (Accurate group = when students' judgments and actual performance match vs. Inaccurate group = when they do not). Accurate students will show more effective metacognitive strategies during TTPM, such as using different forms of organizing and representing the information, more frequently than their inaccurate peers. They will also evaluate the progress and results of their performance more than students in the inaccurate group.
4) Different patterns of calibration accuracy will emerge as a function of students' mathematics achievement level. High-achieving students will be more accurate and under-confident than their peers with lower levels of achievement. This variable is also expected to generate differences in the metacognitive process during TTPM. High-achieving students will use more effective solving-problem strategies (e.g., more frequent use of information organization and representation strategies, better evaluation mechanisms).
5) If grade level has an impact on calibration and the metacognitive process during TTPM, sixth-grade students will show higher levels of calibration accuracy, and better planning and evaluation strategies.

Materials and methods

Participants

Five-hundred and twenty-four students from fifth and sixth grade of elementary school took part in this study. They were recruited from 11 schools in Northern Spain. Ages ranged between 10 and 13 years (*M*=10.991, *SD*=0.716). Of these students, 260 (49.6 %) were female, and 264 (50.4 %) were male. The sample comprised 220 students from the fifth grade (42 %; male=108, female=112) and 304 students from the sixth grade of elementary school (58 %; male=156, female=148).

Students were selected through convenience sampling. They volunteered for the study and presented signed parental permission to participate. Children with a diagnosis of severe learning disabilities or those who showed reading difficulties were excluded from the analyses.

Given that data analyses were conducted separately for each math problem, students were distributed according to the accuracy of their post-performance judgments in each problem (Accurate vs. Inaccurate groups). Post-performance judgments and actual performance were both expressed in dichotomous terms (Success=1, Failure=0). Students were assigned to the accurate group when their post-performance judgments matched their actual performance, and to the inaccurate group when there was a mismatch. There were no statistically significant group differences in age (p=.954), grade level (p=.489), and gender distribution (p=.253) in Problem 1. Regarding Problem 2, gender (p=.177) and grade level (p=.405) were equally distributed in the two groups, and there were no group differences in age (p=.921).

Variables and measurement instruments

In accordance with the objectives proposed in this study, three sets of measures were used: those related to the problem-solving process (TTPM); those applied to calculate post-performance calibration; and a measure of students' achievement in mathematics.

Process measures

Evidence of students' problem-solving processes was obtained by means of the *Triple Task Procedure in Mathematics* (TTPM; García and González-Pienda 2012; García et al. 2015), previously described in the present study. The main characteristics of the TTPM are related to its category system, based on Bransford and Stein's (1993) IDEAL Model and the SRL perspective (Rosário et al. 2008; Zimmerman 2000). It comprises eight sub-processes (reading, drawing or summarizing, recalling similar problems, thinking about a solution, mental calculation, writing a response, reviewing, correcting mistakes) grouped into three main phases (planning, execution, evaluation). Dependent variables were students' relative frequency of election of each category and phase, expressed in percentages.

Post-performance calibration measures

Three measures were established (actual performance, post-performance judgments, and a calibration index):

Actual performance was obtained by asking students to complete two math word problems taken from the book "*Problem-solving and comprehension*" (Whimbey and Lochhead 1999), published in Spanish. Actual performance in the problems was expressed in terms of success (1) or failure (0). Although the second problem involves two different questions (number of dogs and their breeds), no partial scores were awarded in order to unify the measurement system. Thus, both parts of the problem had to be answered correctly to receive credit. Prior to commencing the study, these problems were reviewed by teachers to ensure that they were developmentally appropriate for students. The two math problems are presented below:

Problem 1 "*Beatriz lends €700 to Susana. But Susana borrows €1500 from Esther and €300 from Juana. In addition, Juana owes Esther €300 and Beatriz €700. One day they meet at Beatriz's home to settle their debts. Who went back home with €1800 more than she brought*?"

Problem 2 "*Paula, Mari, and Juana have a total of 16 dogs, 3 of which are poodles, 6 are hounds, and the rest of them are German shepherds and Pekinese dogs. Juana*

> *does not like poodles and Pekinese dogs, but she has 4 hounds and 2 German shepherds, leading to a total of 6 dogs. Paula has a poodle and 2 more dogs, which are German shepherds. Mari has 3 Pekinese dogs and several dogs of other breeds. Which breeds, and how many dogs of each breed, does Mari have*?"

Post-performance judgments were obtained by asking students after each problem whether they considered that they had solved the problem successfully. They had to respond in dichotomous terms (yes/no), and the values of 1/0 were assigned, respectively. The dichotomous nature of the post-performance judgment scores is based on two criteria: (a) the characteristics of the problems (i.e., actual performance was established as success or failure); (b) an attempt was made to provide participants with an easier scoring scheme, given the youth of the sample (i.e., fifth- and sixth-grade students).

Finally, a calibration index was calculated as the correspondence between post-performance judgments and actual performance. As both variables were dichotomous, Phi correlation coefficient ρ_{ϕ} was used for this purpose.

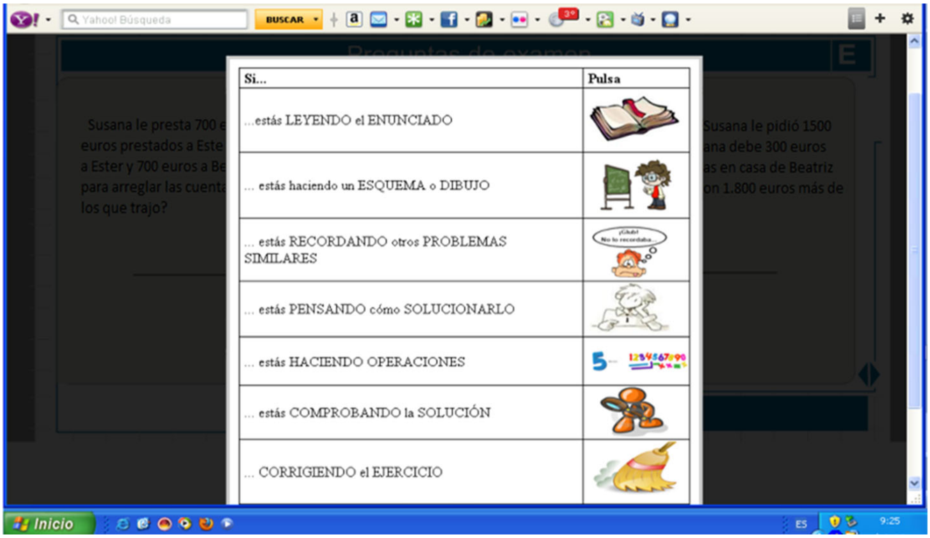
Mathematics achievement level

Students' final academic grades in mathematics were used as an indicator of mathematics achievement. This information was provided by the teachers in charge of mathematics, after the parents' permission was obtained. Academic grades ranged from 0 to 10 points, although three levels of achievement were distinguished: low achievement (grades of 5 or below), medium achievement (grades between 5 and 7.5), and high achievement (grades over 7.5). Due to the dichotomous nature of the calibration measure, and the fact that the process variables did not meet normality conditions (see Table 4), regression analyses with achievement as a continuous variable were unsuitable. Thus, this distinction was made to determine whether different levels of mathematics achievement lead to different patterns of calibration accuracy, as well as to differences in the metacognitive process during TTPM.

Procedure

The study was conducted in accordance with the Helsinki Declaration of the World Medical Association (Williams 2008), which reflects the ethical principles for research involving humans. The evaluation was collectively administered during a regular class, in the class devoted to mathematics. Students performed the TTPM using their personal computers, and the assessment protocol was administered in students' regular classroom. This session was part of a broader research project, and this was the third time that examiners and students worked together. Although two trained examiners carried out the evaluations, different groups of students were assigned to each one, so that the children had only one examiner as a reference. Teachers were not present in this evaluation session. Data collection was implemented through Moodle platform. For this purpose, a multidisciplinary team including psychologists, teachers, and a computer engineer collaborated during the study.

