Metacognition and Learning (2021) 16:89–111
https://doi.org/10.1007/s11409-020-09239-3

Changes in metacognitive monitoring accuracy in an introductory physics course

Jason W. Morphew[1]

Received: 28 October 2019 / Accepted: 20 August 2020 / Published online: 1 September 2020

Abstract
Student learning in introductory science, technology, engineering, and mathematics (STEM) courses is often self-regulated. For self-regulated learning to be effective, students need to engage in accurate metacognitive monitoring to make appropriate metacognitive control decisions. However, the accuracy with which individuals monitor their task performance appears to largely overlap with their ability to perform that task. This study examined the trajectories in the accuracy of students' metacognitive monitoring over the course of a semester, along with the effect of monitoring accuracy feedback. The results indicate that some students improve the accuracy of their predictions over the course of a semester. However, low-performing students are less accurate at predicting their exam grades, and tend not to improve their metacognitive calibration over the course of a semester. In addition, providing low-performing students with calibration feedback may lead to greater overconfidence.

Keywords Metacognitive monitoring · Epistemological beliefs · Academic goal orientation · Low-performing students

Introduction

Learning within authentic contexts such as introductory science, technology, engineering, and mathematics (STEM) courses is considered self-regulated because students act as active participants who largely control how they interact with their course material (Tuysuzoglu & Greene, 2015). This is particularly true for course homework and when studying for exams as these activities typically occur outside of the classroom. Because students actively control their learning in these contexts,

✉ Jason W. Morphew
jmorphew@purdue.edu

1 School of Engineering Education, Purdue University, 701 W. Stadium Drive, West Lafayette, IN, USA

success within introductory STEM courses is largely due to the effectiveness with which a student is able to engage in self-regulated learning, a process that requires effective metacognitive monitoring and control processes (Greene & Azevedo, 2007; Winne & Hadwin, 1998; Zimmerman, 2008).

Imagine a student who has several exams to prepare for over the next two weeks in multiple courses. This student must determine how to allocate their time studying to maximize their performance across all of the classes. To prepare effectively for their upcoming exams, the student needs to consider which courses will require more of their time to study, as well as determine the topics on which they should focus. Students need to know how their current knowledge compares to course expectations, what their academic goals are for each course, and the amount of time it will take to learn the subject material. In other words, to be successful, the student needs to engage in metacognitive monitoring to make judgments about their current ability level, or knowledge state for each course and for each topic within the courses. After making a *metacognitive judgment* about their current state, the student uses their epistemological beliefs (i.e., knowledge and beliefs about the nature of learning), their metacognitive knowledge (i.e., knowledge of, and beliefs about, potential cognitive strategies), and their academic goal orientations (i.e., course specific performance or mastery goals) to plan and enact a study strategy. Along with their academic goals for the course, the effectiveness of their study strategies relies on the accuracy of their metacognitive knowledge and monitoring, their beliefs about the speed at which learning can occur, and their beliefs about the course expectations.

In many introductory STEM courses, a large percentage of a student's grade comes from two to five midterm and final exams. A low score on even one of these exams can lead to a lower course grades, motivation, and even potentially lower persistence within STEM (Cromley, Perez, & Kaplan, 2016; King, 2015). Given the importance of exams for course outcomes, it is important to identify students who are inaccurately monitoring their understanding of course material before the first exam. While previous research has investigated the link between metacognition and other psychological traits, little research has investigated predictors of metacognitive monitoring accuracy and changes in monitoring accuracy.

Introductory STEM courses typically aim to cover a broad sampling of topics leaving little room in the curriculum for additional topic instruction, such as metacognitive skill development. In addition, the primary mode of instruction in these courses is through passive modes of instruction such as lectures (Erdmann, Miller, & Stains, 2020). Further, systemic barriers prevent the implementation of findings from cognitive science within these introductory courses (Henderson, Mestre, & Slakey, 2015). To overcome these barriers, can a simple intervention lead to improvements in monitoring accuracy or exam performance within an introductory STEM course?

This study explores three research questions. First, to what extent do ability, epistemological beliefs, and goal orientations predict accuracy of metacognitive judgments of physics exam performance? Second, what effect does metacognitive accuracy feedback improve the accuracy of the students' metacognitive monitoring? Third, to what extent do epistemological beliefs and goal orientations predict changes in accuracy of metacognitive judgments of physics exam performance over the course of a semester? In the next section, we begin with an overview of metacognition within a self-regulated learning framework, followed by brief reviews of prior research on the relationship between metacognitive monitoring and ability, epistemological beliefs, and goal orientations. Finally, we briefly review the effect of feedback on metacognitive monitoring accuracy.

Literature review

Metacognition

Metacognition, or the act of thinking about and regulating cognitive processes, refers to the ability to monitor one's current learning, evaluate the learning against a criterion, and make and execute plans to maximize one's learning (Tobias & Everson, 2009). Metacognition was initially seen as consisting of metacognitive knowledge (knowledge and informal theories about human cognition in general), metacognitive experiences (individualized experiences one has during their own cognition), goals (the objectives of a cognitive activity), and actions (the specific behaviors used to achieve the goals; Flavel, 1979). Recently, most views of metacognition break metacognition into metacognitive knowledge and metacognitive skills (Dunlosky & Metcalfe, 2008; Scott & Levy, 2013; Veenman, Van Hout-Wolters, & Afflerbach, 2006). Metacognitive knowledge refers to the declarative knowledge of different cognitive strategies, the procedural knowledge of how to implement each cognitive strategy, and the beliefs and heuristics concerning the contextual effectiveness of each strategy, all of which is derived from prior learning experiences (Pintrich, 2002). Metacognitive skills refer to the ways in which learners engage in self-regulated learning processes and include monitoring and control processes (Dunlosky & Bjork, 2008). Learners engage in metacognitive monitoring when they evaluate their current state of learning against a criterion, and engage in metacognitive control when they select study strategies or items for study, or decide when to stop studying.

The predominant framework in which metacognition research occurs assumes that a dynamic and reciprocal relationship occurs between metacognitive monitoring and control processes (Nelson and Narens, 1990, 1994). Within this framework, learners need to be able to accurately monitor their learning, possess effective heuristics for determining when learning has occurred, and utilize effective control strategies for altering their current cognitive processes for learning to be effective. Models of self-regulated learning agree with the assertion that metacognitive control processes are driven by the accuracy of the learners' metacognitive monitoring (Ariel, Dunlosky, & Bailey, 2009; Greene & Azevedo, 2007; Metcalfe & Kornell, 2005; Nelson & Narens, 1990, 1994; Soderstrom, Yue, & Bjork, 2015; Son & Metcalfe, 2000; Winne & Hadwin, 1998, 2008). The relationship between metacognitive monitoring and control is mediated by task conditions, learner goals, and metacognitive knowledge (Koriat, Ma'ayan, & Nussinson, 2006; Koriat, Nussinson, & Ackerman, 2014). For example, learners tend to allocate time to material that they judge to be more difficult when they are not under time pressure (Finn & Metcalfe, 2008; Kelley & Jacoby, 1996), but items that are closest to their current ability level when time pressured (Metcalfe & Kornell, 2005). Learners are also strategic in adapting the time they allocate to items based on learning goals (Wilkinson, Reader, & Payne, 2012), and perceived item value (Ackerman, 2014; Castel, Benjamin, Craik, & Watkins, 2002; Koriat, 2007).

