2016-07-01

Developmental Math Students' Calibrated Judgments of Learning

Brian Lindley Jones
Brigham Young University

Developmental Math Students' Calibrated Judgments of Learning

Brian Lindley Jones

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Randall S. Davies, Chair
Peter J. Rich
David D. Williams

Department of Instructional Psychology and Technology

Brigham Young University

July 2016

ABSTRACT

Developmental Math Students' Calibrated Judgments of Learning

Brian Lindley Jones
Department of Instructional Psychology and Technology, BYU
Master of Science

Calibrated Judgments of Learning (CJOL) represent the degree to which students' judgments of learning (JOL) relate to their actual learning. Although a substantial amount of research has been conducted on calibration and JOL in various domains of psychology, only a growing number of studies have begun to address the use of CJOL in applied educational settings. This study investigated the use of CJOL in university developmental math courses. Study participants included 185 men and 100 women with ages ranging from 18 to 61 years (M = 23.48, SD = 5.95). Study results indicate that these developmental math students were fairly accurate in their perceptions of their math performance. When inaccurate, students most commonly under estimated their performance. Students' accuracy was also greatly influenced by the difficulty of math questions on the tests. High performing students were consistently more accurate than lower performing students. Over the course of the study, students received feedback on their accuracy in an attempt to facilitate improved accuracy. Results indicated that students' accuracy decreased with time; likely this was due to the increase in the difficulty of math questions on each test.

Keywords: judgment of learning, calibration, metacognitive judgments, metacognitive monitoring, self-regulation, developmental math

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Over the past several years, the topic of *college and career readiness* has been at the center of many K-12 educational discussions (Mishkind, 2014). However, the push for higher standards of accountability and achievement are not new. Efforts to improve education have been a notable focus of discourse and policy in the United States since the 1983 publication of the Elementary and Secondary Education report, *A Nation at Risk: The Imperative for Educational Reform* (Miller, Linn, & Gronlund, 2013). Pressures to hold secondary education institutions in the United States accountable for students' readiness for college and career was influential in over 20 states and 25 independent school districts' decision to fund students to take the American College Testing Inc. (ACT) college readiness assessment (Adams, 2014). Unfortunately, according to a report released by ACT in 2014, approximately 74 percent of high school graduates did not meet readiness benchmarks for all four tested subjects on the ACT College Readiness Assessment (Act, 2014). This high percent of underprepared students not only reflects the struggles of K-12 institutions to adequately prepare students to be college and career ready, but also reflects a challenge faced by postsecondary institutions to accommodate these underprepared students.

Many postsecondary institutions accommodate underprepared students by providing opportunities for remedial subject-specific coursework and the development of academic skills (Sparks & Malkus, 2013). The main objective of these remedial courses is to provide students with foundational content knowledge in preparation for postsecondary level coursework. In addition to subject-specific remedial coursework, postsecondary institutions often seek to develop students' general academic abilities (Kuh, Kinzie, Buckley, Bridges, & Hayek, 2006).

Several researchers have found that student success (i.e., academic outcomes) in higher education and lifelong learning are related to a general academic ability of self-regulated learning (de Bruijn-Smolders, Timmers, Gawke, Schoonman, & Born, 2016). Self-regulated learning is the systematic regulation of thoughts, feelings, and actions towards learning goals (Zimmerman & Schunk, 2011). A critical component to many models of self-regulation is a metacognitive monitoring process wherein students accurately understand and interpret their thoughts, feelings, and actions so they can systematically regulate their learning efforts (Boekaerts, Pintrich, & Zeidner, 2005; Greene & Azevedo, 2007; Zimmerman & Schunk, 2001).

This metacognitive monitoring process is crucial for students, since information obtained through monitoring is the basis for subsequent evaluations of learning and the regulation of academic behavior (Stone, 2000). It can be particularly difficult for students to effectively regulate their learning when there is inaccurate or inefficient metacognitive monitoring. For example, students may choose not to study for a final exam if they have incorrectly believed that they had mastered the content for an exam (Dunlosky & Connor, 1997; Son & Metcalfe, 2000). This disconnect between what the student thinks they know and what they actually know can limit the student's ability to effectively evaluate and regulate their learning efforts.