Participants in this study were not given any incentive to take part. However, every attempt was made to design the materials and the evaluation procedure to be appealing for students (e.g., assessment protocol administered through a hypermedia environment -Moodle platform-, category system displayed by means of pop-ups, the use of graphics). Fig. 1 shows an example



Si...	Pulsa
...estás LEYENDO el ENUNCIADO	
... estás haciendo un ESQUEMA o DIBUJO	
... estás RECORDANDO otros PROBLEMAS SIMILARES	
... estás PENSANDO cómo SOLUCIONARLO	
... estás HACIENDO OPERACIONES	
... estás COMPROBANDO la SOLUCIÓN	
... CORRIGIENDO el EJERCICIO	

Fig. 1 Example of category system presentation during TTPM. This figure shows seven of the nine categories (reading, drawing/summarizing, recalling, thinking, calculating, reviewing, and correcting)

of the interface used to present the category system. In addition, the hypothetical case of a boy of their age (Alex), who tried to solve a math problem, was used to introduce the TTPM procedure and guide the training phase. Before the evaluation, students were told that it was not an exam. However, they were encouraged to do their best because the main goal of this study was to know how they solved math problems in order to help them to improve their problem-solving skills and get better grades.

The first step consisted of a training phase to help students to familiarize themselves with the system of categories and the assessment procedure. This phase started by presenting Alex and the different actions he carried out while solving a problem. After training, students performed a category-recognition test consisting of 18 multiple-choice items (2 per category) with four response alternatives. Students had to indicate the category that best expressed each proposed activity, including the "other" category. For instance, for the statement, "Alex realized he made a mistake, so he is erasing", the alternatives were: "Alex is thinking about a solution", "Alex is writing", "Alex is reviewing", or "Alex is correcting mistakes". The length of the training phase is variable and depends on students' engagement (comments, questions, etc.). However, it must never last more than 20–25 min (García et al. 2015). Students with scores below 16 points on the category-recognition test (around 90 % of correct responses) were excluded from the analyses. Originally, the sample was comprised of 567 students, of whom 43 (7.58 %) scored below the cut point. The remaining 524 students (included in this study) obtained a mean score of 16.984 (*SD*=0.493) on this test.

Because of the characteristics of the TTPM, only two math problems were used in the present study. This technique establishes that training and administration of the evaluation protocol must be conducted in the same session. Additionally, this method is designed to match the length of a regular class session (40–45 min). Thus, time has to be carefully assigned to the different TTPM components, taking into consideration that students must perform the problems without time pressure.

Before starting the TTPM administration, students performed a reaction time (RT) task, responding to an image accompanied by a tone, and presented at different time intervals (10–15 s). They had to mouse-click on the stimuli appearing on the screen as quickly as possible. A total of 30 beeps were presented. This task was aimed at familiarizing the students with the response method. Thus, an index of cognitive effort was not calculated in the present study. Students were told that the same stimuli would appear while solving the problems, after which a box with the category system would appear on the computer screen. They had to indicate the sub-process they were engaged in at each moment. During TTPM performance, the visual-auditory stimuli were presented in intervals of 40–45 s. Stimulus presentation during TTPM lasted until students indicated they had finished each problem by clicking on a "finish" button set up for this purpose. This task was designed so that, regardless of each student's response speed and the time they started or finished each problem, the time intervals between probes were the same for all the participants.

An important constraint in the Triple Task technique is the choice of the time interval with which the probe interrupts the primary task. Piolat et al. (1996) examined whether the choice of a particular time interval affects text production, cognitive effort, and temporal organization of the writing process. In their study, three groups completed the Triple Task protocol under different time intervals: a high-rate group (auditory probe every mean interval of 15 s, varying between 10 and 20 s), average-rate group (auditory probe every 30 s, varying between 15 and 45 s), and low-rate group (auditory probe every 45 s, varying between 30 and 60 s). The results showed that there were no differences either in the temporal organization of the writing processes or in text production as a function of time interval. However, secondary RTs, used as a measure of cognitive effort in their study, were affected by this variation. Specifically, longer RTs were found in both the low- and high-rate groups. Subsequently, García and Rodríguez (2007) found in their study that the longer the time interval, the better the performance in composition writing. Results from both studies suggest that high-rate patterns are more intrusive than average- and low-rate patterns in terms of cognitive effort and even of performance. These studies focused on writing composition. However, the number of phases and sub-processes analyzed and the time devoted to them may vary considerably depending on the cognitive task under investigation. In addition, excessively long time intervals may result in a loss of information, as certain processes may occur between probes. In order to reach a balance between informative capacity and the degree of interference caused by the assessment procedure, a time interval of 40–45 s was established in the present study.

Data were collectively gathered, and head-phones were given to the students to prevent them from disturbing each other. Data were collected and stored in a Moodle platform for later analyses. Students accessed this platform through an individual username and password in order to guarantee anonymity. Math problems were presented on the computer screen and on paper. Students could use the paper to write, draw, take notes, etc. They were asked to write their answer on the paper after finishing each problem. After students reported finishing the problem by pressing the "finish" button, a new box appeared on the computer screen displaying the following question: "Do you think that you have solved the problem successfully?" Responses were coded as yes (1) or no (0). Students' achievement in the math problems was established in terms of success (1) or failure (0), according to their written answers.

Once this information was coded, students were assigned to groups as a function of the correspondence between post-performance judgment and actual performance in each problem. Students who judged their performance accurately were assigned to Group 1 (Accurate), whereas those who failed in their judgments were assigned to Group 2 (Inaccurate). The number of students who reported having failed to solve the problem when they actually did solve the problem successfully (under-confident students) was about 1–2 % in both problems. This made it implausible to consider the difference between under- and over-confident groups for further analyses. Group composition varied in both problems, and separate analyses were conducted for each problem.

Process variables were based on relative frequency counts. In order to minimize the effect of students' differences in the total number of reported categories, the frequency of each category was established by dividing the election frequency of that category by the total number of elections across categories. In addition, and given the high variability shown by students in their process profiles, with large differences in frequencies across students and categories (see means and standard deviations in Table 4), frequency counts were then transformed into percentages. Relative frequencies were multiplied by 100 in order to simplify data analysis and interpretation. This high variability also made it necessary to conduct non-parametric analyses, as described below.

Data analysis

To address the proposed goals, the data were analyzed in five steps: (a) calibration, defined as the correspondence between post-performance judgments and actual performance, was calculated with Phi coefficient (ρ_ϕ This statistic was used due to the dichotomous nature of these variables (Adeyemi 2011; Schraw et al. 2012). Although more dichotomous measures of calibration are available, showing similar specificity and sensibility, Phi coefficient was used, as it is provided by SPSS statistical package and easily interpreted in terms of association or correspondence; (b) considering the dichotomous nature of the variables, McNemar's χ^2 statistic was calculated to analyze the stability of post-performance judgments and actual performance across problems. This statistic is based on a 2x2 cross classification of paired responses to a dichotomous item. Data from the concordant cells of the cross-classification table provide an indicator of stability, whereas χ^2 focuses on the change (i.e., non-concordant cells), and establishes whether this change is statistically significant (Adedokun and Burgess 2012). A p value<.05 indicates the existence of statistically significant differences between the responses given at two different moments (e.g., performance judgments in Problem 1 and 2). As this measure informs about two different aspects, it implies that high stability can be found even when the change is significant; (c) prior to analyzing differences in the process between the groups with different accuracy, dependent variables distribution were examined, and students' metacognitive process during mathematical problem-solving was described. Dependent variables included the three SRL phases (planning, execution and evaluation), and the eight TTPM categories or sub-processes (reading, drawing or summarizing, recalling similar problems, thinking about solutions, calculating, writing a response, reviewing, and correcting mistakes); (d) as dependent variable distribution did not meet normality conditions (see Table 4), non-parametric analyses were conducted to examine group differences. Specifically, the Mann-Whitney U-test was applied, using Cliff's delta (δ) as a measure of effect size (Macbeth et al. 2011). The non-parametric nature of this statistic reduces the influence of characteristics such as distribution shape, dispersion differences, and extreme values. It provides a measure of dominance, or the degree of overlapping between two distributions of scores. The value of this statistic ranges from −1 (if scores in Group 2 are larger than scores in Group 1) to +1 (if scores in Group 2 are smaller than scores in Group 1), and takes on the value of zero if the two distributions are similar (i.e., absence of significant group differences in the measured variables). Cohen (1988) established a bridge between Cohen's d and Cliff's δ statistic. In this sense, a δ value of .147 has an effect size of d=.20 (small effect); a δ value of .330 corresponds to an effect size of d=.50 (medium effect); and a δ of .474 has an effect size of d=.80 (large effect); and (e) the emergence of different calibration accuracy patterns and possible differences in the metacognitive process as a function of mathematics achievement and grade level were examined. Phi coefficient (ρ_ϕ) was calculated for each group of students (low, medium, and high achievers; fifth- and sixth-grade students) as an indicator of calibration accuracy. Differences in the metacognitive process among groups with different levels of achievement were examined through the Kruskal-Wallis statistic. As the non-parametric nature of these statistics did not allow conducting post-hoc analyses, differences between pairs of groups were analyzed using the Mann-Whitney U-test. This statistic was also used to analyze differences in the process between fifth- and sixth-grade students. Cliff's delta (δ) was calculated as a measure of effect size.