Within self-regulated learning tasks, it is important that learners have an accurate model of their current understanding and how it relates to their goals for the specific learning task. Research on monitoring processes often focus on the accuracy between monitoring and performance. To study the accuracy of individuals' metacognitive monitoring, learners are asked to make judgments about the state of their learning at various times in the learning process (Dunlosky & Thiede, 2013). In addition, metacognitive judgments can be made on either an item by item-level basis (judgments for each problem), or by having learners provide a single global judgement for the entire task (Dunlosky & Thiede, 2013). Judgment accuracy can be measured using either relative or absolute measures of accuracy. Relative accuracy, or resolution, refers to the ability to distinguish between items that are known and unknown items (Rhodes, 2015), and requires item-by-item judgments for a given exam. Absolute accuracy refers to the magnitude of the discrepancy between judgment and performance. In this study, measures of absolute accuracy are used because students are likely to have an overall goal for an exam (i.e., to earn an A, or to pass) that help to determine when their current level of learning matches their goal for the exam, meaning that they have sufficiently prepared.[1]

Metacognitive judgments are typically more accurate for judgments made after an exam (postdictions) than those made before an exam (predictions), and are usually more accurate for item-level judgments rather than global judgments (Dunlosky & Lipko, 2007). However, students preparing for exams in introductory STEM courses are unlikely to know the individual questions in advance. As such, item-level judgments do not reflect the metacognitive judgments that students are likely to make when preparing for course exams. In addition, predictions are likely to reflect the metacognitive monitoring that influenced their studying decisions. In this study, global predictions were used to measure metacognitive monitoring accuracy because these judgments are more likely to reflect the metacognitive monitoring that learners utilize when preparing for exams.

Metacognitive monitoring and ability

A large body of research has investigated the accuracy of individuals' metacognitive judgments. The accuracy of individual's metacognitive judgments is often found to be related to one's domain knowledge (Fakcharoenphol, Morphew, & Mestre, 2015; Glaser & Chi, 1988; Schneider, 2002). These studies have generally found that individuals tend to overestimate their own performance, with the overestimates being more pronounced for low-performing individuals (Kruger & Dunning, 1999; Morphew, Gladding, & Mestre, 2020; Rebello, 2012; Serra & DeMarree, 2016). The asymmetry in the accuracy of learners' metacognitive judgments is thought to occur because the expertise and skills needed to make accurate metacognitive judgments of performance are the same type of expertise and skills needed to produce good performance on a task (Schlosser, Dunning, Johnson, & Kruger, 2013). From this perspective, low performing students suffer from a double curse of being both unskilled and unaware of their lack of skill (Kruger & Dunning, 1999). However, the less accurate metacognitive monitoring judgments made by low-performing individuals are also likely driven by the desire for positive outcomes and misconceptions about the normative difficulty of the tasks as well as misconceptions about their own performance (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Serra, & DeMarree, 2016; Simons, 2013). Because metacognitive judgements are used to make studying decisions, it is important for low-performing individuals to improve the monitoring accuracy in order to help them to determine when they have sufficiently learned the material.

Changes in metacognitive monitoring accuracy

Laboratory studies have generally demonstrated improvements in calibration over time (e.g, Ariel & Dunlosky, 2011; Tauber & Dunlosky, 2015; Tauber & Rhodes, 2012). Similarly, over

[1] A second reason was pragmatic given that course instructors did not want item-by-item local judgements used on the high-stakes exams for this course.

the course of a semester, one might expect exam predictions to become more accurate as students anchor their predictions based on prior exam performance, and the adjust for their experiences with the new course material (Hacker, et al., 2000; Huff & Nietfield, 2009; Geurten & Meulmans, 2017). However, in classroom contexts, students often make predictions of their exam score using a desire for a positive outcome or their desired grade (Saenz, Geraci, Miller, & Tirso, 2017; Simons, 2013). In fact, students' predictions are often more strongly correlated with their desired grade than with their actual grade (Serra, & DeMarree, 2016), and often fail to use their prior exam performance when making exam predictions (Foster, Was, Dunlosky, & Isaacson, 2017).

The anchoring of exam predictions in a desired grade rather than prior performance may be why the results of previous classroom studies looking at changes in metacognitive monitoring have been mixed, with some studies failing to demonstrate improvement in the accuracy of metacognitive judgments over the course of a semester even with interventions designed to focus student attention on the accuracy of their predictions and incentives for accurate predictions (Foster, et al., 2017; Nietfeld, Cao, & Osborne, 2005). Conversely, other studies have shown improvements in calibration over the course of the semester (e.g., Hacker, Bol, & Bahbahani, 2008; Hacker, Bol, Horgan, & Rakow, 2000; Nietfeld, Cao, & Osborne, 2006). Many of the studies analyzed judgment accuracy using group centered analyses (e.g., repeated measures ANOVA or MANOVA) that do not take individual differences into account. A notable exception is found in Foster, et al. (2017), who used multilevel modeling to examine changes in bias (i.e., metacognitive monitoring accuracy) over the course of thirteen exams. The authors found that on average students are overconfident before taking exams and the magnitude of their overconfidence increased over time. However, they also noted that there was significant variation in students bias before taking the first exam, and marginally significant variation in the change in bias over the course of the semester. In addition, there was significant covariance between the initial bias and the change over time, suggesting that the increase in bias over the course of the semester may be restricted to those who have lower initial bias.

Feedback

Feedback from an external source is essential to development in models of self-regulated learning. External feedback provides information about the contextual task standards and motivates individuals to engage in reflection that critical for more accurate metacognitive monitoring (Zimmerman, 2000). Much of the research on the effect of feedback on monitoring accuracy has been focused on performance feedback. Several studies have found that providing individuals with performance feedback can lead to more accurate monitoring (Baars et al. 2014; Keren, 1990; Laburn, Zimmerman, & Hasselhorn, 2010; Lipko et al. 2009; Rawson & Dunlosky 2007),. However, students in most introductory STEM courses receive detailed correctness feedback on exams for both global (overall exam grade) and local (individual question correctness) judgments, yet many studies have failed to demonstrate improvement in monitoring (e.g., Foster, Was, Dunlosky, & Isaacson, 2017) despite receiving this performance feedback. This failure to demonstrate more accurate judgments may be due to enhanced monitoring being restricted to highly similar or repeated tasks (Thompson, 1998).

More extensive feedback involves providing individuals with calibration feedback that shows the performance feedback, the participants prediction, and a brief interpretation of the accuracy of the prediction can improve monitoring accuracy in lab settings for repeated tasks, (Geurten & Meulmans, 2017; Kim, 2018; Nietfeld, Cao, & Osborne, 2006; Urban & Urban, 2018, 2019), and in classroom settings for exams (Callender, Franco-Watkins, & Roberts, 2015; Miller & Geraci, 2011). In one example, Miller & Geraci (2011) provided students with their exam score and their prediction after every exam in a cognitive psychology course. They found that providing accuracy feedback led to low-performing students becoming more accurate in monitoring over the course of the semester, even though these low-performing students did not improve the exam performance. Callender, Franco-Watkins, & Roberts (2015) showed similar results, although they used a pre-post design rather than a randomized control design. However, the authors note that the improvements in monitoring that they found may have been largely due to performance improvements, rather than simply more accurate monitoring. Thus, the effect that providing students with monitoring feedback (an easy-to-implement intervention for large enrollment introductory STEM courses) has on monitoring accuracy is unclear.

Academic goal orientation, epistemological beliefs, and metacognition

From a self-regulated learning framework, accurate metacognitive monitoring implies two acts. First, the learner must set criteria for what it means to know "enough" for the upcoming test. Second, learners must monitor their learning against the standards they have established. This implies that students' beliefs about the nature of knowledge and learning are likely to influence the accuracy of their metacognitive judgments made before taking an exam. In fact, prior research suggests that overestimates of performance are likely due to individuals overestimating their own ability as well as underestimating the difficulty of the problems on the exam (Metcalfe & Finn, 2008). In addition, an individual whose goal is to earn a certain grade in the course may be more motivated to make an accurate prediction. Conversely a student who is simply focused on not failing may be more likely to inflate their judgement to maintain a positive self-image. This suggests that a student's academic goal orientation may also influence the accuracy of their metacognitive judgments. While no study to date has investigated the link between epistemological beliefs, academic goal orientation, and the accuracy of metacognitive judgments, a few studies have investigated the relationship between students' epistemological beliefs, academic goal orientations, and metacognition in general.

