Contemporary Educational Psychology 49 (2017) 226–237

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

Empirical study

Within and between person associations of calibration and achievement

Teomara Rutherford

Department of Teacher Education and Learning Sciences, College of Education, North Carolina State University, Campus Box 7801, Raleigh, NC 27695, United States

ARTICLE INFO

Article history:
Available online 6 March 2017

Keywords:
Self-regulated learning
Calibration
Metacognition
Mathematics
Educational technology

ABSTRACT

Self-regulated learning (SRL), the ability to set goals and monitor and control progress toward these goals, is an important part of a positive mathematical disposition. Within SRL, accurate metacognitive monitoring is necessary to drive control processes. Students who display this accuracy are said to be *calibrated*, and although calibration is a growing area of research within Educational Psychology, unanswered questions remain about calibration's role as an aspect of metacognition, including the unique association between calibration and academic performance. In this study, calibration is characterized as part of a dynamic system that varies across tasks within the same person; variance in calibration is associated with variance in performance gain for the same student across tasks (quizzes within a year-long mathematics curriculum, ST Math). Both accurate determinations of certainty (Sensitivity) and uncertainty (Specificity) have unique small, yet statistically significant, associations with performance gains from pre to posttest in ST Math. For Specificity, there also remains a contextual association with performance at the Person level. Results are discussed in light of prior research on calibration and of theories of SRL; the data and analyses present a novel approach to studying calibration within a dynamic system and offer insights for future work.

1. Introduction

A wide range of research supports the associations between performance and the match between accuracy and confidence, or *calibration* (Hacker, Bol, & Keener, 2008). Mostly undertaken in the fields of Education or Psychology, calibration research has dealt with topics including eyewitness testimony (e.g., Howie & Roebers, 2007), text comprehension (e.g., Maki & Berry, 1984), and class performance (e.g., Bol & Hacker, 2001). A consistent finding is that higher performers display better metacognitive monitoring, often operationalized as accurate calibration (Stone, 2000; e.g., Bol, Riggs, Hacker, Dickerson, & Nunnery, 2010; Koku & Qureshi, 2004; Ots, 2013). Conversely, poor performers are "doubly cursed" in that not only do they perform poorly, but they are often unaware of their own poor performance, making it unlikely that they will take corrective action (Dunning, Johnson, Ehrlinger, & Kruger, 2003). This relation between calibration and performance is theorized to operate through a system of self-regulated learning (SRL), where accurate monitoring can alert learners to engage in control processes and allocate attention and resources where needed (Pintrich, 2004; Winne, 1995, 2004; Zimmerman, 2008). Although the calibration/performance relation is well-documented, researchers often use methods to study this relation as dispositional, examining performance differences between students who are more and less accurate in their calibration (e.g., Barnett & Hixon, 1997; Chen, 2002; Koku & Qureshi, 2004). This focus may not adequately describe the dynamic nature of monitoring as one that draws on the interaction of person, behaviors, and environments (Bandura, 1986; Zimmerman, 1989). Nor does it distinguish calibration from other individual characteristics that are likely related to both calibration and performance, such as motivation and knowledge. Within this study, a novel approach is taken to examine the relation between monitoring and performance within the same person across multiple tasks, analyzing whether differences in calibration across tasks are associated with differences in performance. In this way, the dynamic nature of calibration can be better represented and the unique contribution of calibration within the SRL system can be better understood.

1.1. Theoretical framework

The model of regulation adopted herein is situated within the larger frame of SRL as described by Zimmerman, in which learners set goals, monitor their progress toward those goals, and adjust as necessary (1986, 1989). In considering the characteristics of a self-regulated learner, Zimmerman noted "most importantly, self-regulated learners are aware of when they know a fact or possess a skill and when they do not" (1990, p. 4). Although this places

E-mail address: taruther@ncsu.edu

http://dx.doi.org/10.1016/j.cedpsych.2017.03.001
0361-476X/

calibration at the heart of Zimmerman's SRL, much of the work on calibration and its role within SRL has been situated in a line of research stemming from Flavell's (1979) and later Nelson (1990) conceptualizations of metacognition. Although typically distinct lines of research, the Flavell and Zimmerman conceptualizations of SRL and metacognition are complementary, jointly emphasizing the role of the learner as an agent of their own learning and acknowledging the contributing role of both learner background and task characteristics. This study also draws on aspects of the Efklides (2011) Metacognitive and Affective Model of Self-Regulated Learning (MASRL model), as well as Efklides' other work explicating the relations between monitoring and control (Efklides, 2008; Efklides & Vlachopoulos, 2012). Although the MASRL model focuses heavily on the role of motivation in SRL, the first tenet of this model is "There are two identifiable levels of functioning in SRL, namely, the Person level and the Task × Person level, with the Task influencing both levels. Moreover, there are reciprocal effects between the two levels" (Efklides, 2011, p. 16). It is this tenet that provides the theoretical justification for the framing of the current study.

Within Efklides' MASRL model, metacognition, motivation, and affect interact across two levels, the Person level and the Task × Person level (Efklides, 2011). The Person level includes personal characteristics such as self-beliefs, ability, and person-level metacognitive knowledge and metacognitive skills—knowledge and skills that apply to a variety of tasks and a sense of when and how to apply them. The Task × Person level is where online metacognition takes place: based on an individual's experience of the task, he or she represents the task in a way that allows him or her to draw on the Person level (e.g., metacognitive knowledge, skills, and motivation) and engage control processes in light of metacognitive experiences. These metacognitive experiences are cues from the individual's interaction with the task, such as an awareness of ease or difficulty of processing and feelings of familiarity or confidence (Efklides, 2008). When a learner becomes conscious of these metacognitive experiences, he or she can take steps to exercise control: increasing resources brought to bear on the task or allocating resources differently across different elements of the task. Learners who consciously monitor and who are calibrated accurately can appropriately allocate resources in the control process (see Efklides, 2008; Koriat, 2012).

1.2. Links between calibration and performance

As typically studied, calibration accuracy is measured with a perception task, such as a providing a judgment indicating the student's perception of the percentage of questions answered correctly on an exam (e.g., Bol, Hacker, O'Shea, & Allen, 2005). This perception of performance is then compared to actual performance to produce a measure of student calibration. Generally, higher achieving students are found to be better calibrated, and have been found by some researchers to exhibit underestimation bias, if any (Stone, 2000). This has been replicated across domains and age groups. With elementary students, the relation between accurate calibration and achievement has been found in math (e.g., Barnett & Hixon, 1997; Tobias and Everson, 1998), reading (e.g., Fajar, Santos, & Tobias, 1996; Romero & Tobias, 1996), social studies and spelling (e.g., Barnett & Hixon, 1997), and in playing computer games (e.g., Nietfeld, Minogue, Spires, & Lester, 2013). This calibration/achievement link has also been found in middle/high school math students (e.g., Bol et al., 2010; Chen, 2002; Chen & Zimmerman, 2007; Pajares & Kranzler, 1995), and in undergraduate students within knowledge-based courses like Research Methods (e.g., Bol & Hacker, 2001; Bol et al., 2005).

Moving beyond correlational studies, there has also been some evidence of the association between calibration and academic performance from longitudinal (e.g., Rinne & Mazzocco, 2014) and experimental studies (e.g., Nietfeld, Cao, & Osborne, 2006; Zimmerman, Moylan, Hudesman, White, & Flugman, 2011). Rinne and Mazzocco (2014) found that fifth grade calibration predicted eighth grade performance on the Woodcock Johnson-Revised measure of mental arithmetic, and that this prediction held even controlling for fifth grade performance on the measure. Nietfeld et al. (2006) found that undergraduate students who participated in a calibration training program improved their calibration scores relative to a control group, and that these improvements in calibration were related to growth in achievement. Zimmerman et al. (2011) training program was on SRL more broadly, but they also found that treatment students improved both in measures of calibration and performance relative to controls.