SPSS v.19 (Arbuckle 2010) was used to carry out the statistical analyses. However, as this program does not provide any non-parametric alternative to obtain an estimation of effect size, Cliff's Delta Calculator (CDC: Macbeth et al. 2011) was used to calculate Cliff's delta. A p-value≤.05 was established as criterion of statistical significance.

Results

Correspondence between post-performance judgments and actual performance

Table 2 shows the frequency distribution of post-performance judgments and actual performance, as well as the Phi correlation coefficients and their statistical significance in the two math problems. In Problems 1 and 2, a total of 221 (42.2 %) and 278 (53.1 %) students, respectively, were accurate in their post-performance judgments, whereas 303 (57.8 %) and 246 (46.9 %) were not. The correlation between the two variables (post-performance judgments and actual performance) was .207 (p<.001) in Problem 1 and .272 (p<.001) in Problem 2. Thus, a low and directly proportional relationship between post-performance

Table 2 Table of cross-classified data frequency according to correspondence between post-performance judgments and actual performance, and phi correlation (ρ_ϕ). Problems 1 and 2

		Post-performance judgments			Phi Correlation	
		Failure	Success	Total	ρ_ϕ	p
Problem 1						
Actual performance	Failure	70	298	368	.207	< .001
	Success	5	151	156		
	Total	75	449	524		
Problem 2						
Actual performance	Failure	107	232	339	.272	< .001
	Success	14	171	185		
	Total	121	403	524		

judgments and actual performance was found. Taking this relationship into consideration, there is a strong mismatch between students' post-performance judgments and their actual performance.

As Table 2 also shows, inaccurate students displayed a tendency towards over-confidence. Specifically, 298 students (56.871 %) in Problem 1 and 232 (44.275 %) in Problem 2 reported having successfully solved the problem when, in fact, they gave an incorrect response. On the other hand, the number of students who erroneously reported having failed to solve the problem (i.e., under-confident students) accounted for less than 1 % (n=5) and 2.671 % (n=14) of the sample in Problems 1 and 2, respectively. At this point, it is also important to note the students' low rates of success in solving both math problems, as only 156 students (29.771 %) gave a correct answer to Problem 1, and 185 (35.305 %) to Problem 2. These data contrast with the high perception of success shown by students, with 449 (85.687 %) and 403 (76.908 %) students reporting having successfully solved Problems 1 and 2, respectively.

Stability of post-performance judgments and actual performance

Data about students' stability in post-performance judgments and actual performance are provided in Table 3. Regarding post-performance judgments, 354 students reported solving Problems 1 and 2 successfully, whereas 26 students reported being unsuccessful in both problems. This means that a total of 380 students (72.5 % of the sample) were stable in their post-performance judgments across problems. Thus, only 27.5 % of the students showed a change in their judgments. McNemar's χ^2 statistic indicated that this change was statistically significant (p<.001). Actual performance followed a similar pattern: 252 students were unsuccessful solving Problems 1 and 2, whereas 69 solved both problems successfully. This indicates that 321 students (61.2 % of the sample) were stable in their actual performance, whereas 38.8 % were not. This change in actual performance across problems was statistically significant (p=.039).

Table 3 Table of cross-classified data frequency according to post-performance judgments and actual performance stability, and McNemar's χ^2 statistic. Problems 1 and 2

		Problem 1			McNemar's χ^2	
		Failure	Success	Total	χ^2	p
Post-performance judgments						
Problem 2	Failure	26	49	75	14.063	< .001
	Success	95	354	449		
	Total	121	403	524		
Actual performance						
Problem 2	Failure	252	116	368	3.862	.039
	Success	87	69	156		
	Total	339	185	524		

Differences in the metacognitive process shown by students with different calibration accuracy

Descriptive statistics for each TTPM sub-process and phase are shown in Table 4. High standard deviations indicated high within-subject variability in their problem-solving process. Kurtosis and skewness values confirmed the unsuitability of using parametric analyses. The presence of high within-subject variability was also observed in the total time that students spent on the problems. Specifically, students spent a mean of 376.794 s (*SD*=151.491) on Problem 1, and 339.084 s (*SD*=125.626) on Problem 2.

Table 4 also provides information about students' metacognitive process before differences in calibration accuracy were considered. As can be observed, profiles are quite similar in both problems. Students reported spending a greater amount of time on calculations in comparison to the rest of sub-processes. Within the planning phase, students reported spending a large amount of time thinking about solutions, in contrast to other sub-processes such as recalling similar problems (i.e., previous knowledge and experiences) or drawing/summarizing (i.e., organization of the information). Finally, the sub-processes of the evaluation phase (i.e., reviewing and correcting mistakes) were the least frequently reported by students in both problems.

With regard to differences between groups with different accuracy, the Mann-Whitney *U*-test indicated the existence of statistically significant differences in some sub-processes and phases. Regarding Problem 1, differences in the sub-processes of drawing or summarizing (U=29143.500, p=.006, δ=.129), writing (U=29760.000, p=.026, δ=−.111), and reviewing (U=29301.000, p=.006, δ=−.125) were found, as well as in the phases of planning (U=29807.500, p=.032, δ=.110) and evaluation (U=28340.500, p=.001, δ=−.153). Means in Table 5 show that students who were accurate in their post-performance judgments reported drawing or summarizing (and also planning) more frequently than their inaccurate peers. Accurate students also reported having spent less time on writing, reviewing, and on the evaluation phase than their inaccurate peers. In Problem 2, a similar pattern of results was

Table 4 Descriptive statistics for dependent variables (TTPM variables). Problems 1 and 2

Process categories and phases	Problem 1			Problem 2		
	M (SD)	Kurtosis	Skewness	M (SD)	Kurtosis	Skewness
Reading	16.438 (15.109)	5.871	1.952	13.454 (14.57)	8.040	2.232
Drawing or summarizing	11.984 (17.352)	2.507	1.663	13.383 (21.138)	3.043	1.868
Recalling similar problems	3.022 (6.697)	8.160	2.693	2.471 (6.510)	16.305	3.553
Thinking about a solution	19.175 (17.181)	1.250	1.111	18.614 (19.467)	2.110	1.439
Calculating	27.616 (22.832)	−.314	.646	30.973 (25.298)	−.847	.438
Writing	11.746 (12.066)	1.942	1.280	13.767 (14.270)	2.435	1.375
Reviewing	6.299 (9.230)	2.319	1.566	4.229 (7.683)	4.877	2.052
Correcting mistakes	3.763 (7.955)	8.431	2.658	3.251 (7.888)	17.255	3.498
Planning	50.637 (23.950)	−.632	.111	47.872 (27.247)	−.860	.327
Execution	39.351 (23.108)	−.449	.291	44.729 (26.452)	−.887	−.005
Evaluation	10.063 (13.121)	2.590	1.504	7.473 (11.936)	6.093	2.141

M Mean of the group in each process variable. It refers to the mean frequency of election of each sub-process and phase divided by the total number of elections and expressed in percentages

Table 5 Means (M) and Standard Deviations (SD) of the groups with different calibration accuracy in TTPM variables. Problems 1 and 2

Process categories and phases	Problem 1		Problem 2	
	Accurate *M* (*SD*)	Inaccurate *M* (*SD*)	Accurate *M* (*SD*)	Inaccurate *M* (*SD*)
Reading	15.895 (15.354)	16.835 (14.941)	13.554 (15.376)	13.341 (13.653)
Drawing or summarizing	14.257 (18.276)	10.326 (16.480)	16.082 (22.906)	10.333 (18.521)
Recalling similar problems	3.601 (7.604)	2.600 (5.927)	2.399 (5.627)	2.552 (7.394)
Thinking about a solution	19.778 (18.213)	18.736 (16.405)	18.960 (19.642)	18.223 (19.300)
Calculating	27.977 (23.054)	27.353 (22.703)	31.370 (25.350)	30.524 (25.284)
Writing	10.479 (11.638)	12.670 (12.306)	11.352 (12.170)	16.495 (15.910)
Reviewing	4.972 (8.112)	7.267 (9.867)	3.712 (7.223)	4.813 (8.147)
Correcting mistakes	2.972 (6.329)	4.339 (8.924)	2.762 (7.055)	3.804 (8.717)
Planning	53.629 (23.955)	48.455 (23.749)	50.920 (27.872)	44.426 (26.153)
Execution	38.448 (23.357)	40.009 (22.941)	42.694 (26.842)	47.028 (25.866)
Evaluation	7.945 (11.547)	11.607 (13.978)	6.467 (10.991)	8.609 (12.850)

M Mean of the group in each process variable refers to the mean frequency of election of each sub-process and phase divided by the total number of elections and expressed in percentages

Accurate group Problem 1 (n=221); Inaccurate group Problem 1 (n=303); Accurate group Problem 2 (n=278); Inaccurate group Problem 2 (n=246)

observed. Statistically significant differences were found in the sub-processes of drawing or summarizing (U=29101.000, p=.001, δ=.149) and writing (U=28067.000, p=.001, *Cliff's* δ= −.179), and in the phases of planning (U=29453.000, p=.006, δ=.139) and evaluation (U= 31088.500, p=.042, δ=−.098). Although differences in correcting mistakes were not statistically significant, a lower proportion of accurate students reported doing this than students in the inaccurate group in both problems.