Unfortunately, many students enter institutions of higher education with a limited ability to monitor and regulate their learning (Ley & Young, 1998). Student success courses seek to help students develop these skills. However, these programs often only indirectly address metacognitive monitoring while focusing primarily on developing students' academic behaviors, such as note taking, study skills, and reading strategies (Kuh et al., 2006). Many methods for explicitly developing students' metacognitive monitoring are also cumbersome to implement, require efforts that can distract from the academic task at hand, and have limited evidence of

effectiveness. Such methods include think aloud protocols, journaling, and self-report questionnaires (Roth, Ogrin, & Schmitz, 2015). Calibrated judgments of learning (CJOL) provide an alternative method for improving and assessing metacognitive monitoring (Stone, 2000).

CJOL are objective measures of metacognitive monitoring that represent the degree to which students' judgments of learning (JOL) mirror actual learning (Alexander, 2013). The more closely a student's JOL mirrors actual learning, the more the student is said to be academically self-aware. CJOL consist of two main components, a JOL and a performance measure. JOL provide a snapshot into students' monitoring processes by asking students to report on the degree to which they believe they have accomplished a specified learning task (e.g., text comprehension, solving an algebraic equation; Van Overschelde & Nelson, 2006). The learning-performance measure indicates the degree to which the students actually accomplished the specified learning task. Calibrating the JOL and the performance measure, through mathematical comparisons, provides an objective measure of the accuracy of the students' beliefs about their learning or performance (i.e., academic self-awareness).

Statement of the Problem

CJOL have been studied in the field of psychology since the 1970s to help psychologists understand various psychological phenomena (Shaughnessy, 1979). In contrast, CJOL are generally unknown in the broader practitioner-based educational community. Despite their lack of prevalence among practitioners, researchers have demonstrated that students' academic outcomes are significantly related to the degree to which they are academically self-aware (Bol, Hacker, O'Shea, & Allen, 2005; Hartwig, Was, Isaacson, & Dunlosky, 2012; Meier, von Wartburg, Matter, Rothen, & Reber, 2011; Roebers, Schmid, & Roderer, 2009). This

relationship suggests that academic outcomes can, in part, be improved through an increase in academic self-awareness. The use of CJOL to assess and track academic self-awareness could be instrumental in positively influencing changes in academic outcomes.

Statement of Purpose

This research aimed to address academic self-awareness of developmental mathematics students through the use of CJOL. More specifically, this study used calibrated judgments of learning (CJOL) to investigate the degree to which :(a) students are academically self-aware and (b) academic self-awareness changes over time.

Research Questions

This study will address the following research questions:

1. To what degree do developmental math students' perceptions of their performance accurately match their actual performance (CJOL accuracy)?
2. To what degree do developmental math students' CJOL become more accurate over time after receiving feedback on their accuracy?
3. To what degree do differences in CJOL accuracy exist amongst disaggregated groups of age, gender, year in school, and course level?

Review of Literature

Calibrated judgments of learning (CJOL) research can be found in education and psychology. In order to identify research in both domains, the ERIC and PsycINFO databases were searched using the EBSCOhost database search tool. CJOL can be conceptually defined as the degree to which student judgments of learning mirror actual learning. The terminology used to address this conceptual definition can vary widely depending on the context of the research. Due to this variation, a diverse collection of search terms were used in order to locate relevant

peer reviewed research. The specific EBSCO search terms used for this review of literature are presented in Table 1.

Table 1

EBSCO Search Terms Used for the Literature Review

("calibration accuracy" AND student) OR	("judgment of performance" AND student) OR
("calibration construct" AND student) OR	("judgments of confidence" AND student) OR
("confidence accuracy" AND student)	("metacognitive accuracy" AND student) OR
("confidence judgement" AND student) OR	("metacognitive calibration" AND student) OR
("confidence judgements" AND student) OR	("metacognitive judgement" AND student) OR
("confidence judgment" AND student) OR	("metacognitive judgements" AND student) OR
("confidence judgments" AND student) OR	("metacognitive judgment" AND student) OR
("JOL" AND student) OR	("metacognitive judgments" AND student) OR
("JOP" AND student) OR	("metacognitive monitoring" AND student) OR
("judgement accuracy" AND student) OR	("metacomprehension accuracy" AND student) OR
("judgement of confidence" AND student) OR	("monitoring accuracy" AND student) OR
("judgement of performance" AND student) OR	("performance judgement" AND student) OR
("judgements of confidence" AND student) OR	("performance judgements" AND student) OR
("judgment accuracy" AND student) OR	("performance judgment" AND student) OR
("judgment of confidence" AND student) OR	("performance judgments" AND student) OR
	("self-assessment accuracy" AND student) OR