1.3. At what level does calibration operate?

The above studies provide evidence for the proposition that students who are better calibrated are able to perform better, and, in the case of the few training studies, provide evidence that calibration is malleable and that improvements to calibration can result in improvements to performance. These studies place calibration at the Person level of the Efklides (2011) MASRL model by examining calibration/achievement associations between students. What is less clear is how calibration operates within a dynamic system of SRL, one that varies within person from task to task, i.e., at the Task × Person level (Efklides, 2011).

There is some evidence that calibration, as a measure of monitoring accuracy, may be more stable across tasks than performance (Hacker et al., 2008; e.g., Johnsson & Allwood, 2003). A relevant line of research has explored whether calibration and other monitoring processes are domain-specific or domain-general. Gutierrez, Schraw, Kuch, and Richmond (2016) found that a model assuming a general monitoring factor best fit their data of college student calibration and performance on knowledge, probability, and spatial tasks. However, they interpret their results as indicating that both domain-specific and domain-general monitoring skills were likely important for monitoring. Greene et al. (2015) found cross-domain differences in their study using think-aloud protocol within a computer-based task. Their results indicated that students participating in a science task more frequently engaged in monitoring processes than did those assigned to a history task. Other researchers have also tested and failed to find support for a domain-general hypothesis. Kelemen, Frost, and Weaver, (2000) found that monitoring accuracy, which they term metacognitive accuracy, was not stable across tasks, although memory and confidence were.

In addition to potential differences across domains, monitoring may vary depending on features of the task as well. Prior research has shown that features such as item difficulty influence calibration accuracy within samples of adolescents and adults (e.g., Nietfeld et al., 2006; Winne & Jamieson-Noel, 2002). Additionally, the format of the item may matter. For example, Pallier et al. (2002) found their participants were better calibrated for open-ended questions than for multiple choice questions. As the specific content of items change, so may student motivation. Students bring with them views of the content and of themselves as learners relevant to the specific domain (see Eccles, Wigfield, Harold, & Blumenfeld, 1993), and these views play a role in monitoring and the application of monitoring toward control processes (Efklides, 2011).

As variance in monitoring and its application is theorized to be influenced by these features of the task, the person, and the interaction between person and task, understanding how this variance relates to performance may be best studied through examining multiple instances of relations between monitoring and performance within the same person. Within-person studies may also

help to deal with potential confounds to understanding monitoring and performance. Although there is evidence that people likely possess a general monitoring skill (e.g., Gutierrez et al., 2016; Hacker et al., 2008), there is also theoretical and empirical evidence that calibration, as measured, may in part reflect stable personal characteristics that have little to do with metacognitive monitoring (Pieschl, 2009; Scheck, Meeter, & Nelson, 2004; Zhao & Linderholm, 2008). Zhao and Linderholm (2008) present a theory wherein individuals, in making monitoring judgments, first anchor their judgment with expectations based on past experiences (from potentially unrelated tasks) and then adjust based on features of the actual task, ending with a judgment that, despite adjustment, is biased toward stable personal characteristics without adequately addressing task-specific considerations. By looking at the link between monitoring and performance within the same person, indicators of task-specific monitoring can be better disentangled from these stable personal characteristics. Research that compares associations between calibration and achievement across domains can help with this disentangling; however, these studies often have a small number of repeated measures, such as comparing between a math test and a language test, and they confound variation in domain-level knowledge and motivation with variation in monitoring.

1.4. Measures of calibration

Isolating Task × Person calibration from Person level aspects of regulation is one step toward better understanding the dynamic nature of monitoring within SRL. Certain operationalizations of monitoring may also shed light on different processes and the different ways in which monitoring can lead to control and improvements in performance. A number of measures of calibration have been used in prior research, many of which are discussed in Schraw, Kuch, and Gutierrez (2013; see also Boekaerts & Rozendaal, 2010; Feuerman & Miller, 2008; Schraw, 2009). Commonly used measures, such as Gamma, the associations between accuracy of judgments and performance, or Discrimination, the comparison of hit rates to false alarms, treat calibration as a single process through which learners make judgments of both confidence and uncertainty (Schraw et al., 2013). This may not be in line with models of SRL in which it is generally assumed that learners who can correctly identify both which content they know and which content they do not know can efficiently apply resources during control phases of SRL (Nelson, 1990; Schraw et al., 2013). Accurately calibrated learners can direct attentional resources away from material already mastered and toward material that has yet to be mastered. In experimental studies, cognitive scientists have demonstrated that individuals do indeed allocate more study time to items they deem as more difficult to learn (e.g., Nelson, 1990), although some research has found that under certain circumstances individuals will choose easier items first (e.g., Thiede & Dunlosky, 1999). Single process measures may conflate learner judgments of confidence and uncertainty and the use of each to guide and direct control processes. A two-process model focusing separately on judgments of confidence and judgments of uncertainty allow for examination of their distinct associations with achievement (see Schraw et al., 2013).

There is empirical support for a two-process model of monitoring. It is a consistent finding that poor performers display overconfidence; however, it is also a consistent finding that although better calibration is associated with better performance, the best performers tend to display underconfidence, although this relation diminishes as task difficulty increases (Stone, 2000). The theorized top-down process by which individuals make confidence judgments (see Zhao & Linderholm, 2008) may illuminate this finding. As individuals draw on general self-beliefs or prior experiences to make their judgments for a given task, those who are unfamiliar with the domain or topic of study may not have access to the information necessary for them to adjust their judgments in consideration of the task, and will therefore likely underestimate the task demands, leading to overconfidence (Kruger & Dunning, 1999). The converse may be true: those with more prior knowledge may have an abundance of resources upon which to draw and may overthink the problem, causing underconfidence. Additionally, metacognitive experiences at the Task × Person level may feed back into more stable beliefs at the Person level (Efklides, 2008), and so it may be protective, especially for those who feel threatened, to bolster their more general sense of self-worth with high confidence judgments (see Ots, 2013). Ots (2013) also offers evidence that high performers may underestimate as a form of defensive pessimism.

Schraw et al. (2013) analysis of ten measures of calibration using simulated data supported a two-process model, finding that including the measures of Sensitivity (proportion confident when correct) and Specificity (proportion uncertain when incorrect) in a model together best accounted for variance within their data. The recommendation that Sensitivity and Specificity be used together to represent the dual processes in calibration was supported with analyses of real data across three tasks in Schraw, Kuch, Gutierrez, and Richmond (2014). Additionally, my own analyses comparing the ten measures investigated in Schraw et al. (2013) found that a model including Sensitivity and Specificity together predicted the most variance within Spatial Temporal (ST) Math quiz data, although differences in explained variance between the models were small (see Rutherford, 2017). Modeling both Sensitivity and Specificity can allow an examination of the potentially different processes surrounding accurate judgments of confidence and uncertainty. Within-student differences in associations between these two measures and performance may indicate that they exert differential influences on control and thus performance, or may indicate that they are biased by different elements at the Task × Person level. For example, if influence on achievement differs between Sensitivity and Specificity, it could be because confidence in correct answers (Sensitivity) allows students to operate more efficiently and proceed more quickly through content, whereas uncertainty in incorrect answers (Specificity) may indicate to students where they need to direct their attention or restudy. Differences in associations *between* students may indicate that stable Person level characteristics associate differently with Sensitivity and Specificity.

1.5. The current study

Relying on a model of SRL in which monitoring accuracy can relate to performance as both a stable personal characteristic, between students (Person Level), and one that changes dynamically with the task, within student (Task × Person Level), the overarching research question addressed in this study is how calibration predicts performance in varying tasks within the same domain. Below, a description of the context for this study is followed by the specific research questions.