Differences in calibration and the metacognitive processes as a function of math achievement level

Table 6 shows the existence of different patterns of correspondence between post-performance judgments and actual performance as a function of students' achievement level (low-medium-high). As can be observed, calibration accuracy improved progressively as students' achievement level increased. This change was more evident in Problem 2, where ρ_ϕ reached values near to .50 in the high-achieving group. This improvement in calibration accuracy was reflected in over-confidence rates. Specifically, 124 (61.083 %) low achievers made over-confident judgments in Problem 1, whereas only 64 (45.714 %) high achievers did so. In Problem 2, the number of over-confident students ranged from 114 (56.157 %) to 37 (26.428 %) in the groups of low and high achievers, respectively.

Regarding group differences in the metacognitive process, the Kruskal-Wallis statistic revealed statistically significant differences in the sub-processes of reading, *K-W* (χ^2)=9.332, p=.009 in Problem 1; and *K-W* (χ^2)=16.243, p<.001 in Problem 2, and drawing/summarizing, *K-W* (χ^2)=6.904, p=.032 in Problem 1; and *K-W* (χ^2)=15.636, p<.001 in Problem 2. High-achieving students in mathematics reported having read less, but

Table 6 Patterns of calibration accuracy as a function of achievement level in math. Problems 1 and 2

			Post-performance judgments			Phi Correlation	
			Failure	Success	Total	ρ_ϕ	p
Problem 1							
Low achievers	Actual performance	Failure	28	124	152	.151	.031
		Success	3	48	51		
		Total	31	172	203		
Medium achievers	Actual performance	Failure	21	110	131	.162	.030
		Success	2	48	50		
		Total	23	158	181		
High achievers	Actual performance	Failure	21	64	85	.338	<.001
		Success	0	55	55		
		Total	21	119	140		
Problem 2							
Low achievers	Actual performance	Failure	45	114	159	.162	.021
		Success	5	39	44		
		Total	50	153	203		
Medium achievers	Actual performance	Failure	32	81	113	.232	.002
		Success	6	62	68		
		Total	38	143	181		
High achievers	Actual performance	Failure	30	37	67	.479	<.001
		Success	3	70	73		
		Total	33	107	140		

using representation strategies more frequently than their low-achieving peers (see Table 7). Mann-Whitney's U-test showed that: (a) the groups with low and medium achievement only differed in the sub-process of drawing/summarizing in Problem 1 (U=16332.500, p=.037, δ=−.111); (b) differences in this sub-process were also found between the groups with medium and high achievement in both problems (Problem 1: U=11007.000, p=.032, δ=−.131; Problem 2: U=10363.500, p=.003, δ=−.182); and (c) the groups with low and high achievement showed the strongest differences in both problems. They differed in drawing/summarizing (Problem 1: U=10938.500, p<.001, δ=−.230; Problem 2: U= 11128.000, p<.001, δ=−.216), as well as in reading (Problem 1: U=11537.500, p=.003, δ=.188; Problem 2: U=11961.500, p=.012, δ=.158). Effect sizes, although higher than in previous analyses, were low.

Differences in calibration and metacognitive processes as a function of grade level

Table 8 shows patterns of calibration accuracy as a function of grade level (fifth and sixth grade) in Problems 1 and 2. As can be observed, calibration accuracy was low, especially in the case of sixth-grade students in Problem 1 (ρ_ϕ=.114). In contrast to mathematics achievement, calibration accuracy did not present a distinguishable pattern of change as a function of grade level. As in previous analyses, calibration patterns were characterized by a marked tendency towards over-confidence. Specifically, 123 (55.909 %) and 104 (47.272 %) fifth-grade

Table 7 Means (M) and Standard Deviations (SD) of the groups with different math achievement levels in TTPM variables. Problems 1 and 2

Process categories and phases	Problem1			Problem 2		
	Low achievers *M* (*SD*)	Medium achievers *M* (*SD*)	High achievers *M* (*SD*)	Low achievers *M* (*SD*)	Medium achievers *M* (*SD*))	High achievers *M* (*SD*)
Reading	19.492 (18.185)	15.243 (12.797)	13.557 (11.935)	16.049 (17.326)	12.314 (12.546)	11.164 (11.953)
Drawing or summarizing	8.536 (13.944)	11.966 (17.176)	17.007 (20.630)	11.640 (21.413)	11.668 (19.196)	18.128 (22.513)
Recalling similar problems	3.354 (7.285)	2.773 (6.541)	2.864 (5.999)	2.773 (7.256)	2.690 (6.720)	1.750 (4.874)
Thinking about a solution	20.182 (18.765)	18.060 (15.607)	19.157 (16.747)	19.335 (20.625)	18.729 (19.362)	17.42 1 (17.890)
Calculating	25.753 (23.003)	29.165 (22.451)	28.314 (23.044)	28.310 (25.044)	33.397 (25.542)	31.700 (25.163)
Writing	12.083 (12.241)	13.044 (13.661)	9.578 (9.014)	13.635 (14.618)	14.198 (15.929)	13.400 (11.256)
Reviewing	6.315 (8.827)	6.243 (9.368)	6.350 (9.675)	4.453 (8.131)	3.966 (7.363)	4.242 (7.454)
Correcting mistakes	4.305 (8.587)	3.591 (7.673)	3.2000 (7.344)	3.891 (9.785)	3.198 (6.469)	2.392 (6.281)
Planning	51.605 (24.799)	48.027 (22.930)	52.607 (23.864)	49.788 (27.295)	45.331 (27.229)	48.378 (27.137)
Execution	37.842 (23.703)	42.198 (22.288)	37.857 (23.092)	41.965 (26.025)	47.580 (26.694)	45.050 (26.529)
Evaluation	10.615 (13.851)	9.828 (12.910)	9.564 (12.347)	8.330 (13.659)	7.160 (10.517)	6.635 (10.946)

M Mean of the group in each process variable refers to the mean frequency of election of each subprocess and phase divided by the total number of elections and expressed in percentages. Low achievers (n = 203); Medium achievers (n = 181); High achievers (n = 140)

Table 8 Patterns of calibration accuracy as a function of grade level. Problems 1 and 2

			Post-performance judgments			Phi Correlation	
			Failure	Success	Total	ρ_ϕ	*p*
Problem 1							
Fifth graders	Actual performance	Failure	47	123	170	.283	<.001
		Success	0	50	50		
		Total	47	173	220		
Sixth graders	Actual performance	Failure	23	175	198	.114	.047
		Success	5	101	106		
		Total	28	276	304		
Problem 2							
Fifth graders	Actual performance	Failure	57	104	161	.265	<.001
		Success	5	54	59		
		Total	62	158	220		
Sixth graders	Actual performance	Failure	50	128	178	.261	<.001
		Success	9	117	126		
		Total	59	245	304		

students were over-confident in Problems 1 and 2, respectively. Sixth-grade students were over-confident in 175 (57.565 %) and 128 (42.105 %) of the cases, respectively.