The initial search, using all search terms, produced 632 results. These results were then systematically reviewed to determine the extent to which they met four criteria for inclusion: (a) Articles must report on research participants making judgments about their learning, (b) Articles must report on the accuracy of the participant's judgments of their learning, (c) Articles must report on the specific methods used for collecting judgments of learning and measuring accuracy, and (d) Articles must be published in a peer-reviewed journal. These criteria reflected the critical information needed in order to adequately address the literature review questions. From the initial search, 185 articles were found that met the four criteria for inclusion.

A secondary search analyzed the reference list for the most relevant articles selected in the initial search. The purpose of the secondary search was to identify seminal articles addressing CJOL that may have been omitted by the initial search. Articles cited by multiple review articles were selected for further analysis. These articles were included in the review if

they met the criteria for inclusion established in the initial search. The secondary search for articles resulted in the inclusion of 57 additional articles in the review. Following the primary and secondary search for literature, a total of 242 articles were selected for review.

To more fully understand the relationship between underprepared university students' academic self-awareness and their academic performance, this review will (1) establish a conceptual definition of CJOL, (2) review how academic self-awareness and CJOL can influence academic outcomes, (3) review CJOL applications in applied educational contexts, and (4) review appropriate CJOL methodological considerations.

Defining Calibrated Judgments of Learning

The limited implementation of calibrated judgments of learning (CJOL) in applied educational contexts may be due, in part, to the dispersed body of research addressing the accuracy of students' perceptions of their learning and assessment performance. A literature base spanning several domains of research presents the challenge of domain-specific terms and definitions. Although useful in their appropriate context, domain-specific terms and definitions can hinder broader conceptual treatments of an idea. This review will define and use practitioner-centered terms and definitions (i.e., more general and conceptual in nature) in order to overcome these challenges and facilitate a cross-disciplinary treatment of the subject. For the purpose of this review, CJOL will be conceptually defined as the degree to which students' judgments of learning correspond to their actual learning. Three main conceptual components make up the CJOL namely, a judgment of learning (JOL), a learning-performance measure, and a comparison between the JOL and performance measure.

Judgments of learning (JOL). JOL will be defined as any judgment made by a student regarding the degree to which they believe they have accomplished a specified learning task.

The order in which JOL and performance measures occur differentiates many domain-specific definitions. Researchers often refer to JOL that occur before the performance measure as predictions (Destan & Roebers, 2015; Lin & Zabrucky, 1998; Maki & McGuire, 2002). In a prediction context, JOL are most often students' judgments of how well they learned or understood instructional material (e.g., how well they feel they will perform on a test before seeing or taking the test). Researchers often refer to JOL which occur after the performance measure as confidence judgments or postdictions (Destan & Roebers, 2015; Hacker, Bol, Horgan, & Rakow, 2000; Koriat, Lichtenstein, & Fischhoff, 1980; Shaughnessy, 1979). In the context of confidence judgments or postdictions, JOL are most often students' judgments of the likelihood that their response to the performance measure was correct (i.e., how well they feel they performed on a test question after answering the question).

Performance measure. Performance measures will be defined as any assessment designed to measure the degree to which students have accomplished a specified learning task. Researchers have used a wide variety of performance measures across various domains (Miller, Linn, & Gronlund, 2013). The appropriateness of the learning-performance measure depends largely on the characteristics of the learning task.

Calibration. Calibration will be defined as the comparison of students' JOL to their performance on the performance measure. There are a wide variety of methods used to calibrate JOL and performance measures (Maki, Shields, Wheeler, & Zacchilli, 2005; Schraw, 2009). The appropriateness of these methods depends on the type of inferences to be drawn from the calibration.

Academic Self-awareness and Academic Outcomes

Academic outcomes can be thought of as the degree to which educational goals are met. At a course level, academic outcomes may be measured by student performance on course-specific assessments. Many institutions and courses also adopt academic goals relating to the affective and social domains of education such as lifelong learning. Researchers in the fields of self-regulation and CJOL have found that academic outcomes can be significantly influenced by students' academic self-awareness. For example, researchers have found that factors related to students' academic self-awareness predicted initial college performance far better than SAT and ACT scores (Credé & Kuncel, 2008; Shivpuri, Schmitt, Oswald, & Kim, 2006). This strong relationship between students' academic self-awareness and their academic performance suggests that both students and educators should be concerned with the degree to which students are academically self-aware. The following sections will briefly highlight research from domains of self-regulation and CJOL to illustrate the importance of academic self-awareness to the achievement of academic outcomes.