1.5.1. ST Math

The context for the study is an online mathematics learning environment, ST Math, created by the MIND Research Institute. Students participating in ST Math proceeded through a grade-level-specific curriculum, divided into 21–25 objectives, depending on grade level. Objectives covered included mathematics topic areas such as "Multi-Digit Multiplication" and "Linear Functions and Equations." Appendix Table 1 provides a description of each of the objectives within the ST Math curriculum, divided by grade. The content was ordered to approximate the progression of

content within a typical mathematics class, but was not aligned with pacing guides or other curricular materials.

Each objective within ST Math was prefaced with a five to 10-question pretest on objective-relevant content (examples provided in Appendix Fig. 1). Within the pretest, once students selected their answer for a question, they were prompted to indicate their confidence in this answer by selecting a cheering icon to represent certainty or a shrugging icon to represent uncertainty (see Fig. 1). This operationalization of confidence was tested by MIND and incorporated into the software after determining that even very young students could understand the icons and what they were measuring. Students were then allowed to review their answers and confidence ratings from the quiz before beginning the main objective content. The main content consisted of a number of learning games leveled by difficulty. Within each game, students solved puzzles to help the ST Math penguin, Jiji, proceed from left to right across the screen through the use of mathematical problem solving. This content between pre and posttests served to both instruct and provide practice on the relevant mathematics concepts. Students had to correctly complete 80% of the puzzles within each level to move on to the next level; however, students were able to replay levels as desired and otherwise proceeded at their own pace: there was no time-limit within the game for the completion of a given puzzle, level, or objective. When replaying, students could select different answer choices; the resulting animations would present the consequences of these choices, allowing for an experimental sandbox of sorts. At the end of each objective, students took a posttest quiz mirroring the content of the pretest quiz. The posttest quiz problems were structured in the same way as the pretest quiz problems: students were asked for confidence judgments on each problem and were allowed a post-quiz period of review.

As students took the pretest quizzes and made judgments of confidence, it was theorized that their attention was directed toward monitoring. They then proceeded to the main learning phase, where, if they were accurately calibrated and perceived a need for performance improvements, they could engage control processes to regulate their learning (such as controlling their attention or replaying games). Because the content of each objective and corresponding quiz varied, confidence judgments and their accuracy were also likely to vary, due in part to features of the task and interactions between person and task. If students were better able to engage in control processes during gameplay for objectives in which they were more accurately calibrated, and these control processes successfully influenced learning, then pre to posttest gains on these objectives would be larger than pre to posttest gains on objectives in which they were more poorly calibrated. By looking only at the joint variation in calibration and performance *within* each student, Person-level characteristics that influence regulation can be eliminated from the model, reducing bias on the estimate of calibration's association with performance (see Allison, 2005). The non-monitoring-aspects of the Task × Person level still remain and are potential confounds to the estimate, as are any task-specific personal characteristics (e.g., self-efficacy for specific ST Math problems); however, because the content across objectives shares a large number of features (in that it is grade-level mathematics problem-solving presented and tested within the same format), it can be assumed that some aspects of the Task × Person level are also controlled. Arguably one of the most important features of the Task × Person level, student familiarity with the particular task (see De Bruin & van Gog, 2012; Dunning et al., 2003) can be controlled by adding a pretest covariate in the model.

1.5.2. Research questions

Within this study, associations between calibration and achievement are explored within an online mathematics learning environment, ST Math, by asking: (1) Are students who are better calibrated at pretest better performers at posttest? This question sets the stage for replication of prior work that links calibration and achievement at the between student level. Moving beyond between-student comparisons and guided by a model of regulation wherein accuracy of metacognitive monitoring, as calibration, affects performance on a specific task through the engagement of control processes, question two asks: (2) Across tasks, do students perform better at posttest when better calibrated at pretest, after controlling for pretest performance? Statistically significant associations between calibration and performance within student will provide evidence that calibration uniquely contributes to variance in performance, net of student factors. An additional question asks: (3) Do associations between calibration and performance *between* students persist after partialling out within student variance? If so, this supports that a more stable metacognitive monitoring trait or that another stable Person level characteristic is associated with both calibration and performance. As specific sample selection may undermine the validity of conclusions drawn (Duncan, 2015), a replication analysis was conducted for a cohort of students participating in the project for a second year. Thus, a fourth research question asks: (4) Do the direction and strength of within and

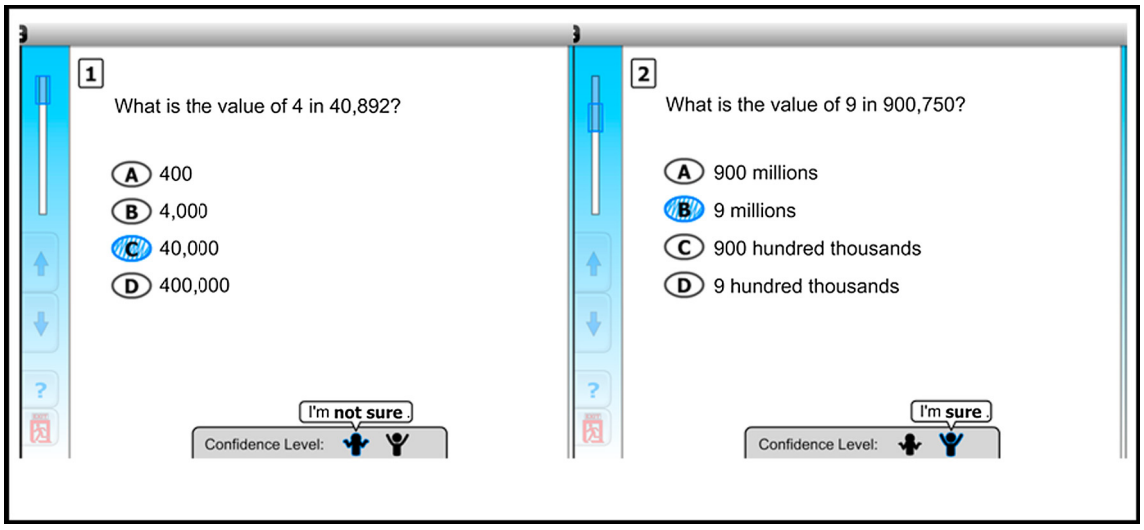

Fig. 1. Quiz questions appearing in ST Math. Students select their answer and then indicate their confidence rating by selecting the appropriate icon. Here, on the left the student is underconfident (right answer, expresses uncertainty) and on the right the student is overconfident (wrong answer, expresses certainty).

Table 1
Grade & demographic information of study students.

	Total sample		Analysis sample	
	Percent	Count	Percent	Count
Grade 2	22%	4137	21%	3912
Grade 3	19%	4137	19%	3912
Grade 4	37%	4137	37%	3912
Grade 5	23%	4137	22%	3912
Male	52%	4006	52%	3912
Asian	3%	4006	3%	3912
Hispanic	85%	4006	85%	3912
White	8%	4006	9%	3912
Other ethnicity	3%	4006	3%	3912
English language learner	66%	4005	66%	3912
Nat'l FREE/reduced lunch	80%	4006	80%	3912
N		4137		3912

Note. Total Sample includes all students in second through fifth grade in the study schools who began at least one objective within ST Math. The analysis sample is limited to those students who had complete demographic information and completed at least two complete objectives (pre and posttest). Count column indicates those students who had valid data on the variable of interest. Because the analysis sample was limited to those with complete data on these variables, the count is 3912 for all in this sample.

between student associations between calibration and achievement replicate in a second sample of students?