Regarding group differences in TTPM sub-processes and phases, means in Table 9 indicate that both groups showed a similar metacognitive process. This was confirmed by the Mann-Whitney *U*-test. Specifically, statistically significant group differences were only found in

Table 9 Means (M) and Standard Deviations (SD) of fifth- and sixth-grade students in TTPM variables. Problems 1 and 2

Process categories and phases	Problem 1		Problem 2	
	Fifth graders *M* (*SD*)	Sixth graders *M* (*SD*)	Fifth graders *M* (*SD*)	Sixth graders *M* (*SD*)
Reading	16.886 (14.961)	16.115 (15.232)	13.040 (13.960)	13.753 (15.026)
Drawing or summarizing	11.904 (18.368)	12.042 (16.609)	13.313 (22.482)	13.434 (20.148)
Recalling similar problems	3.668 (7.268)	2.555 (6.222)	2.700 (7.195)	2.305 (5.972)
Thinking about a solution	20.281 (19.040)	18.375 (15.684)	19.559 (19.222)	17.930 (19.646)
Calculating	25.677 (24.166)	29.019 (21.748)	29.568 (26.278)	31.990 (24.558)
Writing	11.600 (12.922)	11.852 (11.427)	13.727 (14.643)	13.796 (14.018)
Reviewing	5.231 (8.822)	7.072 (9.453)	4.509 (8.519)	4.026 (7.025)
Correcting mistakes	4.781 (9.766)	3.026 (6.246)	3.622 (8.812)	2.983 (7.150)
Planning	52.781 (25.611)	49.085 (22.589)	48.636 (28.196)	47.319 (26.572)
Execution	37.268 (24.562)	40.858 (21.913)	43.286 (27.024)	45.773 (26.025)
Evaluation	10.013 (14.509)	10.098 (12.043)	8.127 (13.447)	7.000 (10.709)

M Mean of the group in each process variable refers to the mean frequency of election of each sub-process and phase divided by the total number of elections and expressed in percentages

Fifth graders (*n*=220); Sixth graders (*n*=304)

Problem 1, in the sub-processes of calculating (U=29533.500, p=.022, δ=−.117), and reviewing (U=29414.500, p=.008, δ=−.120). Sixth-grade students reported calculating and reviewing more frequently than fifth graders in this problem. However, the same pattern of results was not found in Problem 2. Effect sizes for the differences were low.

Discussion and conclusions

The present study was aimed at examining patterns of post-performance calibration and their relationship with the metacognitive process shown by a broad sample of elementary school students while solving two math problems. In order to evaluate the process, the *Triple Task Procedure in Mathematics* (TTPM: García and González-Pienda 2012; García et al. 2015) was administered to 524 fifth- and sixth-grade students in a regular mathematics class. Once the TTPM was completed, the students were asked to judge the results of their performance in terms of success or failure. These judgments were then compared with their actual performance, also expressed in dichotomous terms. The possible influence of students' mathematics achievement and grade level on calibration accuracy patterns and the metacognitive process was analyzed. Separate analyses were conducted for each problem. Results are discussed according to the five initial predictions:

Correspondence between post-performance judgments and actual performance

As hypothesized, students were poorly calibrated in the present study, with a strong tendency towards over-confidence. Nearly 60 and 45 % of the sample in Problems 1 and 2, respectively, reported having successfully solved the math problems when they actually gave an incorrect answer. This finding is coherent with previous studies in mathematics and problem-solving, conducted in elementary school and higher educational stages (Bol et al. 2005, 2010; Hacker et al. 2008b; Özsoy 2012).

Stability of post-performance judgments and actual performance

Previous studies suggest that performance judgments (and biases) are stable over time, and more stable than actual performance (Bol et al. 2012; Dinsmore and Parkinson 2013; Hacker et al. 2008b; Stolp and Zabrucky 2009). In this sense, results from the present study tentatively support this statement, as well as the second hypothesis of the study. Specifically, post-performance judgments and actual performance were both quite stable across problems, with about 70 % of students showing stability in their judgments, and 60 % in their performance. However, as will be discussed later in the present study, the use of only two math problems to estimate stability makes it difficult to generalize these findings.

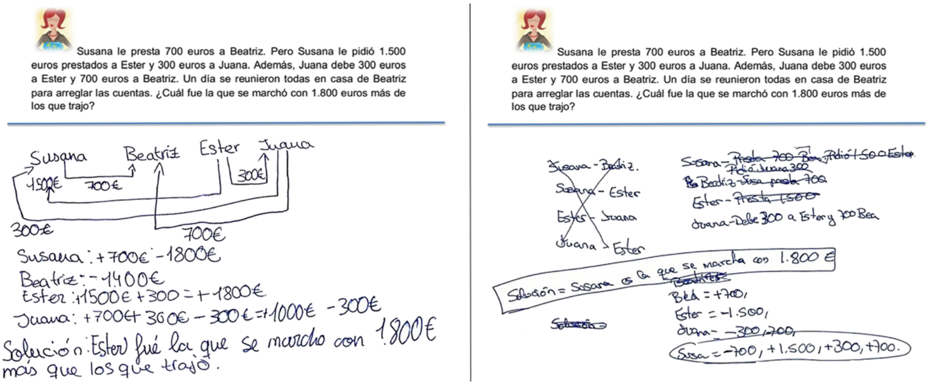
Differences in the metacognitive process shown by students with different calibration accuracy

Descriptive analyses of the metacognitive process during TTPM revealed that students in the overall sample spent a great amount of time performing calculations, but little time organizing or representing information (drawing/summarizing), and even less time reviewing the progress and results of their performance. In addition, the time they spent recalling similar problems (i.e., seeking previous knowledge) was almost negligible in general. These findings indicate an absence of effective metacognitive skills in this group of students, which would made it difficult for them to monitor their own solution process or use different strategies flexibly (e.g., representing problem situations in many forms, such as a graph, a sketch or a table), getting stuck in familiar procedures instead (i.e., calculations). This is coherent with some previous studies pointing out that students at these educational stages tend to show poor metacognitive strategies while solving math problems, which commonly leads to poor performance and low calibration accuracy (Cleary and Chen 2009; Desoete and Roeyers 2006; Kramarski and Gutman 2006). The low rates of success solving the math problems, and the weak correspondence between post-performance judgments and actual performance shown by students in the present study confirmed this pattern of results.

Additionally, differences between accurate and inaccurate students in the metacognitive process during TTPM were found. Specifically, accurate students reported spending significantly more time planning (especially drawing and summarizing) than their inaccurate peers. Accurate students also spent significantly less time writing and reviewing than the inaccurate group. These results initially confirmed the third hypothesis of the present study, but only partially. Thus, accurate students planned more and used different information representation strategies, but they did not evaluate their progress and results more than low calibrated students did. However, the finding that inaccurate students wrote and reviewed significantly more, along with the fact that they also reported correcting (and presumably making) more mistakes than their accurate peers—although this last variable did not generate statistically significant differences—indicates the use of "trial-and-error" mechanisms by this group of students. These mechanisms involve a tendency towards tackling a problem with various methods until a solution is found, and have been reported to be frequent during problem-solving tasks (Callander 2011; Clements et al. 2008; Dijk et al. 2003). A similar pattern of results was found in both problems. Although effect sizes were low in general, these findings initially suggest a relationship between planning strategies (mainly representation and organization of the information) and making accurate post-performance judgments. In fact, there is an inverse relationship between these strategies and the use of trial-and-error as a means to solve the problems.

Evidence from students' metacognitive processes while solving the math problems could be obtained by analyzing performance artifacts (e.g., drawings, notes, or editing from students' drafts on paper). An example of these artifacts is provided below (Fig. 2), corresponding to the performance of two different children in Problem 1 (two 11-year-old boys, the first boy belonging to the accurate group and the second one to the inaccurate group). This figure represents an example of organization of the information and trial-and-error mechanisms, respectively. As can be observed, the student who judged his performance accurately (left side of the image) used different strategies to organize the information, establishing clear relationships between data and facts. He also seems to have tried only one approximation to solve the problem, as a lack of mistakes or corrections indicate. Finally, he successfully solved the problem, giving a unique and correct answer (i.e., "Esther"). Regarding the student who was inaccurate in his performance judgment (right side of the image), his problem-solving process was less smooth than in the former case, with some signs of correcting and editing. A lack of information-organization strategies can be observed, and data and relationships are rather chaotic. Although the student gave a unique answer (i.e., "Susana"), it was incorrect.

This pattern of results indicates the presence of some degree of self-regulatory and metacognitive mechanisms in calibrated students. As Jacobse and Harskamp (2012) pointed



Susana le presta 700 euros a Beatriz. Pero Susana le pidió 1.500 euros prestados a Ester y 300 euros a Juana. Además, Juana debe 300 euros a Ester y 700 euros a Beatriz. Un día se reunieron todas en casa de Beatriz para arreglar las cuentas. ¿Cuál fue la que se marchó con 1.800 euros más de los que trajo?

Susana
Beatriz
Ester
Juana
1500€
700€
300€
300€
700€

Susana: +700€ -1800€
Beatriz: -1400€
Ester: +1500€ +300 = +1800€
Juana: +700€+ 300€ - 300€ = +1000€ -300€

Solución: Ester fué la que se marcho con 1800€ más que los que trajo.