A self-regulation perspective. Researchers of self-regulation have found that students' academic self-awareness is significantly associated with academic performance (Chung, 2000; Paris & Paris, 2001). In one of many studies highliting the relationship beween academic self-awarenss and academic outcomes, Kitsanas (2002) studied the relationship between self-regulatory processes and students' abilities to prepare for and complete course assessments. The self-regulatory process of metacognitive monitoring was particularly relevant to understanding the relationship between academic self-awareness and academic outcomes. In Kitsanas' study, 62 undergraduate psychology students participated in course assessments and structured self-regulation interviews throughout a semester. Course outcomes were measured by three

psychology assessments consisting of 30 multiple-choice questions. The self-regulation interviews followed the 15-item interview questionnaire developed by Zimmerman and Pons (1986). Kitsanas (2002) found that students with high performance on course assessments also reported more self-regulatory processes than that of lower performing students. Among other self-regulatory processes, high performing students reported more frequent monitoring and awareness of their academic performance. From their self-monitoring, high performing students reported making judgments of their learning and took steps to remediate their misunderstandings. Kitsanas' (2002) findings corroborated the findings of many other researchers of self-regulation (Schunk & Zimmerman, 1997; Zimmerman & Kitsantas, 1996, 1997).

In addition to research supporting the influence of metacognitive monitoring on academic outcomes, researchers have found that metacognitive monitoring can be taught and improved. Dignath, Buettner, and Langfeldt (2008) conducted a meta-analysis of instructional practices used to improve self-regulatory processes. The main criteria for the inclusion of studies in the meta-analysis were (a) study participants be students age 12 and under, (b) studies conducted in a real classroom setting, and (c) study methodology included a control group with longitudinal measurement. In total, 48 studies were selected for inclusion. Although the study participants in this meta-analysis were elementary-aged students, the findings associated with metacognitive monitoring (academic self-awareness) are equally relevant to older students populations (Krebs & Roebers, 2010).

In this study, Dignath et al. (2008) found there was substantial evidence that self-regulatory processes can be taught and learned. Their findings confirmed earlier research that indicates that self-regulation instruction is most effective when carried out within the context of authentic learning environments (i.e., integrated into a course such as mathematics) (Perels,

Gurtler, & Schmitz, 2005). The range of effect sizes also indicate that not all instructional practices are equal in their effectivness. Dignath et al. found that combining the instruction of metacognitive monitoring strategies with motivational or cognitive strategies resulted in the highest effect size for developing academic self-awareness. In addition, the combination of instructional strategies produced significantly higher gains in academic outcomes compared to the use of any single strategy. The findings from the 48 studies reported by Dignath et al. support the idea that metacognitive monitoring strategies can be taught and learned to improve academic self-awareness and in turn academic outcomes.

A calibrated judgment of learning perspective. Researchers of CJOL have found that the accuracy of academic self-awareness is significantly associated with academic outcomes and performance (Bol et al., 2005; Hartwig et al., 2012; Meier et al., 2011; Roebers et al., 2009). Many researchers studying CJOL have investigated the accuracy of students' metacognitive monitoring. This perspective is notably different from a self-regulation perspective of metacognitive monitoring where researchers are generally more concerned with the behaviors resulting from monitoring and not the accuracy of the monitoring.

In a study of 27 undergraduate students in a teacher education program, Nietfeld, Cao, & Osborne (2005) studied the relationship between academic outcomes and students' CJOL accuracy. Throughout the course of the semester, academic outcomes were measured with three 25-question multiple-choice tests and one 50-question comprehensive test. Students reported judgments of learning (JOL) both at a global level (i.e., confidence in their overall performance on the test) and item level (i.e., confidence in their performance on each test question). Reports of JOL were collected using a continuous scale ranging from 0% accurate to 100% accurate.