2. Method

2.1. Sample

The sample for this study comprises approximately half of the second through fifth graders at 18 schools in Southern California. The participating schools were largely Hispanic (85%) and low-income (80% eligible for free/reduced lunch), and on average, lower performing than the other county and state schools. The schools had been randomly assigned to receive ST Math in either second and third or fourth and fifth grades. Because the year of data collection for this paper (2010) is the second year of the study, those third grade students who were in schools assigned to second/third treatment in the first year of the study (2009) continued to receive treatment when they progressed to fourth grade. Therefore, all fourth graders in the participating schools were included in this sample. There were 4137 students in grades 2 through 5 who used ST Math within the study schools. Because the analysis focuses on data collected within ST Math, none of the control students were included. The current analyses were further limited to the 3912 students (95%) who were using their on-grade curriculum (excluding fifth graders using fourth grade curriculum, for example), and who had completed at least two of the 20+ objectives within a year's ST Math curriculum. Comparisons between the samples and descriptive statistics on both can be seen in Table 1. Excluded students were more likely to be in fifth grade and male; no other differences between excluded and retained students rose to the level of statistical significance.

2.2. Procedure

Students played ST Math for 45 min at a time during twice weekly visits to the computer lab throughout the academic year. Student selections on multiple choice quiz questions along with their ratings of confidence were collected and compiled by MIND and provided to the study team, who were able to match them with state identifiers and demographic information.

2.3. Measures

2.3.1. Quiz data

Calibration and performance were drawn from student answers to the quiz questions as seen in Fig. 1 and described in Section 1.5.1. Although students made judgments of confidence for each individual test question, the "task" level for this study was each objective quiz as an aggregate of these items. For each objective, accuracy was represented separately for pre and posttest quizzes as percentage correct. Calibration was operationalized as recommended by Schraw et al. (2013), with Sensitivity (percent of correct items where students noted confidence) and Specificity (percent of incorrect items where students noted uncertainty), based on the distribution of data within the 2 × 2 contingency table of confidence and accuracy (see Fig. 2). Specific information regarding the calculation of calibration is discussed in Section 2.4.1.

2.3.2. Demographics

Demographic information was provided by the participating school districts. This information included student gender, grade-level, ethnicity (categorized in analyses to represent the largest groups: Hispanic, Asian, White, and Other), English Language Learner (ELL) status, and free/reduced lunch eligibility as a measure of socioeconomic status.

2.4. Analysis

2.4.1. Calculation of calibration measures

For each quiz, Sensitivity was calculated by totaling the number of questions answered accurately where students also indicated confidence. This number was divided by the total of all questions answered correctly on that quiz, whether with confidence or uncertainty. For example, a student who answered four questions correctly, but only indicated confidence on two of them would have a value of 0.5. Specificity was calculated by dividing the number of questions answered incorrectly with uncertainty by the number of total questions answered incorrectly, either with confidence or uncertainty. Because of this calculation, a number of

A. Confident & Correct	B. Confident & Incorrect
C. Not Confident & Correct	D. Not Confident & Incorrect

Fig. 2. 2 × 2 contingency table expressing the relations between accuracy and confidence. Sensitivity is A/(A + C), Specificity is D/(B + D).

quizzes of students who were otherwise included in the sample did not have data for either Sensitivity (N = 2919, 5% of student/quiz pairs) or Specificity (N = 9490, 16% of student/quiz pairs), because, for these quizzes, the students either answered all questions correctly or all questions incorrectly. Listwise deletion would have removed all student/quiz pairs even when a quiz contained valid data on one of the two calibration measures. Because the data are NMAR, methods such as multiple imputation were not an option (Allison, 2002). A decision was made to assign these quizzes the same value as the relevant student-level mean. Additionally, in all multilevel analyses, two dummy variables were included to indicate whether either Sensitivity or Specificity was missing from that observation. Further analyses were conducted to see if the results were robust to a different missing data handling method, namely listwise deletion.

2.4.2. Person-level analysis

To answer the first research question, these data were analyzed in a way typical to calibration data: examining zero-order correlations between accuracy and calibration. Multiple regression analyses were then conducted looking between students to examine the associations between calibration and accuracy controlling for other observed student characteristics. This single-level analysis was conducted to both replicate prior work and to examine the impact of covariates on associations between posttest and calibration. Sensitivity and Specificity were included together in the model to represent accurate identifications of both confidence and uncertainty (see Schraw et al., 2013, 2014). To view the association between calibration and posttest achievement net of pretest achievement, a model was estimated also controlling for pretest accuracy. In these single student-level models, each student's pretest means for accuracy, Sensitivity, and Specificity were calculated as was each student's mean posttest accuracy (as outcome). The final single student-level model is represented by the following equation:

$$\overline{\text{PosttestAcc}}_i = \beta_0 + \beta_1\overline{\text{Sensitivity}}_i + \beta_2\overline{\text{Specificity}}_i + \beta_3\overline{\text{PretestAcc}}_i + \beta_3\text{Covariates}_i + r_i \quad (1)$$

2.4.3. Multilevel models

To answer the second and third research questions and to address the dynamic nature of calibration and its role within a model of regulation that varies between tasks, a random intercepts two-level hierarchical linear model with objectives nested within students was analyzed to determine whether student calibration at pretest (Sensitivity and Specificity) was associated with posttest performance. To isolate within-student effects and to eliminate bias from unobserved student characteristics (see Allison, 2005; Hofmann & Gavin, 1998; Park, 2008), group-mean centering around each student's mean across all quizzes was used for Level 1 predictors. In this way, the question of whether the same student scored higher during objectives when he/she was better calibrated could be answered. Unchanging student characteristics were entered as covariates at Level 2 along with student means for pretest accuracy and calibration. Models were built and tested starting with a non-HLM model then moving through: unconditional, covariates only, pretest score and covariates, random intercepts with calibration variables, random slopes. Only models that resulted in statistically significant changes in Deviance statistics were retained. Although three of the Level 1 variables had statistically-significant variability around Level 2 units, a model with random slopes resulted in an increase in Level 2 residual and did not change the direction or strength of the associations with the variables of interest; therefore, the final model reflects the more parsimonious random intercept model.

Level 1

$$\begin{aligned}\text{PosttestAcc}_{ti} = {} & \beta_{0i} + \beta_{1i}(\text{Sensitivity}_{ti} - \overline{\text{Sensitivity}}_{ti}) \\ & + \beta_{2i}(\text{Specificity}_{ti} - \overline{\text{Specificity}}_{ti}) \\ & + \beta_{3i}(\text{PretestAcc}_{ti} - \overline{\text{PretestAcc}}_{ti}) \\ & + \beta_{Ni}\text{ObjectiveCovariates}\ldots + r_{ti}\end{aligned} \quad (2)$$

Level 2

$$\begin{aligned}\beta_{0i} = {} & \gamma_{00} + \gamma_{01}\overline{\text{Sensitivity}}_i + \gamma_{02}\overline{\text{Specificity}}_i + \gamma_{03}\overline{\text{PretestAcc}}_i \\ & + \gamma_{0N}\text{StudentCovariates}_i + u_{0i}\beta_{1i} \\ = {} & \gamma_{10};\ \beta_{2i} = \gamma_{20};\ \beta_{3i} = \gamma_{30};\ \beta_{Ni} = \gamma_{40}\ldots\end{aligned} \quad (3)$$

The Level 1 associations are represented by formula (2), in which individual student mean (e.g., $\overline{\text{Sensitivity}}_i$) was subtracted from the student's score at that time-point (e.g., Sensitivity_{ti}) to arrive at the student group-mean centered value for that variable. Three additional Level 1 predictors were included: a variable indicating what number objective the quiz came from as a loose time/order variable and two variables noting whether the observation was missing Sensitivity or Specificity information, coded as a 0/1 dummy with 1 for missing, represented in the formula as β_{Ni}ObjectiveCovariates. In formula (2), posttest accuracy is a function of the student intercept (β_{0i}), these predictors (β_{1i}, β_{2i}, β_{3i}, β_{Ni}), and a time-varying student error (r_{ti}). The Level 2 or student-level predictors are represented within formula (3) above. In this formula, the student-level intercept is a function of the grand intercept (γ_{00}), the means of the calibration and pretest accuracy Xs (γ_{01}, γ_{02}, γ_{03}), the non-time-varying student-level characteristics, such as gender, ethnicity, ELL, free/reduced lunch, grade-level (γ_{0N}), and a student error (u_i).