Epistemological beliefs The relationship between metacognition and epistemological beliefs has been studied by a number of researchers (e.g., Bromme, Pieschl, & Stahl, 2010; Hofer, 2004; Kitchner, 1983; Muis, & Franco, 2010). Kitchner (1983) viewed the relationship between cognition, metacognition, and epistemology as a three-level system where at the epistemological level individuals think on a "meta-meta level" (Barzilai & Zohar, 2014, p. 17), reflecting on the limits of their knowing and the nature of knowledge in general. More recently epistemology has been conceived as an integral component of metacognition. Hofer (2004) utilized the classic distinction between metacognitive knowledge and skills or processes and conceptualized epistemological beliefs as components of metacognition. She locates epistemological beliefs dealing with the nature of knowledge (the certainty of knowledge and simplicity of knowledge) within metacognitive knowledge. Beliefs about the nature of knowing (the source of knowledge and the justification for knowing) are located with metacognitive processes or skills. From this perspective an individual's epistemological beliefs will affect metacognitive processes. For example, consider an individual's belief about the nature of intelligence. Individuals adopt beliefs about the nature of intelligence, specifically whether intelligence is fixed (entity mindset) or is changeable through effort (incremental mindset; Dweck, 1999). Beliefs about the malleability of intelligence likely affect how students engage in learning strategies that require more effort. In other words, if a student believes that their ability in physics is changeable, then they are more likely to adopt an incremental learning goals approach (Dweck, 1999). This mindset can affect how they make predictions of the performance, as individuals with an entity mindset tend to make lower predictions of their performance than those who adopt an incremental mindset (Hong, Chiu, & Dweck, 1995).

Within self-regulated learning contexts, the relationship between epistemological beliefs, metacognition, and learning is a primary concern. Several models have been proposed which view learning as occurring in a cyclical process. The key assumption of these models is that an individual's epistemological beliefs are implicitly activated during the initial phases of a task (e.g., studying for an exam). Once activated, these beliefs (along with the nature of the task) determine the standards for learning. These standards in turn impact the amount of information processed by learners (Pieschl, Stahl, & Bromme, 2008), the metacognitive knowledge which is activated, and the metacognitive processes that one employs to complete a learning task (Bromme, Pieschl, & Stahl, 2010; Mason, Boldrin, & Ariasi, 2010; Muis & Franco, 2010).

The ability to accurately monitor depends on the individual having substantial knowledge both within and about the domain in which the task is situated (Veenman, Van Hout-Wolters, & Afflerbach, 2006). This includes knowledge about what it means to know within a particular domain. In other words, when monitoring cognition individuals must make two simultaneous and related judgements. Learners must first decide what it means to know within a subject (e.g., is knowledge in the domain complex and interrelated or is it simple and segmented). Then learners must decide how their current level of knowledge aligns with the expectations of the domain. Similarly, when selecting control strategies learners must draw upon their domain specific epistemology in order to determine which strategies best align with the subject and task demands.

Using the framework of epistemic metacognition, we can imagine that two students with the same metacognitive competency, and with the same ability, can make drastically different judgments of their learning. These different judgments can lead to different strategy use even if they are equally capable of employing the same metacognitive strategies. When individuals make metacognitive judgments of learning before taking an assessment, they are actually making two judgements; one epistemic and one metacognitive. Individuals must make an epistemic judgement about what it means to know in a particular subject area, and then make a metacognitive judgement to determine how their current knowledge relates to the expectations of the domain. These judgments are likely used to determine when to study, how much time and effort to put into studying, and what strategies to employ while studying.

Achievement goal orientation The way in which students approach learning within a given context is, in part, related to the goals they set for learning. Achievement goal orientations (AGOs) are the general orientation a student adopts when engaging in learning within a specific context. There are two types of learning goals that a learner may choose to adopt; *mastery* goals and *performance* goals (Dweck & Leggett, 1988; Elliot & Murayama, 2008). For each of these goals, learners may adopt either an *approach* or *avoidance* valence resulting in four distinct achievement goal orientations (Elliot & Murayama, 2008). Learners with mastery-approach goals tend to focus on attaining task-based competence or mastering conceptual understanding. Learners with mastery-avoidance goals tend to focus on avoiding conceptual misunderstandings or developing task-based incompetence. Learners with performance-approach goals tend to focus on demonstrating normative competence (e.g., performing better than average). Learners with performance-avoidance goals tend to focus on avoiding demonstrations of normative incompetence. Learners may adopt multiple achievement goals simultaneously and to different extents. For example, a student enrolled in an introductory biology course, and who plans to apply to medical school, may adopt both mastery-approach goals (focusing on developing a mastery of the material) and performance-approach goals (wanting to score in the top quartile of the class).

AGOs have been found to be related to measures of course performance (e.g., Elliot & Murayama, 2008) and transfer (Belenky & Nokes-Malach, 2013), however few studies have looked at the relationship between AGOs and metacognition in general. Both Coutinho (2007) and Gul and Shehzad (2012) conducted surveys of students' metacognition using the metacognitive awareness index, which measures metacognitive knowledge and strategies. Both studies found that mastery goal orientation was correlated with academic performance (as measured by self-reports of GPA) and with metacognitive awareness. Bipp, Steinmayr, and Spinath (2012) investigated the link between AGOs and metacognitive monitoring by having students complete measures of these constructs, estimate their intelligence, then complete an intelligence test. They found that students' performance, but not mastery goals, were related to estimates of intelligence. In addition, students with performance approach goals tended to overestimate their intelligence while those with performance-avoidance goals tended to underestimate their intelligence. However, the estimates of intelligence were made using both percentile and Likert scales, making measures of absolute calibration unwarranted.

Method

Participants

Participants were 284 Undergraduate students enrolled in an algebra-based introductory physics course at a large Midwestern university who completed consent forms at the beginning of the semester agreeing to participate in this study. Due to a glitch in the online survey delivery platform demographics data are only available for 164 students. The demographics indicated that the sample was relatively evenly distributed for gender (43% female, 57% male), and representative of the course distribution for ethnicity (3.0% African American, 21.3% Asian American, 12.8% International, 44.5% Caucasian, 5.5% Hispanic, 12.9% other ethnicities). Neither the mean age nor socio-economic data were available. Students were randomly assigned to either receive feedback about the accuracy of their predictions (feedback condition; $N = 141$) or to a control conditions where they did not receive feedback (no feedback condition; $N = 143$).

Procedure

Participants completed the surveys during the first week of the semester as part of the course delivery system. Students completed three computerized midterm exams and one comprehensive final exam (heretofore referred to as exam 4) on paper during the semester. The four exams were all constructed in a similar format in that the questions were multiple-choice and contained both calculational and conceptual questions. Before beginning each of the four exams, students were prompted to make a prediction about their expected performance on the exam using the prompt: "Before you begin the exam, please take a second to think about what grade you anticipate getting on this exam (0 - 100%). Try to be as accurate as you can with your prediction." To motivate accurate metacognitive judgments, students who predicted within 3% of their actual exam grade were entered into a drawing for one of three $30 prizes on each exam.

Students in the feedback condition received calibration feedback similar to the feedback in Miller and Geraci (2011). After every exam, students in the feedback condition were sent an email that indicated their exam score, their predicted score on the exam, and whether their prediction was an overestimate or underestimate. Students were also instructed to compare the prediction and the exam score. Students in the no feedback condition were given their exam scores through the course management system, but did not receive any information about their predictions.

Measures

At the beginning of the semester, students completed surveys that measured their goal orientations and epistemological beliefs. The surveys are discussed in more detail below. Due to a glitch in the online survey delivery platform, survey data were only available for 170 of the students.

Goal orientation Participants' achievement goal orientations were measured using the Revised Achievement Goal Questionnaire (AGQ) (Elliot & Murayama, 2008). This questionnaire is intended to measure the participants' approach and avoidance behaviors on two different goal orientations; performance and mastery goals. The questionnaire asks students to consider their goals for the introductory physics course and then rate their agreement to statements reflecting the range of goal orientations using a 5-point Likert scale. All four subscales have been reported to display high reliability (mastery-approach, $\alpha = .84$, mastery-avoidance, $\alpha = .88$, performance-approach, $\alpha = .92$, and performance-avoidance, $\alpha = .94$). The reliabilities were lower for this sample (mastery-approach, $\alpha = .74$, mastery-avoidance, $\alpha = .69$, performance-approach, $\alpha = .76$, and performance-avoidance, $\alpha = .75$). The scores were slightly negatively skewed. To aid in the interpretation of the results, the scores on all four subscales were normalized by subtracting the mean and dividing by the standard deviation.