Nietfeld et al. (2005) found students' CJOL accuracy was significantly related to test performance and grade point average (GPA). This relationship suggests the presence of a more general academic self-awareness extending beyond a single course (Hartwig et al., 2012; Schraw & Dennison, 1994). The strong relationship between the accuracy of students' CJOL and academic outcomes merits attempts to improve CJOL accuracy to improve student outcomes (Bol et al., 2005; Hartwig et al., 2012; Meier et al., 2011; Nietfeld et al., 2005; Roebers et al., 2009).

Researchers have long recognized the importance of improving academic self-awareness and have sought to research the improvement of CJOL accuracy. However, findings from these studies have been mixed. Several researchers have reported significant improvements in academic self-awareness (Hacker et al., 2000; Koriat & Goldsmith, 1994, 1996; Nietfeld & Schraw, 2002; Pressley, Snyder, Levin, Murray, & Ghatala, 1987; Walczyk & Hall, 1989). In contrast, others have reported no significant gains in academic self-awareness (Bol & Hacker, 2001; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Koriat et al., 1980; Koriat, 1997). These discrepancies curtail a definitive conclusion on whether or not CJOL accuracy can be improved. Discrepancies in research findings may be due, in part, to differing research domains and research methodologies. A reduction of methodological differences and measurement error could possibly be achieved through an increase in research specifically targeting CJOL methodology (Lin, Moore, & Zabrucky, 2001; Schraw, 2009; Was, 2014).

One methodological shortcoming of CJOL research is the lack of longitudinal studies. The majority of CJOL studies are either single-sitting laboratory studies or studies that do not extend beyond the length of a normal college semester. One exception to this is the three-year longitudinal study conducted by Fitzgerald, White, and Gruppen (2003). With a sample of 500

medical students, Fitzgerald et al. studied CJOL to gain a better understanding of the stability of CJOL over time and lay the groundwork for studying how to improve academic self-awareness. During the first two years of the study, students provided judgments of learning (JOL) following the completion of cognitive tasks, namely multiple-choice quizzes, labs, and examinations. Students then provided JOL on objective-structured clinical exams during the third year of the study. Students were asked to report JOL by estimating what percentage correct (0-100%) they would receive on the performance task.

Fitzgerald et al. (2003) found CJOL accuracy to be relatively stable across time. This stability is favorable to efforts to study the improvement in the accuracy of students' CJOL because the stability allows for researchers to establish baseline measures of CJOL. Researchers can then empirically test the effects of various interventions for improvement against this baseline measure. The work of Fitzgerald et al. (2003) should serve as a framework for future longitudinal research to establish baseline measures of CJOL within the broader populations of undergraduate and secondary education students. More robust research into baseline measures may, in turn, help to establish a clearer understanding of efforts to improve academic self-awareness.

Research from the perspective of self-regulation and CJOL suggests that academic self-awareness is significantly related to academic outcomes. Self-regulation perspectives emphasize the importance of being academically self-aware in order to self-regulate the learning process to produce desired academic outcomes. CJOL perspectives emphasize the importance of having an accurate awareness in order to have high performance on academic outcomes. Academic self-awareness can also contribute to achieving affective and social goals of many institutions of higher education such as lifelong learning (Cohen, 2012; Luftenegger et al., 2012). Although

findings from the fields of self-regulation and CJOL are promising, not a single article falling within the scope of this literature review specifically addressed the relationship between academic self-awareness and students' academic outcomes in developmental university courses. Moreover, this literature review found that a disproportionate number of research studies on academic self-awareness and academic outcomes were conducted in non-authentic laboratory settings corroborating researchers' findings (Carpenter et al., 2015; Hacker, Bol, & Bahbahani, 2008; Miller & Geraci, 2011a).

Calibrated Judgments of Learning in Applied Educational Contexts

Calibrated judgments of learning (CJOL) have a wide range of possible uses in applied educational contexts. The literature was reviewed to better understand how CJOL have previously been used in applied educational contexts to support underprepared university students' academic self-awareness and academic outcomes. Two prominent areas in which CJOL may be used in an applied context include, (a) the use of CJOL results as a guide for the allocation and use of study time, and (b) coupling CJOL with incentives to help improve academic self-awareness.

Using CJOL results to guide the allocation and use of study time. The appropriate allocation and use of study time has been shown to influence academic outcomes (Credé & Kuncel, 2008). Researchers have found that students' use of study time is greatly influenced by their academic self-awareness (Metcalfe & Finn, 2008a), though this is not the only contributing factor (Son & Metcalfe, 2000). Consequently, students with more accurate CJOL tend to be more capable of appropriately allocating and using their study time, which can result in more favorable academic outcomes. Findings from several studies indicate that the use of CJOL in

applied contexts may support underprepared students' academic outcomes through the appropriate use of study time.