In this way, the within student effect for Sensitivity (β_w) is represented by γ_{01} and the between student effect (β_b) is represented by γ_{10}. The effect for the individual at Level 2 is the difference between β_w and β_b—this "compositional" effect[1] is the extent to which the between student effect (Person level) remains once the individual quiz effect within student (Task × Person level) is controlled (see Raudenbush & Bryk, 2002). However, it is worth noting that because this compositional effect does not control for omitted variables unique to the person, it should not be interpreted as definitive (see Allison, 2005).

To test for these differences between β_w and β_b, Wald post-estimation tests were conducted to compare within and between student coefficients for Sensitivity, Specificity, and pretest accuracy. Differences, if statistically significant, were quantified and expressed as differences in standardized effect sizes. All standardized effect sizes were calculated using the relevant level-specific standard deviation for each variable using the formula: $(B * SD_X)/SD_Y$. In text, these are referred to as β.

3. Results

3.1. Question 1: Are students who are better calibrated at pretest better performers at posttest?

3.1.1. Describing the data

The means and standard deviations of each measure of calibration along with pre and posttest accuracy are presented in Table 2, divided by grade. The top half of the table presents descriptive statistics at the observation level and the bottom half presents

[1] Although the term "compositional" is a more natural description for nesting units such as classrooms and schools, the term is retained here to represent the student, in order to maintain consistency with the Raudenbush and Bryk (2002) calculation language.

Table 2
Quiz accuracy and calibration measures, by grade.

	Grade 2		Grade 3		Grade 4		Grade 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Observation/objective-level quiz descriptives (N = 56,962)								
Pretest accuracy	0.61	0.29	0.61	0.28	0.56	0.30	0.58	0.28
Pretest sensitivity[a]	0.84	0.32	0.85	0.29	0.80	0.35	0.84	0.31
Pretest specificity[b]	0.29	0.41	0.29	0.40	0.34	0.42	0.31	0.40
Posttest accuracy	0.69	0.28	0.70	0.27	0.67	0.30	0.70	0.27
N (Observations)	12,935		11,146		21,263		11,618	
Student-level quiz descriptive statistics (N = 3912)								
Pretest accuracy	0.60	0.13	0.59	0.16	0.54	0.14	0.57	0.13
Pretest sensitivity	0.84	0.19	0.85	0.18	0.80	0.20	0.83	0.18
Pretest specificity	0.29	0.27	0.29	0.25	0.33	0.27	0.32	0.25
Posttest accuracy	0.68	0.14	0.66	0.16	0.64	0.17	0.68	0.15
No. of objectives	15.47	6.53	14.88	6.63	14.63	6.54	13.29	5.36
N (Students)	836		749		1453		874	

Note. Data from analysis sample presented with objective data nested within students. Curricular and quiz content differs across grades.
[a] Objective-level sample limited to those who have at least one question correct (Grade 2 N = 12,290, Grade 3 N = 10,664, Grade 4 N = 19,959, Grade 5 N = 11,137).
[b] Objective-level sample limited to those who have at least one question incorrect (Grade 2 N = 10,533, Grade 3 N = 9410, Grade 4 N = 17,628, Grade 5 N = 9901).

them at the student level. On average, students were more accurate in their posttest answers than their pretest answers and were better able to correctly identify when they were correct (Sensitivity) than they were able to correctly identify when they may have been incorrect (Specificity).

Variation in calibration is distributed across the within and between student levels. Random effects Analyses of Variance revealed that 25% of the variance in Sensitivity and 29% of the variance in Specificity was associated with the student. Similarly, 17% of variation in pretest accuracy was between students.

3.1.2. Person level comparisons

Zero-order correlations between calibration and accuracy are presented in the bottom half Table 3. Full correlations between calibration, accuracy, and other student characteristics are presented in Appendix Table 2. The top half presents zero-order correlations for student-centered variables. At the person level, Sensitivity and Specificity had a strong inverse correlation. Sensitivity had moderate correlations with both pretest and posttest accuracy; Specificity had no correlations with either. The pre and posttest accuracy measures were strongly correlated with each other.

Table 4 presents the results from the student-level regressions of posttest accuracy on pretest accuracy and calibration. The first model presents the regression of posttest accuracy only on pretest accuracy, the models then include calibration measures and progress sequentially from a model that does not control for other observable student characteristics (Model 2) to one that has a full complement of demographic controls (Model 4). The third model is seen as an intermediate step as it controls partially for student characteristics through the addition of grade-level dummy variables, which also control for some characteristics of the task, as students at different grades received different curricula and quiz questions. Adding all observed student covariates does little to change explained variance. Model 4's r-squared was only a 0.007 improvement upon Model 2. The regression coefficients for calibration and pretest accuracy also changed little with the addition of student covariates. In the final model, there was a strong association between pretest and posttest accuracy: a one standard deviation increase in pretest accuracy was associated with a 0.77 standard deviation increase in posttest accuracy. The association between calibration and posttest accuracy was much smaller: $\beta = 0.12$ for Sensitivity and $\beta = 0.10$ for Specificity.

3.2. Question 2: Across tasks, do students perform better at posttest when better calibrated at pretest, after controlling for pretest performance?

The models in Table 4 look only between students and present calibration as a dispositional characteristic of the student, relating each student's average level of calibration to their average level of posttest performance within the ST Math curriculum. A further series of models were estimated to account for both the dynamic nature of calibration and to estimate associations between calibration and achievement net of unmeasured student characteristics. Based

Table 3
Correlations between calibration and accuracy measures.

N = 3912	Sensitivity	Specificity	Pretest Acc	Posttest Acc
Sensitivity	1	−0.45	0.21	0.16
Specificity	−0.83	1	−0.10	−0.05
Pretest Acc	0.21	−0.02[a]	1	0.37
Posttest Acc	0.20	−0.01[b]	0.81	1

Note. All correlations statistically significant at the $p < 0.001$ level, except those marked with superscript.
Bottom half displays correlations between person-level variables; top half displays correlations between individual-centered variables. Pairwise correlations were used for this level; Ns are reduced for Sensitivity (N = 54,050) and Specificity (N = 47,472).
[a] $p = 0.13$.
[b] $p = 0.72$.

Table 4
Student-level regressions of posttest accuracy on pretest accuracy and calibration.