Epistemological beliefs Investigations on epistemological beliefs initially focused on developmental issues and assumed that epistemology was unidimensional. Schommer (1990) noted that this assumption was unlikely and proposed that that personal epistemology was composed of a set of independent beliefs which may be thought of as existing on a continuum from less-adaptive to more adaptive positions. Schommer proposed that epistemological beliefs are comprised of five dimensions of beliefs that are interrelated; (a) the simplicity/complexity of knowledge, (b) the certainty of knowledge, (c) the source of knowledge, (d) innate ability, and (e) the speed of learning. Participants' epistemological beliefs were measured using the Connotative Aspects of Epistemological Beliefs questionnaire (CAEB) developed by Stahl and Bromme (2007), and the Theories of Intelligence (TOI) scale designed by Dweck (1999).

The CAEB is intended to measure individuals' epistemological beliefs about the simplicity/complexity of knowledge, the certainty of knowledge, and the source of knowledge. The questionnaire asks students to indicate how they believe knowledge in physics might be best described using a seven-point (1–7) Likert-scale with 24 adjective pairs (e.g., negotiated-discovered) as the end-points. On this scale, a score of 1 indicates that knowledge in physics is represented by only the first adjective (e.g., negotiated), while a score of 7 indicates that knowledge in physics is represented by only the first adjective (e.g., discovered). A score of 4 indicates that knowledge in physics is represented by both adjectives equally. Two questions were removed from the calculation of the scores because their inclusion lowered the reliability of the scales. The reliabilities for this sample were acceptable (Simple, $\alpha = .74$, Certainty, $\alpha = .62$, and Source, $\alpha = .61$). The scores were normally distributed. To aid in the interpretation of the results, the scores on all four subscales were normalized by subtracting the mean and dividing by the standard deviation.

The TOI asks students to think about learning in Physics, read 7 philosophical statements, and rate their level of agreement using a six-point Likert scale ranging from strongly agree to strongly disagree. For example; "Your intelligence is something about you that you can't change very much." The reliability was relatively high for this sample ($\alpha = .89$), and the scores were normally distributed. To aid in the interpretation of the results, the scores were normalized by subtracting the mean and dividing by the standard deviation.

Ability group The exams for this course varied in difficulty as the means for the four exams were 72.2%, 61.3%, 78.5%, and 65.1% respectively.[2] To determine whether high or low performing students show better metacognitive calibration or more improvement over time the ability level of each student was estimated by calculating their exam average across the four exams. Students were divided into quartiles using the average of the four exams. The average exam scores for each quartile were as follows: First quartile [35% - 61%], second quartile [61% - 69%], third quartile [69% - 79%], and fourth quartile [79% - 99%].

Bias Students' metacognitive bias was calculated by subtracting their exam score from their prediction so that positive scores represent overconfidence and negative scores represent underconfidence.

Data analysis

For ethical reasons, students were not required to make predictions, therefore, not all students made predictions for every exam. The majority of the students made predictions for every exam. Of the 284 students who consented, 279 made predictions for the first exam, 268 made predictions for the second exam, 258 made predictions for the third exam, and 241 made predictions for the final exam. One student provided letter grade predictions rather than percentage predictions. This student's letter grade predictions were converted to the mean of the letter grade (e.g., scores between 86% and 88% are awarded a B, thus an estimate of a B was converted to a prediction of 87%). Two students did not make a prediction for any of the four exams given in the course, and an additional 11 students dropped the course after having made at least one prediction. The data for these 13 students were not included in the data

[2] The Physics department generally aims to write exams that have a mean score between 70 and 75%. The second and final exams had lower means than desired by course instructors. While course instructors aim for a mean in this range, students are not made aware of this goal. Historically, the mean varies and may fall outside of this this desired range

analysis. The responses to the survey data were matched to the exam and prediction data for each individual.

Pearson correlations, Analyses of Variance (ANOVAs), Multivariate Analyses of Variance (MANOVAs), and Chi-Square tests were used to analyze differences for all variables except goal orientation, which were analyzed using Kruskal-Wallis tests. Descriptive statistics, Chi-Square tests, Kruskal-Wallis tests, and ANOVAs were calculated with SAS Version 9.4. The trajectory of metacognitive bias of exam predictions were tested in a structural equation modeling framework to analyze change over time. Since there was missing data Mplus Version 7.11 was used to conduct the growth curve modeling and growth mixture modeling since the default analysis utilizes full information maximum likelihood estimation which handles missing data well.

Examination of spaghetti plots suggested non-linear trajectories, so a sequence of models was tested to fit the shape of growth. A no-growth model, unconditional growth model, and a quadratic growth model were fit using the maximum likelihood (ML) estimator. The models were compared using the change in chi-squared test. To assess model fit, we analyzed the CFI, RMSEA, and Chi-Square Goodness of fit tests. Cutoffs of CFI $\geq .95$ and the lower bound of the 90% CI for RMSEA $\leq .05$ were used to determine the model fit (Hooper, Coughlan, & Mullen, 2008). These results, along with the multiple trajectories suggested by the spaghetti plots indicated that a model with a single trajectory for all students was inappropriate, so several growth mixture models were run to determine the optimum number of trajectories (classes). Multiple models with different variance and covariance structures were fit. However, due to convergence problems with the other models, Nagin models were fit to the data where intraclass variances and covariances were fixed at zero (Nagin, 2005; Nagin & Odgers, 2010). In the final model, the intercept, slope, quadratic, and cubic slopes were freely estimated using the ML estimator for three of the classes, while only the intercept was estimated for the fourth class (i.e., the linear, quadratic, and cubic slopes were fixed for this class) because examination of the spaghetti plots suggested a large group of students that exhibited no change in bias across the exams.

Results

To address the first research question, Pearson correlations were conducted and the results are shown in Table 1. Students' bias scores among the exams were all positively correlated suggesting that the students were relatively consistent in their calibration. However, neither academic goal orientations nor epistemological beliefs were consistently correlated with bias scores.

To address the second research question, Two-way ANOVAs (Ability x Feedback) were conducted for each exam. Descriptive statistics for the bias scores for each feedback condition are shown in Table 2. The ANOVAs showed differences in bias between ability groups for exam 1, $F(3, 260) = 46.89$, $p < .001$, $\eta_p^2 = .34$, exam 2, $F(3, 256) = 19.87$, $p < .001$, $\eta_p^2 = .18$, exam 3, $F(3, 247) = 8.02$, $p < .001$, $\eta_p^2 = .09$, and exam 4, $F(3, 233) = 36.00$, $p < .001$, $\eta_p^2 = .31$.[3] Post-hoc Tukey's HSD tests indicated that the low-ability group was more overconfident than the high-ability group and the medium-high ability group on all four exams. The low-

[3] The same conclusions are reached if only those who made predictions on all four exams are used in the analysis.

Table 1 *Pearson correlations among measured variables*

	1	2	3	4	5	6	7	8	9	10	11
1. Bias (Exam 1)											
2. Bias (Exam 2)	.263**										
3. Bias (Exam 3)	.236**	.356**									
4. Bias (Exam 4)	.453**	.459**	.250**								
5. TOI	.070	.102	.008	.158							
6. Mastery Approach	−.072	−.044	−.050	−.049	.187*						
7. Mastery Avoid	−.088	−.024	−.063	−.033	.124	.227**					
8. Performance Approach	−.081	−.059	.005	−.013	.061	.319**	.237**				
9. Performance Avoid	−.036	−.047	−.162*	−.081	.093	.200**	.419**	.527**			
10. Simple	−.029	−.135	.083	−.211**	−.004	.208**	.097	−.010	−.012		
11. Certain	−.060	−.046	−.084	.100	.154*	−.026	.090	.007	.050	−.300**	
12. Source	.044	.086	−.050	.189*	.074	−.119	−.055	.064	.063	−.664**	.503**

Note: * $p \leq .05$, ** $p < .01$,

Table 2 *Means and Standard Deviation of the Bias Scores for Each Ability Group by Feedback Condition*