Susana le presta 700 euros a Beatriz. Pero Susana le pidió 1.500 euros prestados a Ester y 300 euros a Juana. Además, Juana debe 300 euros a Ester y 700 euros a Beatriz. Un día se reunieron todas en casa de Beatriz para arreglar las cuentas. ¿Cuál fue la que se marchó con 1.800 euros más de los que trajo?

Susana - ~~Presta 700 Bea~~ ~~Pidió 1.500 Ester.~~
~~Pidió Juana 300~~
~~Be~~ Beatriz - ~~Susa presta 700~~
Ester - ~~Presta 1.500~~
Juana - Debe 300 a Ester y 700 Bea

Solución = Susana es la que se marcha con 1.800 €

~~Beatriz~~
Bea = +700,
Ester = -1.500,
Juana = -300, -700,
Susa = -700, +1.500, +300, +700.

~~Solución =~~

Fig. 2 Example of students' artifacts during TTPM (Problem 1). Two 11-year-old students: accurate (*left*) and inaccurate (*right*)

out, an important cognitive sub-process in mathematical problem-solving is building a representation of the problem situation. However, few elementary school students use this strategy spontaneously. The use of schematic visualizations allows students to establish those much-needed relationships between variables that help to solve the problem. For these authors, making these kinds of representations can be interpreted as an expression of sophisticated metacognitive regulation mechanisms, which provide insight into the episodes of analyzing and exploring a problem (i.e., metacognitive regulation in the first episodes of the problem-solving process). Moreover, schematic representations have been shown to predict students' successful problem-solving (Fagnant and Vlassis 2013).

Differences in calibration and the metacognitive process as a function of math achievement level

Different patterns in calibration accuracy, as well as differences in the metacognitive process during TTPM, emerged as a function of mathematics achievement level. These results confirmed the fourth hypothesis of the present study. Regarding calibration accuracy, high-achieving students made more accurate judgments than their peers with lower levels of achievement. Specifically, the higher the achievement level, the more precise the post-performance judgments, and the lower the tendency towards over-confidence shown by students. The same pattern of results was obtained in both problems. These findings are consistent with previous studies showing that high-achieving students are commonly more accurate in their judgments of performance (and somewhat under-confident) than their low-achieving peers (Bol et al. 2005, 2010; Hacker et al. 2008b; Özsoy 2012). However, no conclusions about under-confidence were reached in the present study, as the proportion of under-confident students was negligible (about 1–2 %). This could be related to the fact that most students in the current sample were low achievers, as problem-solving scores evidenced. Results also indicated the existence of some differences in the metacognitive process during TTPM as a function of achievement level. In this sense, previous differences between accurate and inaccurate students in drawing/summarizing are at least partially explained by this variable. This is not surprising, as high achievers have been referred to as being more highly calibrated and self-regulated than low achievers (Bol et al. 2005; Dunlosky and Rawson 2012;

Nietfeld et al. 2005, 2006). The low effect sizes found in the previous analyses may also be a consequence of the influence of this variable.

Differences in calibration and the metacognitive process as a function of grade level

With regard to grade level, results did not allow to confirm the fifth hypothesis of the present study. First, no distinguishable patterns of calibration accuracy were obtained as a function of this variable. In general, the overall sample showed low calibration accuracy. Additionally, sixth-grade students' calibration accuracy was considerably lower in the first problem. Second, regarding the metacognitive process during TTPM, statistically significant differences between fifth- and sixth-grade students were mainly found in reviewing, in Problem 1. In this sense, sixth-grade students reported reviewing more than their fifth-grade peers. These differences could be due to the fact that sixth-grade students were considerably less accurate in this first problem. Thus, these results reflect the previously described differences between accurate and inaccurate students in this sub-process. This pattern of results shows that grade level did not influence calibration accuracy or metacognitive processes in this sample of students. This finding could be related to the fact that both school grades are very close in time and constitute a unique educational cycle in the current Spanish Educational System. This makes it difficult to observe developmental or cognitive differences between fifth- and sixth-grade students. However, as there is also a good deal of evidence suggesting that low calibration accuracy and over-confident judgments are present across different educational stages, an absence of differences can also be expected.

Implications

The main implication of the present study is related to the impact of inaccurate performance judgments on learning processes and outcomes, and how these self-evaluation biases can be corrected. In this sense, as students show a strong tendency towards over-confidence (also corroborated in the present study), most research has focused on whether this positive bias is adaptive or maladaptive (Dupeyrat et al. 2011). Results are mixed. On the one hand, based on Bandura's social cognitive theory (Bandura 1986), positive self-evaluation bias is adaptive because it engages motivation, persistence in the face of failures, and protects against negative emotions such as anxiety (Stolp and Zabrucky 2009). On the other hand, authors such as Bouffard and Narcis (2011), Narciss et al. (2011), and Schunk (2008) emphasize the negative effects of unrealistic judgments. There is also some evidence that positive illusions are either adaptive or maladaptive depending on variables such as the domain or the temporariness of the related outcomes (Bouffard et al. 2011; Gonida and Leondari 2011; Gramzow et al. 2003). This debate has prompted important efforts to improve calibration accuracy, many of them from the perspective of metacognition and SRL.

Some of these studies have shown that gains in calibration accuracy are difficult to achieve or are not durable (Bol and Hacker 2001; Bol et al. 2005; Nietfeld et al. 2005), whereas others have reported some improvement (DiGiacomo 2014; Hacker et al. 2008b; Nietfeld et al. 2006; Zimmerman et al. 2011). These studies agree that more comprehensive and adapted interventions must be designed and implemented. For this purpose, and as self-evaluation biases are present at all academic stages, the first step consists of analyzing the factors involved in making performance judgments at early educational stages. Results from the present study suggest the usefulness of process-based measures, specifically the Triple Task Procedure in

Mathematics (García and González-Pienda 2012; García et al. 2015), to study the metacognitive mechanisms involved in post-performance judgments of math problem-solving. This procedure also reflected the influence of the mathematics achievement level on calibration accuracy and the metacognitive process itself. This aspect is relevant because the relationship between these components has been shown to have important implications for intervention (Hacker et al. 2000). These authors analyzed the effectiveness of an intervention based on feedback, practice tests, and course instruction to improve calibration accuracy and academic performance in a sample of undergraduate students. The results revealed that calibration accuracy improved, but only for higher achieving students. Similar results were obtained by Nietfeld et al. (2006) and Hacker et al. (2008b). These studies suggest that intervention strategies must be adapted to students' characteristics, and achievement level is an important factor moderating the effects of the intervention.

Limitations

Finally, some limitations in the present study must be acknowledged. First, the low effect sizes found call for some degree of caution concerning the scope of the findings. This result could be related to the high inter-subject variability observed in the students' metacognitive process. These low effect sizes may also indicate that other variables account for differences in the metacognitive process. This should be addressed in further studies, establishing more homogeneous groups based on additional variables. Second, the use of only two math problems must also be considered. As explained before, the features of the assessment method used (TTPM) justified this decision. However, including more problems would lead to a clearer measure of post-performance judgments and actual performance stability. Similarly, and given the mentioned time constraints, the possibility of administering the problems across days must be considered. Taking into account that the dichotomous nature of post-performance judgments and actual performance may be a limitation in the present study, more continuous measures of calibration could be obtained by increasing the number of math problems. In this same line, retaining the classical measure of cognitive effort in future studies would help to establish the appropriate number of problems by examining the extent to which they are challenging for students. Third, another limitation in the present study is that the TTPM may be intrusive for students. Specifically, it may lead to a "reactivity" effect (Bowles and Leow 2005), that is, that the type of measure employed triggers changes in learners' cognitive and metacognitive processes while performing the task. This effect has been substantially studied in other on-line measures, mainly Think-aloud protocols (Bannert and Mengelkamp 2008; Ericsson and Simon 1996; Fox et al. 2011; Veenman et al. 1993). These authors concluded that, although the use of this sort of measure may be related to an increase in the time to complete the task, simply instructing participants to verbalize their thoughts during a task does not alter the sequence of the cognitive processes or task performance. In the same line, Kellog (1987), using Triple Task procedures, reached similar results. This author found no differences in written fluency or quality between writers engaged in directed retrospection and 30 control writers who made a written composition as a single task, suggesting that directed introspection in Triple Task was not reactive. However, the use of different tasks (i.e., math problems in the present study) must be considered. In this sense, whereas there is extensive tradition in the study of mathematics problem-solving processes though Think-aloud protocols (see Montague et al. 2011), the flexibility of the Triple Task technique and its suitability for larger samples motivated its application in the present study. Given its design, however, it may be more intrusive for students than Think-aloud. In order to reduce its possible effect on students' performance, the TTPM places special emphasis on the training phase so that the categorization task (directed introspection) becomes more automatic, generating less interference and involving less cognitive effort. Additional research is being conducted to examine the possible reactivity of this technique. Fourth, students were not provided with any incentive for taking part in the study. This is relevant because students' performance was low. Providing incentives in the form of prizes or extra credits may increase motivation, improving performance. In fact, motivation has been shown to be important for students to engage in and successfully complete a task, as well as to evaluate their performance (Sheldrake et al. 2014). This is relevant, as Zimmerman's SRL model (2000, 2008) includes cognitive, behavioural, emotional, and motivational components. Thus, exploring motivation and its relationship to calibration accuracy and performance is an aspect to be considered in future studies. Fifth, the categorization of the variable mathematics achievement in different levels must be considered. The dichotomous measure of calibration, and mainly the non-parametric nature of the process variables, with high levels of variability among values, made it unfeasible to conduct regression analyses considering this variable as continuous. Thus, in order to obtain information about the influence of mathematics achievement in the mentioned variables, different levels of achievement were established. This decision may have led to a loss of statistical power, given the lower effect sizes found in the present study. In this sense, additional studies focused on specific sub-processes or phases would help to reduce variability, making it possible to conduct regression analyses properly. Finally, although this study was initially aimed at increasing ecological validity, some contextual aspects such as the fact that teachers were not present during the evaluation, or the above-mentioned absence of incentives, may have compromised this attempt. Future studies will be conducted in order to address these issues.