N = 3912	(1) B (SE)	Beta	(2) B (SE)	Beta	(3) B (SE)	Beta	(4) B (SE)	Beta
Pretest accuracy	0.877***	0.812	0.852***	0.788	0.856***	0.792	0.840***	0.777
	(0.010)		(0.011)		(0.011)		(0.011)	
Sensitivity			0.104***	0.128	0.101***	0.124	0.099***	0.120
			(0.014)		(0.014)		(0.014)	
Specificity			0.072***	0.119	0.068***	0.114	0.064***	0.107
			(0.010)		(0.010)		(0.010)	
Grade 2					−0.005	−0.012	−0.005	−0.013
					(0.004)		(0.004)	
Grade 3					−0.013**	−0.033	−0.013**	−0.033
					(0.004)		(0.004)	
Grade 5					0.015***	0.041	0.015***	0.040
					(0.004)		(0.004)	
Eng. lang. learner							−0.009*	−0.028
							(0.004)	
Male							−0.007*	−0.023
							(0.003)	
Asian							0.008	0.009
							(0.008)	
White							0.001	0.001
							(0.006)	
Other ethnicity							0.002	0.002
							(0.008)	
Free/reduced lunch							−0.014***	−0.036
							(0.004)	
Constant	0.164***		0.070***		0.071***		0.104***	
	(0.006)		(0.014)		(0.014)		(0.015)	
R2	0.659		0.664		0.668		0.671	
Change in R2			0.005		0.004		0.003	

Note.
Bs are unstandardized regression coefficients; Betas are standardized coefficients. Standard errors in parentheses. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

on a random effects Analysis of Variance, 22% of variance in the posttest was associated with student as the grouping variable. This was confirmed with the unconditional hierarchical model specifying student as the nesting variable. Proportion of variance between the two levels and descriptions of incremental model fit are provided in Appendix Table 3. Student-level covariates produced a small but statistically significant improvement in explaining variance in posttest accuracy (2.61%). The addition of the pretest accuracy variables at both levels along with a variable indicating the quiz number resulted in a larger improvement over the unconditional model (29.72%), the addition of calibration variables also resulted in a statistically significant improvement over this model, but the incremental improvement was small (0.98%).

Table 5 displays the results from the hierarchical regressions. Compared to the unconditional model, the full conditional model explained 87% of the variance between students and 14% of the variance within students. Pretest accuracy was a moderately strong predictor of posttest accuracy from quiz to quiz ($\beta = 0.38$; standardized coefficients calculated using the level-specific standard deviations of predictor and outcome variables). Both calibration measures were statistically significant predictors of within student differences in posttest performance (Sensitivity: $\beta = 0.07$; Specificity: $\beta = 0.01$), and these coefficients were different from each other ($p < 0.001$ based on post-estimation Wald test). This indicates that when students were one standard deviation higher on their confidence when correct (Sensitivity) they were 0.07 of a standard deviation higher on posttest accuracy; when they were one standard deviation higher on uncertainty when incorrect (Specificity) they were 0.01 of a standard deviation higher on posttest accuracy.

3.3. Question 3: Do associations between calibration and performance between students persist after partialling out within student variance?

The mean of student pretest accuracy at Level 2 was similar in magnitude to the association as calculated with the one-level model ($\beta = 0.81$). This Level 2 effect size combines both the quiz and student levels. Subtracting the within effect size from the between effect size resulted in a 0.43 contextual effect for student (see Raudenbush & Bryk, 2002). Post-estimation Wald tests revealed that this contextual effect was statistically significant at the $p < 0.001$ level. As means, both calibration variables were statistically significant predictors at the student level of mean posttest achievement (Sensitivity: $\beta = 0.13$; Specificity: $\beta = 0.11$). The difference between Level 1 and 2 Sensitivity was relatively small ($\beta = 0.06$) and not statistically significant ($p = 0.28$). For Specificity, post-estimation Wald tests did indicate a statistically significant difference between the Level 1 and Level 2 coefficients ($p < 0.0001$), with a difference of $\beta = 0.10$; this contextual effect was different from that of Sensitivity ($p = 0.0002$).

3.4. Question 4: Do the direction and strength of within and between student associations between calibration and achievement replicate in a second sample of students?

As a robustness check, a replication was run using data from students who were in the study schools and participated in ST Math during the 2011–2012 school year. Because of the design of the study, this sample included those students who were in the 2010 sample and did not age out or otherwise leave their schools. It also included students who were in grades that became ST

Math-using grades at the start of the 2011 school year. Table 4 in the Appendix displays demographic information on this new sample. Starting in 2011, all students in grades 2–5 in all study schools used ST Math, and so the grade distribution is more even than in 2010, where fourth graders dominated (compare with Table 1). Also of note is the ethnic makeup of this new sample: the sample was less Hispanic (74% vs. 85%) and more White (20% vs. 8%).

There were differences in calibration and accuracy by grade between the samples, but all differences were smaller than 0.08. Although small in magnitude, most differences were statistically significant. Full descriptives are available in Appendix Table 5. Correlations between quiz variables are in Appendix Table 6.

As with the 2010 data, around 20% of the variance in posttest was associated with student as the grouping variable (see Table 6). As variables are added to the model, the model fit improved at levels of statistical significance, with the final model improving on the unconditional model by 27.49% (compare with 30.41% for the 2010 sample). Hierarchical regression results for the 2011 sample are presented in Table 6. Effect sizes for pretest accuracy at both levels were within β = 0.07 of the 2010 model; calibration effect sizes were within β = 0.008 at Level 1 and β = 0.06 at Level 2, with direction and relative strength of Sensitivity and Specificity comparable between the two years. Specifically, effect sizes for Level 1 were stronger for Sensitivity than Specificity (β = 0.06 vs. β = 0.01). Also replicating the 2010 analysis, there was no statistically significant contextual effect of Sensitivity (β = 0.007, *ns*), but a contextual effect for Specificity that was larger than the within-student effect (β = 0.05, $p = 0.002$).

3.5. Sensitivity to missing data handling

Student calibration data were missing for some quizzes in a way that was assumed correlated with outcome variables (NMAR). To test the sensitivity of results to decisions regarding handling of missing data (see Section 2.4.1), final hierarchical regressions were conducted using listwise deletion of observations that were missing either Sensitivity or Specificity measures due to students answering all questions correctly or incorrectly. Appendix Tables 8 and 9 contain the hierarchical regressions from these analyses on the 2010 and 2011 samples, respectively. All effect sizes for within-person and contextual associations were within 0.005 of those from the full analysis samples.

4. Discussion

4.1. Different associations between calibration and achievement at two levels

Analyses conducted with multiple instances of measures of calibration and performance across a year-long mathematics curriculum were able to disentangle the associations between calibration

Table 5
Results from hierarchical regressions of post-test accuracy on pre-test accuracy, calibration, & covariates, 2010 sample.

	Beta	B	SE	*p*
Fixed parameters				
Level 1 (β)				
Sensitivity	0.068	0.077	0.004	<0.0001
Specificity	0.013	0.013	0.004	0.001
Pretest accuracy	0.376	0.359	0.006	<0.0001
Objective no.	0.021	0.001	<0.0001	<0.0001
No sensitivity	0.031	0.052	0.005	<0.0001
No specificity	−0.011	−0.009	0.014	0.802
Level 2 (γ)				
Sensitivity	0.126	0.093	0.014	<0.0001
Specificity	0.114	0.059	0.010	<0.0001
Pretest Accuracy	0.806	0.868	0.011	<0.0001
Grade 2	−0.045	−0.015	0.003	<0.0001
Grade 3	−0.054	−0.019	0.004	<0.0001
Grade 5	0.036	0.012	0.004	0.001
ELL	−0.021	−0.006	0.003	0.050
Male	−0.029	−0.008	0.003	0.002
Asian	0.003	0.002	0.007	0.759
White	<0.001	−0.001	0.005	0.890
Other Ethnic	0.002	0.009	0.007	0.228
Free/Reduced Lunch	−0.066	−0.011	0.004	0.002
Intercept		0.090	0.014	<0.0001
Random parameters				
Between (u_{0i})	0.003	0.0001		
Within (r_{ti})	0.054	0.0003		
% Variance explained				
L2	0.832			
L1	0.148			

Note. Variance explained as calculated in comparison to null two-level model (see Appendix Table 3). R2 as calculated by Stata as per Raudenbush and Bryk (2002) is 0.144 for L1 and 0.867 for L2. Beta provides standardized regression coefficients; B, unstandardized regression coefficients. Level 1 quiz accuracy and calibration variables are group-mean centered around student means. Level 2 quiz variables represent student means. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch.