In a study of 66 university students, Thiede, Anderson, and Therriault (2003) studied CJOL and study time in the context of text comprehension. Students were asked to read texts, rate their comprehension, and respond to a comprehension assessment. Students rated their comprehension by responding to the prompt, "How well do you think you understood the passage whose title is listed above? 1 (very poorly) to 7 (very well)" (p. 68). Following the comprehension assessment, students were given their overall score on the assessment without scores and feedback for individual questions. After reviewing their overall score, students were given the opportunity to select texts for restudy. This procedure sought to simulate a situation where students had the opportunity to appropriately allocate and use study time. Following restudy, students completed a second comprehension assessment.

Thiede et al. (2003) found that students with more accurate CJOL selected poorly comprehended text for restudy. Spending additional time studying unlearned material resulted in an increase in overall performance on the second comprehension assessment. In contrast, students with inaccurate CJOL were not as effective in the use of their study time. Their efforts to restudy text resulted in insignificant gains in performance. These findings support similar research that indicates that the accuracy of CJOL influences students' appropriate allocation and use of study time (Metcalfe & Finn, 2008b; Metcalfe, 2009; Nelson, Dunlosky, Graf, & Narens, 1994; Thiede & Dunlosky, 1999).

In studies addressing the way students interpret the use study time, Koriat, Nussinson, and Ackerman (2014) investigated the relationship between JOL, study time, and two predominant student interpretations of study time. These interpretations were data-driven and

goal-driven interpretations. The data-driven interpretation based the allocation and use of study time on the amount of effort required to accomplish a learning task. The goal-driven interpretation based the allocation and use of study time on the amount of effort voluntarily given to accomplish the students' learning goals (e.g., studying for a set amount of time). Koriat et al. recruited 42 and 56 undergraduate students to participate in their first and second studies respectively. Study participants were assigned to conditions eliciting either data-driven or goal-driven study time interpretations. Koriat et al. (2014) found that students' performance and perceptions of their learning were associated with their interpretation of study time. When students held data-driven interpretations of study time, perceptions of their learning were inversely proportional to the amount of time they spent studying. That is, students who spent less time studying an item perceived that the item was well learned; while extended time studying an item was perceived as being not well learned. Son and Metcalfe (2000) attributed this pattern to students' perceptions of item difficulty. In other words, less difficult items were perceived to require less time to complete and more difficult items were perceived to require more time for completion. In contrast, Koriat et al. (2014) found that when students held goal-driven interpretations of study time, their perceptions of learning were directly proportional to the amount of time studying an item. That is, students who spent less time studying an item perceived that the item was not well learned; while extended time studying an item was perceived as being well learned.

Both data-driven and goal-driven interpretations of study time can, under some circumstances, be detrimental to student learning. Using the duration of study time or arbitrary goals as indicators of perceived learning can lead to premature termination or an unnecessary extension of study time. This can occur when students' perceptions of performance do not

coincide with actual their performance (Mihalca et al., 2015). However, when trained, students can override the common data-driven and goal-driven interpretations of study time (Ariel, Dunlosky, & Bailey, 2009). Rather than having students default to judge their learning based on the amount of study time or arbitrary goals, students should be trained to allocate and use their study time based on their current level of understanding. CJOL provide an alternative method for students to objectively assess their perceptions of their understanding and appropriately allocate and use their study time (Koriat et al., 2014; Metcalfe, 2009; Mihalca, Mengelkamp, Schnotz, & Paas, 2015).

Considering previous research on the relationship between CJOL and the allocation and use of study time, underprepared students in developmental university courses may improve their study habits and consequently academic performance by improving the accuracy of their academic self-awareness. However, this review of the literature did not identify any previous research that specifically addressed the use of an intervention aimed to improve the academic self-awareness of underprepared college and university students.