Acknowledgments Grants awarded to the authors from the Spanish Ministry of Science and Innovation (Ref.: EDU2010-19798) and the "Severo Ochoa" program of the Principality of Asturias (Ref.: 11–067) supported this work. We also thank Stephen J. Loew and Kenneth Watson for their assistance in the preparation of this manuscript.

Conflict of interest The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Ackerman, P. L., & Wolman, S. D. (2007). Determinants and validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied, 13*, 57–78.

Adedokun, O. A., & Burgess, W. D. (2012). Analysis of paired dichotomous data: a gentle introduction to the McNemar test in SPSS. *Journal of Multidisciplinary Evaluation, 8*(17), 125–131.

Adeyemi, O. (2011). Measures of association for research in educational planning and administration. *Research Journal of Mathematics and Statistics, 3*(3), 82–90.

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

Arbuckle, J. L. (2010). *SPSS (version 19.0) [Computer program]*. Chicago: SPSS.

Azevedo, R., & Aleven, V. (2013). *International handbook of metacognition and learning technologies* (Springer International Handbooks of Education, Vol. 28). New York: Springer. doi:10.1007/978-1-4419-3_11.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs: Prentice-Hall.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalization method affect learning? *Metacognition and Learning, 3*, 39–58.

Boekaerts, M. (1999). Motivated learning: studying student situation transactional units. *European Journal of Psychology of Education, 14*(1), 41–55. doi:10.1007/bf03173110.

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration measures in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20*(4), 372–382. doi:10.1016/j.learninstruc.2009.03.002.

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*, 133–151.

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*, 269–290. doi:10.3200/JEXE.73.4.269-290.

Bol, L., Riggs, R., Hacker, D. J., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education, 21*, 81–96.

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology, 37*, 280–287. doi:10.1016/j.cedpsych.2012.02.004.

Boston, M. D., & Smith, M. S. (2009). Transforming secondary mathematics teaching: increasing the cognitive demands of instructional tasks used in teachers' classrooms. *Journal for Research in Mathematics Education, 40*(2), 119–156.

Bouffard, T., & Narciss, S. (2011). Benefits and risks of positive bias in self-evaluation of academic competence. *International Journal of Educational Research, 50*(4), 205–256.

Bouffard, T., Vezeau, C., Roy, M., & Lengelé, A. (2011). Stability of biases in self-evaluation and relations to well-being among elementary school children. *International Journal of Educational Research, 50*, 221–229. doi:10.1016/j.ijer.2011.08.003.

Bowles, M. A., & Leow, R. P. (2005). Reactivity and type of verbal report in SLA research methodology. *Studies in Second Language Acquisition, 27*, 415–440.

Bransford, J. D., & Stein, B. S. (1993). *The ideal problem solver: A guide for improving thinking, learning and creativity* (2nd ed.). New York: W.H. Freeman.

Butler, R. (2011). Are positive illusions about academic competence always adaptive, under all circumstances? New results and future directions. *International Journal of Educational Research, 50*, 251–256. doi:10.1016./j.ijer.2011.08.001.

Callander, S. (2011). Searching and learning by trial and error. *American Economic Review, 101*(6), 2277–2308. doi:10.1257/aer.101.6.2277.

Cleary, T. J. (2009). Monitoring trends and accuracy of self-efficacy beliefs during interventions: advantages and potential applications to school-based settings. *Psychology in the Schools, 46*, 154–171. doi:10.1002/pits.20360.

Cleary, T. J., & Chen, P. (2009). Self-regulation, motivation, and math achievement in middle school: variations across grade level and math context. *Journal of School Psychology, 47*, 291–314. doi:10.1016/j.jsp.2009.04.002.

Clements, D. H., Sarama, J., & Liu, X. (2008). Development of a measure of early mathematics achievement using the Rasch model: the research-based early maths assessment. *Educational Psychology, 28*(4), 457–482. doi:10.1080/01443410701777272.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic.

Desoete, A., & Roeyers, H. (2006). Metacognitive macro-evaluations in mathematical problem-solving. *Learning and Instruction, 16*, 12–25.

DiGiacomo, G. (2014). Enhancing self-monitoring and self-reflection through a self-regulatory skills intervention embedded in a middle school mathematics curriculum. *Dissertations and Theses, 2014-Present.* Paper 201. http://works.gc.cuny.edu/etd/201.

Dijk, I. M. A. W., Van Oers, B., & Terwel, J. (2003). Providing or designing? Constructing models in primary maths education. *Learning and Instruction, 13*, 53–72. doi:10.1016/S0959-4752(01)00037-8.

Dinsmore, D., Alexander, P., & Loughlin, S. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*, 391–409.

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. doi:10.1016/j.learninstruc.2012.06.001.

Dunlosky, J., & Rawson, K. A. (2012). Over-confidence produces underachievement: inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280. doi:10.1016/j.learninstruc.2011.08.003.

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

Dupeyrat, C., Escribe, C., Huet, N., & Régner, I. (2011). Positive biases in self-assessment of mathematics competence, achievement goals, and mathematics performance. *International Journal of Educational Research, 50*(4), 241–250. doi:10.1016/j.ijer.2011.08.005.

Efklides, A., & Misailidi, P. (2010). *Trends and prospects in metacognition research*. New York: Springer.

Ericsson, K. A., & Simon, H. A. (1996). *Protocol analysis: Verbal reports as data* (Revisedth ed.). Cambridges: MIT Press.

Fagnant, A., & Vlassis, L. (2013). Schematic representations in arithmetical problem-solving: analysis of their impact on grade 4 students. *Educational Studies in Mathematics, 84*(1), 149–168.

Finn, B., & Metcalfe, J. (2014). Over-confidence in children's multi-trial judgments of learning. *Learning and Instruction, 32*, 1–9. doi:10.1016/j.learninstruc.2014.01.001.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*, 316–344. doi:10.1037/a0021663.

García, T., & González-Pienda, J. A. (2012). Evaluación del proceso de aprendizaje autorregulado en el área de las matemáticas mediante pizarras digitales. [Assessment of the self-regulated learning process in Mathematics through electronic blackboards]. In J. Dulac & C. Alconada (Eds.), *(Coords.), III Congreso pizarra digital. Publicación de comunicaciones* (pp. 105–117). Madrid: Pluma y Arroba.

García, J. N., & Rodríguez, C. (2007). Influence of the recording interval and a graphic organizer on the writing process/product and on other psychological variables. *Psicothema, 19*(2), 198–205.

García, T., Betts, L., González-Castro, P., González-Pienda, J. A., & Rodríguez, C. (2015). On-line assessment of the process involved in Maths problem-solving in fifth and sixth grade students: Self-regulation and achievement. *Revista Latinoamericana de Investigación en Matemática Educativa.*

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist, 21*, 99–120.

Gonida, E. N., & Leondari, A. (2011). Patterns of motivation among adolescents with biased and accurate self-efficacy beliefs. *International Journal of Educational Research, 50*(4), 209–211. doi:10.1016/j.ijer.2011.08.002.