	No Feedback Condition Mean(SD)				Feedback Condition Mean(SD)			
	Low	Med-Low	Med-High	High	Low	Med-Low	Med-High	High
Exam 1	21.7 (11.4)	12.8 (12.2)	9.8 (10.6)	3.1 (9.3)	24.1 (12.0)	13.3 (10.4)	11.5 (11.1)	−1.5 (8.6)
Exam 2	19.2 (14.4)	17.9 (14.6)	15.1 (15.5)	4.9 (10.1)	27.2 (15.4)	19.4 (15.5)	11.7 (12.7)	6.8 (10.4)
Exam 3	7.2 (13.2)	−2.1 (12.2)	2.3 (12.6)	0.8 (9.7)	7.9 (14.4)	−0.2 (9.2)	−3.1 (13.1)	−3.2 (9.0)
Exam 4	17.8 (14.3)	12.3 (10.1)	6.7 (8.4)	2.3 (11.1)	25.8 (16.4)	18.1 (10.3)	7.0 (8.4)	0.3 (10.3)

Note: Sample sizes for the no feedback condition by each exam respectively: Low ability ($n = 33, 31, 30, 29$), Medium-Low ability ($n = 34, 32, 32, 30$), Medium-High ability ($n = 34, 34, 32, 32$), High ability ($n = 30, 31, 30, 27$). Sample sizes for the feedback condition for each exam respectively: Low ability ($n = 34, 33, 30, 28$), Medium-Low ability ($n = 31, 32, 32, 30$), Medium-High ability ($n = 36, 36, 33, 30$), High ability ($n = 36, 35, 36, 35$)

ability group was also more overconfident than the medium-low ability group on all but the second exam. The high-ability group was less overconfident than the other three ability groups, except for exam 3, where they were only less overconfident than the low-ability group.

A small, but significant main effect for feedback condition was found for exam 4, $F(1, 232) = 4.82$, $p = .03$, $\eta_p^2 = .01$, indicating that those who received metacognitive monitoring accuracy feedback were more overconfident on the final exam compared to those who did not receive accuracy feedback. However, no differences in bias between feedback conditions were detected for exam 1, $F(1, 260) < 0.01$, $p = .98$, $\eta_p^2 < .001$, exam 2, $F(1, 256) = 1.68$, $p = .20$, $\eta_p^2 = .005$, or exam 3, $F(1, 247) = 1.40$, $p = .24$, $\eta_p^2 = .005$.[4] A marginally significant interaction between ability group and feedback was found for exam 4, $F(3, 233) = 2.25$, $p = .08$, $\eta_p^2 = .02$, and is visualized in Fig. 1. The interaction was not significant for exam 1, $F(3, 260) = 1.44$, $p = .23$, $\eta_p^2 = .01$, exam 2, $F(3, 256) = 1.87$, $p = .13$, $\eta_p^2 = .02$, or exam 3, $F(3, 246) = 1.50$, $p = .22$, $\eta_p^2 = .02$.[5] To investigate the marginally significant interaction term, differences in the bias scores between the feedback conditions for each ability group was tested using two independent samples t-tests. The results indicate that students who received accuracy feedback were more overconfident than students who did not receive feedback for the low-ability and medium-low ability groups, $t(115) = 2.77$, $p < .01$, $d = 0.51$, but not for the high-ability and medium-high ability groups, $t(121) = -0.46$, $p = .65$, $d = 0.08$.

To address the third research question, growth curve modeling analysis was conducted following the procedure outlined above. The fit indices and change in chi squared tests are shown in Table 3. Chi-square tests indicate that the unconditional linear model fit better than the no-growth model. In addition, the quadratic model fit significantly better than the linear model. However, the model fit statistics (CFI and RMSEA) were unacceptable, indicating that one single model was inappropriate to model these data. Because a cubic model could not be fit with only four time points, and the spaghetti plots suggest both a cubic model and multiple

[4] The same conclusions are reached if only those who made predictions on all four exams are used in the analysis.

[5] The same conclusions are reached if only those who made predictions on all four exams are used in the analysis though the marginal interaction for exam 4 results in $p = .10$.

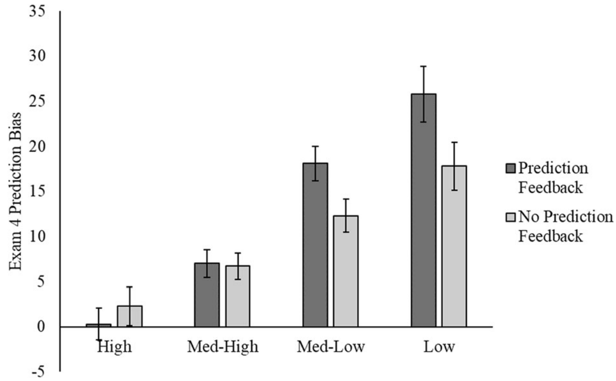

Fig. 1 Exam 4 Prediction Bias by Ability and Feedback Condition

trajectories, growth mixture modeling was conducted to determine the optimum number of classes. The results can be found in Table 4.

A four-class model was selected because, although the sample-size adjusted BIC was slightly higher than the three-class model, the entropy was higher, and the resultant models reflected the observed patterns in the data from the spaghetti plots more accurately. In the four-class model 20.7% belonged to the first class, 52.0% belonged to the second, 15.9% belonged to the third class, and 11.4% to the fourth. Fitted growth trajectories for the four classes are shown in Fig. 2. Individuals in class 1 had the lowest bias scores on three of the fours exams and were relatively consistent across the four exams compared to the other three classes. Individuals in both classes 2 and 3 showed similar patterns of large alternating increases and decreases in overconfidence across the four exams. Both groups were least overconfident on the third exam, and much more overconfident on the other three exams. However, individuals in class 2 were between 10 and 20 percentage points more overconfident on every exam than individuals in class 3. Finally, individuals in class 4 were about 20 percentage points overconfident on the first exam, then became more accurate on the second exam, and even displayed underconfidence on the third exam. This class was overconfident on the final exam, but notably reduced their bias by more than ten percentage points on average as compared to the first exam. To investigate the demographic characteristics of the individuals in each class trajectory three Chi-Square tests of independence were conducted. There were no differences in class membership between feedback condition, $\chi^2(3) = 4.40$, $p = .22$, indicating that providing feedback did not seem to affect the trajectory group that an individual was in. However, there were significant differences in class membership by ability group, $\chi^2(9) = 135.51$, $p < .001$. Students in class 3, the most overconfident class, were primarily students in the bottom quartile (77%) and did not include any students in the top quartile. Conversely,

Table 3 *Comparative Model Fit across a Series of Models*

Model	AIC	BIC	χ2	*df*	*p*	CFI	RMSEA	(90% CI)
A. No-growth model	8365.97	8376.78	240.13	11	< .001	−.517	.277	.242, .314
B. Unconditional linear growth model	8314.31	8335.92	182.47	8	< .001	−.156	.284	.241, .327
χ2 test of difference vs. Model A			57.66	3	< .001			
C. Unconditional quadratic model	8293.20	8329.22	153.35	4	< .001	.011	.371	.312, .432
χ2 test of difference vs. Model B			29.12	4	< .001			

Table 4 *Comparative Model Fit across a Series of Models*

Models	AIC	BIC	Sample-Size Adjusted BIC	Entropy	Smallest Class Percentage
Model 1 (1 class)	Did not converge				
Model 2 (2 class)	8196.25	8253.89	8203.16	.611	33.1
Model 3 (3 class)	8176.81	8266.86	8187.59	.601	14.6
Model 4 (4 class)	8180.20	8306.07	8195.10	.615	11.4
Model 5 (5 class)	8189.15	8347.64	8208.13	.613	0.4

students in class 1, the most accurate class, were primarily students in the top quartile (66%), and only included two students from the bottom quartile. The four ability groups were represented in class 4, as 23% were from the lowest quartile, 29% from the second quartile, 32% from the third quartile, and 16% from the top quartile. Students in class 2 were also relatively evenly represented among the ability groups, with 18%, 35%, 30%, and 18% respectively.