Table 6
Results from hierarchical regressions of post-test accuracy on pre-test accuracy, calibration, & covariates, 2011 replication sample.

	Beta	B	SE	*p*
Fixed parameters				
Level 1 (β)				
Sensitivity	0.060	0.066	0.003	<0.0001
Specificity	0.011	0.010	0.003	<0.0001
Pretest accuracy	0.390	0.358	0.004	<0.0001
Objective no.	0.002	0.002	<0.0001	<0.0001
No sensitivity	0.033	0.056	0.004	<0.0001
No specificity	−0.014	−0.011	0.003	<0.0001
Level 2 (γ)				
Sensitivity	0.068	0.049	0.010	<0.0001
Specificity	0.062	0.034	0.007	<0.0001
Pretest accuracy	0.743	0.723	0.009	<0.0001
Grade 2	−0.057	−0.019	0.003	<0.0001
Grade 3	0.023	0.008	0.003	0.009
Grade 5	<0.001	0.000	0.003	0.996
ELL	−0.051	−0.015	0.002	<0.0001
Male	0.025	0.007	0.002	0.001
Asian	0.023	0.019	0.005	<0.0001
White	0.026	0.010	0.003	<0.0001
Other ethnic	0.015	0.013	0.006	0.026
Free/reduced lunch	−0.012	−0.005	0.003	0.091
Intercept		0.224	0.011	<0.0001
Random parameters				
Between (u_{0i})	0.002	0.0001		
Within (r_{ti})	0.049	0.0002		
% Variance explained				
L2	0.822			
L1	0.141			

Note. Variance explained as calculated in comparison to null two-level model (see Appendix Table 7). R2 as calculated by Stata as per Raudenbush and Bryk (2002) is 0.145 for L1 and 0.833 for L2. Beta provides standardized regression coefficients; B, unstandardized regression coefficients. Level 1 quiz accuracy and calibration variables are group-mean centered around student means. Level 2 quiz variables represent student means. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch.

and other Person level variables related to regulation in a manner not previously undertaken within calibration research. Analysis related to the first research question indicated that calibration and achievement were associated in this sample. Additionally, with respect to the second research question, when better calibrated at pretest, students *did* perform better at posttest, net of associations with pretest. These calibration/posttest associations were replicated with an additional sample of students, but were small, with effect sizes less than one tenth of a standard deviation. This is in contrast to the larger effects of calibration reported in previous research, many using between-person comparisons and zero-order correlations (e.g., Bol et al., 2010; Maki, Shields, Wheeler, & Zacchilli, 2005; Koku & Qureshi, 2004; Ots, 2013).

Within the current study, effect sizes were sensitive to the mode of analysis: Table 7 compares effect sizes for Sensitivity and Specificity across methods. Sensitivity in particular appears to be inflated when using zero-order correlations, indicating that person characteristics associated with both Sensitivity and performance can bias results when analysis is conducted with this typical method. Although estimates for Sensitivity were reduced between the one-level model (Table 4) and the hierarchical linear model (Table 5), the largest reduction was from the zero-order correlations to the single-level model adjusted for pretest. It may be that Sensitivity's association with pretest accuracy is the relation that causes the greatest statistical bias—this bias was eliminated in both regression models by the addition of a pretest covariate. For Specificity, associations with posttest were largest in the single level model, and these results were close to the Level 2 contextual associations between Specificity and performance in the hierarchical linear model. This could indicate that for students at similar levels of performance, a dispositional person-level characteristic is driving this association. The lack of zero-order correlations may be because of a suppression relation between Sensitivity and Specificity.

Table 7
Comparison of standardized effect sizes for 2010 data across analysis methods.

	Zero-order correlations	One-level model	Hierarchical model
Sensitivity	0.20	0.12	0.07
Specificity	−0.01	0.10	0.01

Sensitivity and Specificity also had statistically significantly different associations with performance gains in the hierarchical linear model. Within student, the effect size for Sensitivity was more than three times as large as that for Specificity. It is a common assumption of theoretical metacognitive and SRL models that students who are aware of what they do not know should engage in behaviors to direct and control learning, for example, students in ST Math could adjust attention, replay puzzles, or seek help (see Efklides, 2008; Nelson, 1990; Zimmerman, 2008). The small effect size for Specificity suggests that such awareness of errors may not be triggering these control processes or control processes are not being enacted well, at least when Specificity is measured within students at the Task × Person level, where student and task characteristics interact. It appears more important at this level that students are confident when they are correct, as measured by Sensitivity. It is less clear what control mechanism may be at play with regards to this result. It may be that very task-specific self-efficacy is driving this association. Students may be more willing to persist on ST Math content and quizzes dealing with topics about which they feel confident. Although this is merely speculative, it would be in line with the positive association between self-efficacy and performance that is a consistent finding in the motivation literature (e.g., Bandura, 1997; Pajares, 1996).

Sensitivity and Specificity also differed in their associations with performance gains between students at the Person level. The between-student effect for Sensitivity was no different than the within-student effect, but once the within-student association was partialled out, the contextual effect for Specificity was statistically significantly different from zero and three times as large as the within-student effect. This could indicate that a more general awareness of errors within a domain is beneficial to students. However, as noted in Allison (2005), this estimate of the contextual effect may be biased by other stable characteristics of the student and should be interpreted with caution.

These results contribute to the literature in two important ways. First, they highlight the important distinctions between Sensitivity and Specificity. Although studies of simulated data show these measures as uncorrelated (e.g., Schraw et al., 2013), other studies using non-simulated data also find an inverse correlation (e.g., Gutierrez et al., 2016). That Sensitivity and Specificity showed this inverse correlation in these data and that they had different relations both between and within students provides further evidence that they represent separate metacognitive evaluation processes. Second, within this study, multiple instances of calibration and achievement within the same student are leveraged to estimate a less-biased association between calibration and achievement—one that is net of person-level factors. The Allison (2005) method of group mean centering together with the use of a multilevel model gives us insight into how Sensitivity and Specificity relate to achievement at two levels. The results have implications for future research on calibration, suggesting that a two-process model with both Sensitivity and Specificity may be needed to understand monitoring accuracy and its association with performance. The change in associations between both measures and achievement depending on the model used (Table 7) suggests that researchers should carefully consider omitted variables that may inflate associations, in the case of Sensitivity, or mask them, in the case of Specificity.

4.2. Limitations

Although this study contributes to the calibration literature by isolating calibration from other person-level factors, even the within-person associations may be biased by other variables at the Task × Person level. For example, the effect of metacognitive knowledge about math tasks within ST Math in general would have been removed from the model; however, the difference between what a student knows about shape problems and what a student knows about fraction problems could be picked up by the calibration estimates between objectives, biasing the results. Task-specific metacognitive knowledge is just one example of omitted variables that reduce my ability to make causal claims. As posited by the MASRL model (Efklides, 2011), a number of factors operate within students at the Task × Person level that were not incorporated in the current study, such as task-specific motivation and individual representation of the task, including personal benchmarking for success as needed to make confidence judgments. Using content that is more closely related may solve this problem, but a complete solution may remain elusive: as young students make great leaps in their math learning across the school year, the nature of their interaction with the task is likely to evolve, fundamentally changing aspects at the Task × Person level even in tasks that appear similar. More comprehensive measuring of these variables may be needed in order to discuss the unique contribution of each.

These data allowed an investigation of within-person associations between calibration and performance, but did not provide insight into how students might use their accurate monitoring to exercise control. To truly understand the dynamic process of SRL, research linking control processes with monitoring and performance is needed. Lastly, these conclusions are also limited by the specific sample. The young students within this study may not be fully able to accurately monitor. Children's ability to monitor metacognition may suffer from working memory demands in complicated tasks (Ghetti, Hembacher, & Coughlin, 2013; Hacker, Dunlosky, & Graesser, 1998). Even when they do display the same mean-level of calibration accuracy as their older counterparts, younger children may not be as able to use information from metacognitive monitoring to influence control processes (Destan, Hembacher, Ghetti, & Roebers, 2014).