Gramzow, R. H., Elliot, A. J., Asher, E., & McGregor, H. A. (2003). Self-evaluation bias and academic performance: some ways and some reasons why. *Journal of Research in Personality, 37*, 41–61.

Hacker, D. J., & Bol, L. (2004). Metacognitive theory: Considering the social-cognitive influences. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited* (Research on sociocultural influences on motivation and learning, Vol. 4, pp. 275–297). Greenwich: Information Age.

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

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

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

Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: examining the nature of judgments of confidence. *Learning and Instruction, 24*(12), 37–47. doi:10.1016/j.learninstruc.2012.10.001.

Hattie, J. (2013). Calibration and confidence: where to next? *Learning and Instruction, 24*, 62–66. doi:10.1016/j.learninstruc.2012.05.009.

Jacobse, A. E., & Harskamp, E. G. (2012). Towards efficient measurement of metacognition in mathematical problem-solving. *Metacognition and Learning, 7*(2), 133–149.

Kellog, R. T. (1987). Writing performance: effects of cognitive strategies. *Written Communication, 4*, 269–298.

Kramarski, B., & Gutman, M. (2006). How can self-regulated learning be supported in mathematical E-learning environments? *Journal of Computer Assisted Learning, 22*, 24–33. doi:10.1111/j.1365-2729.2006.00157.x.

Labuhn, A., Zimmerman, B., & 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. doi:10.1007/s11409-010-9056-2.

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

Macbeth, G., Razumiejczyk, E., & Ledesma, R. D. (2011). Cliff's delta calculator: a non-parametric effect size program for two groups of observations. *Universitas Psychologica, 10*(2), 545–555.

Maki, R. H., Shields, M., Wheller, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723–731.

Montague, M., Enders, G., & Dietz, S. (2011). Effects of cognitive strategy instruction on math problem-solving of middle school students with learning disabilities. *Learning Disability Quarterly, 34*(4), 262–272. doi:10.1177/073i9487M421762.

Moos, D. C., & Ringdal, A. (2012). Self-regulated learning in the classroom: A literature review on the teacher's role. *Education Research International, 2012*, Article ID 423284. doi:10.1155/2012/423284.

Narciss, S., Koerndle, H., & Dresel, M. (2011). Self-evaluation accuracy and satisfaction with performance: are there affective costs or benefits of positive self-evaluation bias? *International Journal of Educational Research, 50*, 230–240. doi:10.1016/j.ijer.2011.08.004.

Newton, S. J., Horner, R. H., Algozzine, R. F., Todd, A. W., & Algozzine, K. M. (2009). Using a problem-solving model to enhance data-based decision making in schools. In W. Sailor, G. Dunlap, G. Sugai, & R. Horner (Eds.), *Handbook of positive behaviour support. Issues in clinical child psychology* (pp. 551–580). New York: Springer.

Newton, S. J., Horner, R. H., Todd, A. W., Algozzine, R. F., & Algozzine, K. A. (2012). A pilot study of a problem-solving model for team decision making. *Education and Treatment of Children, 35*(1), 25–49. doi:10.1353/etc.2012.0001.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*, 7–28.

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*, 159–179.

Olive, T., & Piolat, A. (2002). Suppressing visual feedback in written composition: effects on processing demands and coordination of the writing processes. *International Journal of Psychology, 37*(4), 209–218. doi:10.1080/00207590244000089.

Özsoy, G. (2012). Investigation of fifth grade students' mathematical calibration skills. *Educational Sciences: Theory and Practice, 12*(2), 1190–1194.

Panadero, E., & Alonso-Tapia, J. (2014). How do students self-regulate? Review of Zimmerman's cyclical model of self-regulated learning. *Annals of Psychology, 30*(2), 450–462. doi:10.6018/analesps.30.2.167221.

Parkinson, M. M., Dinsmore, D. L., & Alexander, P. A. (2010). *Calibrating calibration: Towards conceptual clarity and agreement in calculation. Paper presented at the annual Meeting of the American Educational Research Association*. CO: Denver.

Pennequin, V., Sorel, O., Nanty, I., & Fontaine, R. (2010). Metacognition and low achievement in mathematics: the effect of training in the use of metacognitive skills to solve mathematical word problems. *Thinking and Reasoning, 16*(3), 198–220. doi:10.1080/13546783.2010.509052.

Pereis, F., Dignath, C., & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education, 24*(1), I7–I29. doi:10.1007/BF03173472.

Piolat, A., Roussey, J. Y., Olive, T., & Farioli, F. (1996). Charge mentale et mobilisation des processus redactionnels: examen de la procédure de Kellogg [Mental load and time processing of writing: test of Kellogg's procedure]. *Psychologie Française, 41–4*, 339–354.

Piolat, A., Kellogg, R. T., & Farioli, F. (2001). The triple task technique for studying writing processes: on which task is attention focused? *Current Psychology Letters: Brain Behavior and Cognition, 4*, 67–83.

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology, 19*, 291–312. doi:10.1002/acp.1086.

Polya, G. (1954). *Mathematics and plausible reasoning*. Princeton: Princeton University Press.

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: calibration of confidence predicts the development of accuracy. *PLoS ONE, 9*(7), e98663. doi:10.1371/journal.pone.0098663.

Rosário, P., Mourão, R., Núñez, J. C., González-Pienda, J. A., & Solano, P. (2008). Storytelling as a promoter of self-regulated learning (SRL) throughout schooling. In A. Valle, J. C. Núñez, R. G. Cabanach, J. A. González-Pienda, & S. Rodríguez (Eds.), *Handbook of instructional resources and their applications in the classroom* (pp. 107–122). New York: Nova Science.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33–45. doi:10.1007/s11409-008-9031-3.

Schraw, G., Kuch, F., & Gutiérrez, A. P. (2012). Measure for measure: calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57. doi:10.1016/j.learninstruc.2012.08.007.

Schunk, D. (2008). Metacognition, self-regulation, and self-regulated learning: research recommendations. *Educational Psychology Review, 20*, 463–467.

Schunk, D., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35–53). New York: Routledge.

Sheldrake, R., Mujtaba, T., & Reiss, M. J. (2014). Calibration of self-evaluations of mathematical ability for students in England aged 13 and 15, and their intentions to study non-compulsory mathematics after age 16. *International Journal of Educational Research, 64*, 49–61. doi:10.1016/j.ijer.2013.10.008.

Stoeger, H., & Ziegler, A. (2008). Evaluation of a classroom based training to improve self-regulation in time management tasks during homework activities with fourth graders. *Metacognition and Learning, 3*(3), 207–230. doi:10.1007/s11409-008-9027-z.

Stolp, S., & Zabrucky, K. M. (2009). Contributions of metacognitive and self-regulated learning theories to investigations of calibration of comprehension. *International Electronic Journal of Elementary Education, 2*(1), 3–31.

Throndsen, I. (2011). Self-regulated learning of basic arithmetic skills: a longitudinal study. *British Journal of Educational Psychology, 81*, 558–578. doi:10.1348/2044-8279.002008.

Tillema, M., Van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2011). Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning, 6*(3), 229–253. doi:10.1007/s11409-011-9072-x.

Veenman, M. V. J. (2011). Learning to self-monitor and self-regulate. In R. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 197–218). New York: Routledge.

Veenman, M. V. J., Elshout, J. J., & Groen, M. G. M. (1993). Thinking aloud: Does it affect regulatory processes in learning? *Tijdschrift voor Onderwijs Research, 18*, 322–330.

Whimbey, A., & Lochhead, J. (1999). *Problem-solving and comprehension*. Hillsdale: Erlbaum.

Williams, J. R. (2008). Revising the declaration of Helsinki. *World Medical Journal, 54*, 120–125.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement* (pp. 153–190). New York: Erlbaum.

Winne, P. H. (2004). Students' calibration of knowledge and learning processes: implications for designing powerful software learning environments. *International Journal of Educational Research, 41*(6), 466–488. doi:10.1016/j.ijer.2005.08.012.

Winne, P. H., & Muis, K. (2011). Statistical estimates of learners' judgments about knowledge in calibration of achievement. *Metacognition and Learning, 6*(2), 179–193. doi:10.1007/s11409-011-9074-8.

Zimmerman, B. (2000). Attaining self-regulation. A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego, CA: Academic.

Zimmerman, B. (2008). Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183. doi:10.3102/0002831207312909.

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling, 53*(1), 141–160. doi:10.1080/00461520.2013.794676

Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance. In B. J. Zimmerman and D. H. Schunk (Eds.), Handbook of self-regulation of learning and performance (pp. 1–12). New York, NY: Routledge.