To investigate potential characteristics that might determine class membership, differences between the trajectory classes were investigated. A one-way ANOVA was conducted to investigate differences on the TOI, and a one-way MANOVA was conducted to investigate differences on the CAEB. Class membership did not differ in scores one the TOI, $F(3, 163) = 1.56$, $p = .20$, or the CAEB, $F(3, 163) = 0.37$, $p = .78$. Because the distributions for academic goal orientations were not normally distributed, four Kruskal-Wallis tests were conducted to investigate differences in academic goal orientations with a Bonferroni correction made to the critical alpha level such that $\alpha = 0.012$. Class membership did not differ by Mastery Approach, $\chi^2(3) = 2.35$, $p = .50$, Mastery Avoidance goals, $\chi^2(3) = 7.78$, $p = .05$, Performance Approach, $\chi^2(3) = 1.45$, $p = .69$, or Performance Avoidance, $\chi^2(3) = 1.10$, $p = .78$.

Discussion

This study is the first to model changes in metacognitive calibration using growth mixture modeling, and, along with Foster, et al., (2017), among the first studies to investigate changes in metacognitive calibration using person-centered techniques. Similar to Foster, et al. (2017),

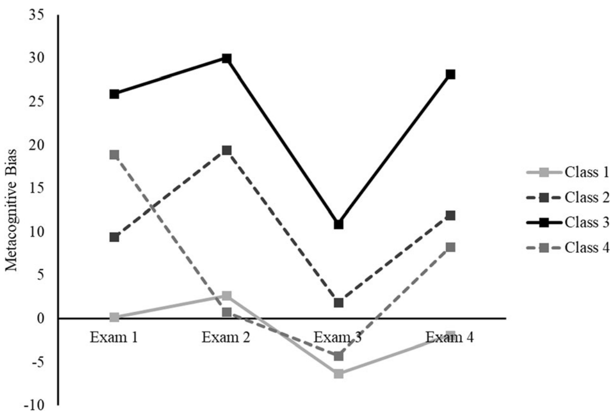

Fig. 2 Fitted Growth Trajectories for Calibration for the Four Classes

this study found that individual changes in metacognitive calibration vary greatly across individuals. As such, person centered analyses (e.g., multi-level or hierarchical modeling) may be more appropriate for identifying individual factors which are related to the ability to improve one's metacognitive monitoring. In addition, four classes were found to model changes in metacognitive bias across four exams. Although three of these classes did not exhibit significant decreases in bias, the fourth class – consisting of just over 11% of the students – demonstrated improved monitoring accuracy over the course of the semester.

Many models of self-regulated learning indicate that metacognitive monitoring is related to the goals, motivations, and epistemological beliefs of the learner (e.g., Hofer, 2004; Kitchner, 1983; Winne & Hadwin, 1998). These models suggest that interventions aimed at improving metacognitive monitoring and calibration may also need to address these other factors as well. When making predictions about upcoming exam grades, students must have substantial knowledge both within and about the domain in which the task is situated in order to make accurate predictions (Veenman, et al., 2006). This suggests that prediction accuracy should rely on the accuracy of beliefs about knowledge within a domain as well as accurate metacognitive monitoring. In this study, students who were more likely to report that they held a mastery-approach orientation were also more likely to report that ability was changeable and that knowledge in physics was simple. However, academic goal orientations largely did not correlate with epistemological beliefs. There was also not a consistent correlation between metacognitive bias, and epistemological beliefs, or academic goal orientations. In addition, these constructs were not predictive of class membership. This suggests that metacognitive monitoring accuracy (at least as measured by bias using single global exam predictions) may be orthogonal to these constructs. Alternatively, it could be that the relationship between these constructs and metacognitive monitoring accuracy may be obscured by the variation in exam difficulty, which made it more difficult to use prior exam performance to make accurate predictions about future exam performance.

Consistent with prior research, this study found that metacognitive bias was related to ability (Dunning, Heath, & Suls, 2004; Kruger & Dunning, 1999). Specifically, the lowest performing students were the least accurate, and were consistently overconfident, while the highest performing individuals were the most accurate and tended towards underconfidence. This study also found that students in the top quartile were more likely to be in class 1, which was consistently well calibrated (i.e., bias close to zero) on every exam. In contrast, students in the bottom quartile were more likely to be in class 3, which was consistently poorly calibrated and overconfident by one to three letter grades on every exam. In other words, students who earned D's and F's on the exams, generally came into the exam believing that they had prepared enough to earn B's and C's. While students in classes 2 and 4 were represented by students of all ability groups, at least 60% of the class membership was from the two middle ability groups. Students in class 4 improved their grades on the second and third exams by about ten percentage points compared to the first exam, however their bias decreased by more than 20 percentage points. This suggests that the improvements in monitoring accuracy may have been the result of improvements in both performance and metacognition monitoring. Alternatively, the observed improvements in performance could be the result of improved monitoring accuracy. Future research should look to investigate these possibilities.

One common explanation for the pattern where individuals overestimate their own performance on exams, with the overestimates being more pronounced for low-performing students, is that the expertise and skills needed to produce good performance on a task are the same type of expertise and skills needed to produce accurate judgments of performance (Schlosser, et al., 2013). In other words, low-performing students suffer from the dual curse of being both unskilled and unaware of their lack of skill (Kruger & Dunning, 1999). This interpretation suggests that providing students with feedback about the accuracy of their exam predictions may help students gain metacognitive awareness, which could either help students improve their performance, or at least to regulate their overconfidence. However, in contrast to Miller and Geraci (2011), providing students with feedback did not lead to improvements in monitoring accuracy. In fact, paradoxically, the results suggest that providing students with accuracy feedback may have resulted in greater overconfidence by low-performing students.

While this study was not designed to investigate the causes for inflated overconfidence, there are a few possible explanations for this finding. One explanation is that providing students with feedback about prediction accuracy does not help them develop productive study strategies. Prior research has found that low-performing students tend to utilize more passive and less effective study strategies when studying for exams (Hartwig & Dunlosky, 2012; Karpicke, Butler & Roediger, 2009). Passive study strategies such as reviewing notes, or rewatching online lectures can lead to increased familiarity with the content without leading to enhanced learning. Because familiarity and fluency are heuristic-based cues commonly used to make metacognitive judgments, these study methods may result in overconfident predictions about one's preparation (Koriat, 1997). Providing students with prediction feedback without social interaction or instruction on how to use this feedback to make changes to their studying may prompt students to engage in more studying using the same ineffective strategies that they were using, then use the increase in the amount of study time to make higher predictions about exam performance. As noted by Labuhn, Zimmerman, and Hasselhorn (2010), feedback may prompt individuals to use self-regulatory processes, but does not necessarily direct individuals on which processes will be helpful or productive for self-regulation. In other words, low-performing students may make overconfident predictions on subsequent exams because they engaged in more studying than for previous exams, even though the additional studying was not helpful in improving performance. This explanation is consistent with recent experimental findings showing that providing low-performing students with accuracy feedback can lead to greater overconfidence and less accurate control choices (Raaijmakers, Baars, Paas, van Merrienboer, & van Gog, 2019). All of this suggests that future research should focus on using metacognitive strategy training and self-assessment monitoring to help students improve metacognitive monitoring (e.g., Dunlosky & Metcalfe, 2008; Kostons, VanGog & Paas, 2012).

An alternative explanation for this finding is that providing feedback about prediction accuracy may motivate students to make overconfident predictions in an attempt to maintain a positive self-image. A final explanation is that students who received feedback may have been more likely to use the feedback from the third exam, which was also the easiest exam, when making predictions for the final exam compared to those who did not receive feedback. Because this was an unanticipated finding, future research should attempt to replicate this finding using experimental methods designed to investigate students' study habits and rationale for student predictions.

Should the findings from this study prove to be robust, it would suggest that interventions aimed at improving calibration need to be sensitive to differences in ability levels. Lower-performing students may need interventions that incorporate reality checks, such as required practice tests that ask students to predict their performance followed by accuracy feedback and suggested study strategies. While much of the literature on metacognitive calibration has focused on the correlations with performance, the extent to which overconfidence or underconfidence are related to availing constructs such as self-efficacy or persistence has not been extensively studied. For example, students who make overconfident predictions may exhibit overconfidence, in part, to maintain a positive self-image. If their overconfidence is related to their self-efficacy for a task, overconfidence could potentially encourage persistence, as long as the student can maintain a reasonable level of success.