5. Conclusion

This study demonstrates the potential role of calibration within the system of SRL, showing how differences in the accuracy of student metacognitive monitoring are related to differences in performance within an online mathematics curriculum. As a main contribution, this study explores the dynamic nature of calibration as it varies from task to task within the same student. For the elementary school students within this study, calibration emerged as a statistically significant predictor of posttest performance net of pretest performance. Within students, the accurate identification of correct answers (Sensitivity) had larger associations with performance gains than did the accurate identification of uncertainty for incorrect answers (Specificity), although both had small statistically significant coefficients. This finding was replicated with a second sample of students. These results can help calibration researchers in education and other fields start to disentangle the unique contribution of metacognitive monitoring within SRL—better understanding how calibration works in this dynamic system can help support the improvement of calibration and the improvement of SRL, both thought foundational to many learning activities.

Acknowledgements

This research was supported in part by grants from the Institute of Education Sciences to the University of California, Irvine (Grant R305A090527) and the National Science Foundation [grant number DGE-0808392].

Additional thanks to the participating districts, schools, teachers, and students, as well as George Farkas, Jacquelynne Eccles, Greg J. Duncan, Deborah Lowe Vandell, Elizabeth Loftus, John Nietfeld, and Gregory Schraw, who provided feedback on the study and earlier versions of this manuscript. Thank you also to MIND Research Institute for provision of data and assistance during the project.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cedpsych.2017.03.001.

References

Allison, P. D. (2002). *Missing data series: Quantitative applications in the social sciences*. Thousand oaks, CA: Sage.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. SAS Publishing.

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

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Barnett, J. E., & Hixon, J. E. (1997). Effects of grade level and subject on student test score predictions. *The Journal of Educational Research, 90*(3), 170–174. http://dx.doi.org/10.2307/27542087.

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20*(5), 372–382. http://dx.doi.org/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*(2), 133–151. http://dx.doi.org/10.2307/20152656.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290. http://dx.doi.org/10.2307/20157403.

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

Chen, P. (2002). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*(1), 77–90. http://dx.doi.org/10.1016/j.lindif.2003.08.003.

Chen, P., & Zimmerman, B. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *Journal of Experimental Education, 75*(3), 221–244. http://dx.doi.org/10.3200/jexe.75.3.221-244.

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

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology, 126C*, 213–228. http://dx.doi.org/10.1016/j.jecp.2014.04.001.

Duncan, G. J. (2015). Toward an empirically robust science of human development. *Research in Human Development, 12*(3–4), 255–260. http://dx.doi.org/10.1080/15427609.2015.1068061.

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

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64*(3), 830–847. http://dx.doi.org/10.1111/j.1467-8624.1993.tb02946.x.

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277–287. http://dx.doi.org/10.1027/1016-9040.13.4.277.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL Model. *Educational Psychologist, 46*(1), 6–25. http://dx.doi.org/10.1080/00461520.2011.538645.

Efklides, A., & Vlachopoulos, S. P. (2012). Measurement of metacognitive knowledge of self, task, and strategies in mathematics. *European Journal of Psychological Assessment, 28*(3), 227–239. http://dx.doi.org/10.1027/1015-5759/a000145.

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: Sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Practice, 14*(5), 930–933. http://dx.doi.org/10.1111/j.1365-2753.2008.00984.x.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. http://dx.doi.org/10.1037/0003-066X.34.10.906.

Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling uncertain and acting on it during the preschool years: A metacognitive approach. *Child Development Perspectives, 7*(3), 160–165. http://dx.doi.org/10.1111/cdep.12035.

Greene, J. A., Bolick, C. M., Jackson, W. P., Caprino, A. M., Oswald, C., & Mcvea, M. (2015). Domain-specificity of self-regulated learning processing in science and history. *Contemporary Educational Psychology, 42*, 111–128. http://dx.doi.org/10.1016/j.cedpsych.2015.06.001.

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction, 44*, 1–10. http://dx.doi.org/10.1016/j.learninstruc.2016.02.006.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 429–455). New York: Psychology Press.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (1998). *Metacognition in educational theory and practice*. Routledge.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*(5), 623–641. http://dx.doi.org/10.1177/014920639802400504.

Howie, P., & Roebers, C. M. (2007). Developmental progression in the confidence-accuracy relationship in event recall: Insights provided by a calibration perspective. *Applied Cognitive Psychology, 21*(7), 871–893. http://dx.doi.org/10.1002/acp.1302.

Koku, P. S., & Qureshi, A. A. (2004). Overconfidence and the performance of business students on examinations. *Journal of Education for Business, 79*(4), 217–224. http://dx.doi.org/10.3200/JOEB.79.4.217-224.

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

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

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 10*(4), 663–679. http://dx.doi.org/10.1037/0278-7393.10.4.663.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In Gordon H. Bower (Ed.). *Psychology of Learning and Motivation* (Vol. 26, pp. 125). . Retrieved from <http://www.sciencedirect.com/science/article/pii/S0079742108600535>.

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

Ots, A. (2013). Third graders' performance predictions: Calibration deflections and academic success. *European Journal of Psychology of Education, 28*(2), 223–237. http://dx.doi.org/10.1007/s10212-012-0111-z.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543–578. http://dx.doi.org/10.3102/00346543066004543.

Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology, 20*(4), 426–443. http://dx.doi.org/10.1006/ceps.1995.1029.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129*(3), 257–299. http://dx.doi.org/10.1080/00221300209602099.

Park, H. S. (2008). Centering in hierarchical linear modeling. *Communication Methods and Measures, 2*(4), 227–259. http://dx.doi.org/10.1080/19312450802310466.

Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning, 4*(1), 3–31. http://dx.doi.org/10.1007/s11409-008-9030-4.

Pintrich, P. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385–407. http://dx.doi.org/10.1007/s10648-004-0006-x.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

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. http://dx.doi.org/10.1371/journal.pone.0098663.

Rutherford, T. (2017). The measurement of calibration in real contexts. *Learning and Instruction, 47*, 33–42. http://dx.doi.org/10.1016/j.learninstruc.2016.10.006.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51*(1), 71–79. http://dx.doi.org/10.1016/j.jml.2004.03.004.

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

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57. http://dx.doi.org/10.1016/j.learninstruc.2012.08.007.

Schraw, G., Kuch, F., Gutierrez, A. P., & Richmond, A. S. (2014). Exploring a three-level model of calibration accuracy. *Journal of Educational Psychology*. http://dx.doi.org/10.1037/a0036653.

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

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25*(4), 1024–1037. http://dx.doi.org/10.1037/0278-7393.25.4.1024.

Tobias, S., & Everson, H. T. (1998, April). Research on the assessment of metacognitive knowledge monitoring. Paper presented at the annual convention of the American Educational Research Association, San Diego.

Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist, 30*(4), 173–187. http://dx.doi.org/10.1207/s15326985ep3004_2.

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.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27*(4), 551–572. http://dx.doi.org/10.1016/S0361-476X(02)00006-1.

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review, 20*(2), 191–206. http://dx.doi.org/10.1007/s10648-008-9073-8.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology, 11*(4), 307–313. http://dx.doi.org/10.1016/0361-476X(86)90027-5.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339. http://dx.doi.org/10.1037/0022-0663.81.3.329.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3–17. http://dx.doi.org/10.1207/s15326985ep2501_2.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183. http://dx.doi.org/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. Available online at <http://www.psychologie-aktuell.com/journale/psychological-test-and-assessment-modeling.html>.