In addition to students' general belief for when they are sufficiently prepared for an exam, the ability to resolve easy and difficult items is equally important for students as they prepare for an exam. In fact, Schwartz and Efkides (2015) suggest that metacognitive judgments of learning are very effective when students use them for making decisions on what to study, suggesting that measures of relative accuracy are important for investigating the effect of metacognition on studying. This study used measures of absolute accuracy to investigate the magnitude of the discrepancy between judgment and performance for both practical and theoretical reasons. However, measures of absolute and relative accuracy sometimes yield different findings. For example, relative accuracy does not seem to vary by ability the same way as absolute accuracy (e.g., Keleman, Winningham, & Weaver, 2007; Maki, Shields, Wheeler, & Zacchilli, 2005; Ozuru, Kurby, McNamara, 2012). An area for future research would be to incorporate local item-by-item judgments on exams to investigate how measures of relative accuracy change over the course of a semester, in addition to, correlations between measures of relative accuracy and goal orientations and epistemological beliefs.

Another area for future research is the degree to which low-performing students demonstrate metacognitive awareness. While inaccurate metacognitive predictions could indicate a lack of awareness, it may also simply reflect a lack of understanding of the material that will appear on the exam. One way to measure metacognitive awareness could be to look at how individuals change their estimate of performance after taking an exam. A student who initially overpredicts may demonstrate metacognitive awareness by adjusting their estimate downward after completing an exam. If Kruger and Dunning (1999) are correct in asserting that lower performing individuals are less accurate in their metacognitive monitoring, then we would expect higher performing individuals to make more appropriate adjustments in their performance estimates from before the exam to after the exam. Future research should have students make predictions both before and after an exam to determine the type of student who is more accurate in adjusting their performance estimates and what factors are associated with accuracy in these adjustments.

There were limitations to this study. First, measures of goal orientation, epistemological beliefs, and demographic information was obtained for only about 60% of the sample. While the demographic data of those that completed the surveys was representative of the entire course, there may be a systematic difference on these measures between those who complete assignments early and those who wait until later in the week to complete their assignments. The lack of association between the motivational variables we collected and the trajectories of growth suggests the need to search for other covariates. For example, it is likely that an individual's prediction before knowing the specific questions on an exam is influenced by both their metacognitive abilities and their epistemological beliefs about the complexity of knowledge and the speed of learning. Future work should investigate the effect of epistemological beliefs on initial metacognitive accuracy and improvement over time. Another limitation of the study is that there was not a way to measure how strongly (or if) students considered the feedback. It is possible that some of the students did not meaningfully consider the accuracy feedback because of a lack of social interaction or scaffolding to ensure that the feedback was meaningfully attended to and considered. Future research should investigate how providing accuracy feedback with scaffolding and instruction for students impacts the development of metacognitive monitoring accuracy.

A final limitation is that no model that was run for the growth curve modeling exhibited ideal model fit. One reason for this lack of fit likely lies in the large variation in monitoring accuracy trajectories between individuals. This suggests that models assuming a single trajectory type my not be sufficient for analyzing change in monitoring accuracy over time. Another reason for the lack of fit during growth curve monitoring is the variation in exam difficulty between exams 2 and 3. The exams for this course, were written by course instructors with the goal of having the mean on the exam between 70 and 75%. Individual variations by course instructors result in exams with mean scores higher or lower than this target. Future work should investigate monitoring by using exams with known psychometric properties to control for differences in exam difficulty between exams over the course of a semester.

Compliance with ethical standards

Conflict of interest The author declares that they have no conflicts of interest.

References

Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General, 143*, 1349–1368. https://doi.org/10.1037/a0035098.

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39*, 171–184. https://doi.org/10.3758/s13421-010-0002-y.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138*, 432–447. https://doi.org/10.1037/a0015928.

Baars, M., Vink, S., Van Gog, T., De Bruin, A., & Paas, F. (2014). Effects of training self-assessment and usingassessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107. https://doi.org/10.1016/j.learninstruc.2014.04.004.

Barzilai, S., & Zohar, A. (2014). Reconsidering personal epistemology as metacognition: A multifaceted approach to the analysis of epistemic thinking. *Educational Psychologist, 49*, 13–35. https://doi.org/10.1080/00461520.2013.863265.

Belenky, D. M., & Nokes-Malach, T. J. (2013). Mastery-approach goals and knowledge transfer: An investigation into the effects of task structure and framing instructions. *Learning and Individual Differences, 25*, 21–34. https://doi.org/10.1016/j.lindif.2013.02.004.

Bipp, T., Steinmayr, R., & Spinath, B. (2012). A functional look at goal orientations: Their role for self-estimates of intelligence. *Learning and Individual Differences, 22*, 280–289. https://doi.org/10.1016/j.lindif.2012.01.009.

Bromme, R., Pieschl, S., & Stahl, E. (2010). Epistemological beliefs are standards for adaptive learning: A functional theory about epistemological beliefs and metacognition. *Metacognition and Learning, 5*, 7–26. https://doi.org/10.1007/s11409-009-9053-5.

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2015). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*, 215–235. https://doi.org/10.1007/s11409-015-9142-6.

Castel, A. D., Benjamin, A. S., Craik, F. I., & Watkins, M. J. (2002). The effects of aging on selectivity and control in short-term recall. *Memory & Cognition, 30*, 1078–1085. https://doi.org/10.3758/BF03194325.

Cromley, J. G., Perez, T., & Kaplan, A. (2016). Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights From the Behavioral and Brain Sciences, 3*, 4–11. https://doi.org/10.1177/2372732215622648.

Coutinho, S. A. (2007). The relationship between goals, metacognition, and academic success. *Educate, 7*, 39–47.

Dunlosky, J., & Bjork, R. (2008). The integrated nature of memory and metamemory. In J. Dunlosky & R. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 11–28). New York, NY: Psychology Press.

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

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Los Angeles: SAGE.

Dunlosky, J., & Thiede, K. W. (2013). Metamemory. In D. Reisberg (Ed.), *The Oxford handbook of psychology* (pp. 283–298). Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0019.
Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x.
Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Taylor and Francis/Psychology Press.
Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256–273. https://doi.org/10.1037/0033-295X.95.2.256.
Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002.
Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*, 613–628. https://doi.org/10.1037/0022-0663.100.3.613.
Erdmann, R., Miller, K., & Stains, M. (2020). Exploring STEM postsecondary instructors' accounts of instructional planning and revisions. *International Journal of STEM Education, 7*(1), 1–17. https://doi.org/10.1186/s40594-020-00206-7.
Fakcharoenphol, W., Morphew, J.W., & Mestre, J.P. (2015). Judgments of physics problem difficulty among experts and novices. *Physical Review Physics Education Research, 11*(2), 020108. https://doi.org/10.1103/PhysRevSTPER.11.02012.
Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19–34. https://doi.org/10.1016/j.jml.2007.03.006.
Flavel, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist, 34*, 906–911. https://doi.org/10.1037/0003-066X.34.10.906.
Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition & Learning, 12*, 1–19. https://doi.org/10.1007/s11409-016-9158-6.
Geurten, M., & Meulemans, T. (2017). The effect of feedback on children's metacognitive judgments: A heuristic account. *Journal of Cognitive Psychology, 29*, 184–201. https://doi.org/10.1080/20445911.2016.1229669.
Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. Farr (Eds.), *The Nature of Expertise* (pp. xv-xxvii). Hillsdale, NJ: Erlbaum.
Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*, 334–372. https://doi.org/10.3102/003465430303953.
Gul, F., & Shehzad, S. (2012). Relationship between metacognition, goal orientation and academic achievement. *Procedia – Social and Behavioral Sciences, 47*, 1864–1868. https://doi.org/10.1016/j.sbspro.2012.06.914.
Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*, 101–121. https://doi.org/10.1007/s11409-008-9021-5.
Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170. https://doi.org/10.1037/0022-0663.92.1.160.
Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin and Review, 19*, 126–134. https://doi.org/10.3758/s13423-011-0181-y.
Henderson, Mestre, & Slakey, 2015.
Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online-searching. *Educational Psychologist, 39*, 43–55. https://doi.org/10.1207/s15326985ep3901_5.
Hong, Y-Y., Chiu, C-Y., & Dweck, C. S. (1995). Implicit theories of intelligence: Reconsidering the role of confidence in achievement motivation. In M. H. Kernis (Ed.), Efficacy, agency, and self-esteem (pp. 197–216). New York, NY: Springer Science and Business Media. DOI: https://doi.org/10.1007/978-1-4899-1280-0_10.
Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*, 53–60.
Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471–479. https://doi.org/10.1080/09658210802647009.

Kelemen, W. L., Winningham, R. G., & Weaver III, C. A. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology, 19*, 689–717. https://doi.org/10.1080/09541440701326170.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*, 157–175. https://doi.org/10.1006/jmla.1996.0009.

Keren, G. (1990). Cognitive aids and debiasing methods: Can cognitive pills cure cognitive ills? *Advances in Psychology, 68*, 523–552. https://doi.org/10.1016/S0166-4115(08)61341-2.

Kim, J. H. (2018). The effect of metacognitive monitoring feedback on performance in a computer-based training simulation. *Applied Ergonomics, 67*, 193–202. https://doi.org/10.1016/j.apergo.2017.10.006.

King, B. (2015). Changing college majors: Does it happen more in STEM and do grades matter? *Journal of College Science Teaching, 44*, 44–51. https://doi.org/10.2505/4/jcst15_044_03_44.

Kitchner, K. S. (1983). Cognition, metacognition, and epistemic cognition: A three-level model of cognitive processing. *Human Development, 26*, 106–116.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370. https://doi.org/10.1037/0096-3445.126.4.349.

Koriat, A. (2007). Metacognition and consciuousnees. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–325). Cambridge, UK: Cambridge University Press.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36–69. https://doi.org/10.1037/0096-3445.135.1.36.

Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1624–1637. https://doi.org/10.1037/xlm0000009.

Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*, 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004.

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*, 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.

Laburn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*, 173–194. https://doi.org/10.1007/s114909-010-9056-2.

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*, 307–318. https://doi.org/10.1037/a0017599.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723–731. https://doi.org/10.1037/0022-0663.97.4.723.

Mason, L., Boldrin, A., & Ariasi, N. (2010). Epistemic metacognition in context: Evaluating and learning online information. *Metacognition and Learning, 5*, 67–90. https://doi.org/10.1007/s11409-009-9048-2.

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

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477. https://doi.org/10.1016/j.jml.2004.12.001.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*, 303–314. https://doi.org/10.1007/s11409-011-9083-7.

Morphew, J.W., Gladding, G.E., & Mestre, J.P. (2020). Effect of presentation style and problem-solving attempts on metacognition and learning from solution videos. *Physical Review Physics Education Research, 16*(1), 010104. https://doi.org/10.1103/PhysRevPhysEducRes.16.010104.

Muis, K. R., & Franco, G. M. (2010). Epistemic profiles and metacognition: Support for the consistency hypothesis. *Metacognition and Learning, 5*, 27–45. https://doi.org/10.1007/s11409-009-9041-9.

Nagin, D. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology, 6*, 109–138. https://doi.org/10.1146/annurev.clinpsy.121208.131413.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York, NY: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–26). Cambridge, MA: MIT Press.

Nietfeld, J. L., Cao, L., & Osborne, J. (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. https://doi.org/10.1007/s10409-006-9595-6.

Ozuru, Y., Kurby, K. A., & McNamara, D. S. (2012). The effect of metacomprehension judgment task on comprehension monitoring and metacognitive accuracy. *Metacognition and Learning, 7*, 113–131. https://doi.org/10.1007/s11409-012-9087-y.

Pieschl, S., Stahl, E., & Bromme, R. (2008). Epistemological beliefs and self-regulated learning with hypertext. *Metacognition and Learning, 3*, 18–27. https://doi.org/10.1007/s11409-007-9008-7.

Pintrich, P. P. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Into Practice, 41*, 219–225. https://doi.org/10.1207/s15430421tip4104_3.

Raaijmakers, S. F., Baars, M., Paas, F., van Merrienboer, J. J. G., & van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning, 14*, 21–42. https://doi.org/10.1007/s11409-019-09189-5.

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*, 559–579. https://doi.org/10.1080/09541440701326022.

Rebello, N. S. (2012). How accurately can students estimate their performance on an exam and how does this relate to their actual performance on the exam? In N. S. Rebello, P. V. Engelhardt, & C. Singh (Eds.), *AIP Conference Proceedings (Vol. 1413, pp. 315–318)*. AIP. https://doi.org/10.1063/1.3680058.

Saenz, G. D., Geraci, L., Miller, T. M., & Tirso, R. (2017). Metacognition in the classroom: The association between students' exam predictions and their desired grades. *Consciousness and Cognition, 51*, 125–139. https://doi.org/10.1016/j.concog.2017.03.002.

Simons, D. J. (2013). Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin & Review, 20*, 601–607. https://doi.org/10.3758/s13423-013-0379-2.

Schlosser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning-Kruger effect in self-evaluation of performance. *Journal of Economic Psychology, 39*, 85–100. https://doi.org/10.1016/j.joep.2013.07.004.

Schneider, W. (2002). Memory development in childhood. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 236–256). Malden, MA: Blackwell.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*, 498–504. https://doi.org/10.1037/0022-0663.82.3.498.

Schwartz, B. L., & Efklides, A. (2015). Metamemory and memory efficiency: Implications for student learning. *Journal of Applied Research in Memory and Cognition, 1*, 145–151. https://doi.org/10.1016/j.jarmac.2012.06.002.

Scott, B. M., & Levy, M. G. (2013). Metacognition: Examining the components of a fuzzy concept. *Educational Research, 2*, 120–131. https://doi.org/10.5838/erej.2013.22.04.

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

Soderstrom, N. C., Yue, C. L., & Bjork, E. L. (2015). Metamemory and education. In J. Dunlosky and S. K. Tauber (Eds.) *The Oxford Handbook of Metamemory*. Oxford, UK: Oxford University Press. Retrieved from http://www.oxfordhandbooks.com/view/10.1093 /oxfordhb /9780199336746.001.0001/oxfordhb-9780199336746-e-6.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 204–221. https://doi.org/10.1037/0278-7393.26.1.204.

Stahl, E., & Bromme, R. (2007). The CAEB: An instrument for measuring connotative aspects of epistemological beliefs. *Learning and Instruction, 17*, 773–785. https://doi.org/10.1016/j.learninstruc.2007.09.016.

Tauber, S. K., & Dunlosky, J. (2015). Monitoring of learning at the category level when learning a natural concept: Will task experience improve its resolution? *Acta Psychologica, 155*, 8-18. DOI: 10/1016/j.actpsy.2014.11.011.

Tauber, S. K., & Rhodes, J. (2012). Multiple bases for young and older adults' judgments of learning in multiblock learning. *Psychology and Aging, 27*, 474–483. https://doi.org/10.1037/a0025246.

Thompson, W. B. (1998). Metamemory accuracy: Effects of feedback and the stability of individual differences. *The American Journal of Psychology, 111*, 33–42.

Tobias, S., & Everson, H. T. (2009). The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 107–127). New York, NY: Routledge.

Tuysuzoglu, B. B., & Greene, J. A. (2015). An investigation of the role of contingent metacognitive behavior in self-regulated learning. *Metacognition and Learning, 10*, 77–98. https://doi.org/10.1007/s11409-014-9126-y.

Urban, K., & Urban, M. (2019). Improving the accuracy of the self-evaluation during on-screen self-regulated learning through calibration feedback. In *INTED2019 proceedings (pp. 9002-9007)*. Spain.

Urban, K., & Urban, M. (2018). Influence of fluid intelligence on accuracy of metacognitive monitoring in preschool children fades with the calibration feedback. *Studia Psychologica, 60*, 123–136. https://doi.org/10.21909/sp.2018.02.757.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14. https://doi.org/10.1007/s11409-006-6893-0.

Wilkinson, S. C., Reader, W., & Payne, S. J. (2012). Adaptive browsing: Sensitivity to time pressure and task difficulty. *International Journal of Human-Computer Studies, 70*, 14–25. https://doi.org/10.1016/j.ijhcs.2011.08.003.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah: Erlbaum.

Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). New York, NY: Erlbaum.

Zimmerman, B. J. (2000). Attainment of self-regulation: A social cognitive perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Self-regulation: Theory, research, and applications* (pp. 13–39). Orlando: Academic.

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