Coupling CJOL with incentives to help improve academic self-awareness. Many researchers have studied the coupling of CJOL with incentives to improve academic self-awareness (Epley & Gilovich, 2005; Lin & Zabrucky, 1998; Miller, Duffy, & Zane, 1993; Miller & Geraci, 2011a; Nietfeld & Schraw, 2002; Schraw & Others, 1993). In these studies, CJOL played a crucial role by providing an objective measure of academic self-awareness. Emphasizing the importance of using educational measurement to improve academic outcomes, Resnick and Resnick (1992) articulated three premises associated with performance assessment. These premises were, (1) “what you test is what you get”, (2) “you do not get what you do not assess,” and (3) “make assessment worth teaching to” (p. 59). Resnick and Resnick’s premises

stress the idea that academic self-awareness should be measured if it is a desired student outcome. Integrating CJOL in applied educational settings allows practitioners to incentivize the improvement of academic self-awareness. Incentivizing academic self-awareness can take on many forms, such as: integrating CJOL as part of assignment, quiz, and test questions; assigning points based off CJOL accuracy; awarding extra-credit for CJOL accuracy; and requiring the remediation of CJOL inaccuracies. This section will highlight relevant research on the use of CJOL to incentivize academic self-awareness and suggest considerations for future research in applied settings.

In a 15-week introductory educational psychology course, Hacker, Bol, and Bahbahani (2008) studied the effects of incorporating CJOL as an incentivized measure of academic self-awareness. The course consisted of 137 students enrolled in teacher education programs. JOL were collected prior to each exam (predictions) and immediately after the completion of each exam (postdictions). Calibrations were calculated as the absolute difference between the students' JOL and actual exam performance. Two fully-crossed quasi-experimental research conditions were used to study the effect of incentivizing academic self-awareness. In the first research condition, students' incentive was a requirement to provide reflections describing the reasons for any discrepancies between their JOL and exam performance. In the second research condition, students' incentive was extra credit on the exam based on the overall accuracy of their CJOL. Students were randomly assigned to one of the four possible combinations of these two research conditions namely, (1) reflections only, (2) extra credit only, (3) reflections with extra credit, or (4) no reflection and no extra credit.

In this study, Hacker et al. (2008) found that there was a significant difference between the CJOL for high and low performing students, corroborating many other findings (Bol &

Hacker, 2001; DiFrancesca, Nietfeld, & Cao, 2016; Dunning, Heath, & Suls, 2004; Hacker et al., 2000; Nietfeld et al., 2005; Shake & Shulley, 2014; Valdez, 2013). High performing students had significantly more accurate CJOL than lower performing students. Under all experimental conditions, high performing students were consistently about 94% accurate across all three course exams. This already high level of accuracy may account for lack of improvement throughout the course for these students. In contrast, lower performing students assigned to the extra-credit incentive condition improved in both performance and accuracy of their postdiction CJOL across all three exams. These findings suggest that using CJOL measures as a basis for some forms of incentives can help low performing students become more academically self-aware and positively influence academic outcomes.

Methodological Considerations for Calibrated Judgments of Learning

Several methodological considerations should be taken into account when using calibrated judgments of learning (CJOL) in an applied educational context. CJOL results represent the degree to which students' perceptions of their performance on an academic task mirror their actual performance. The derivative nature of CJOL results dictates that the quality of CJOL measures cannot exceed the quality of the individual JOL and performance measures. As the quality of the JOL and performance measures increase, the quality of inferences drawn from CJOL results will also increase. Therefore, care should be taken when developing JOL and performance measures. In addition, the methods used to compare measures of JOL and performance (calibration) determine the type of inferences that can be drawn from CJOL results. This section will provide a brief overview of methodological considerations for the individual components of CJOL namely, judgments of learning (JOL), performance measures, and calibration calculations.

Considerations for JOL. Judgments of learning have been defined in this review as any judgment made by a student regarding the degree to which they believe they have accomplished a specified learning task. Two main methodological decisions must be made in order to measure JOL. First, the appropriate level of detail or *granularity* of the judgment must be determined. Second, an appropriate measurement scale must be selected.

There are two main levels of detail that are often used in the measurement of JOL. Global JOL refer to judgments of the overall outcome of the performance such as the final score on an exam. Local JOL refer to judgments of the individual components that make up the entire performance measure such as exam questions. Researchers have found that students are generally more capable of providing more accurate global JOL (Händel & Fritzsche, 2016; Nietfeld et al., 2005; Schraw, 1994). With the increased accuracy of global JOL comes a decrease in the amount of information provided by the measure. Global JOL do not provide information about how academically self-aware students are on specific tasks. For example, a student may feel that their performance on an exam was average and could be accurate in this perception of their performance. However, the student might not be as accurate in identifying which portions of the exam they did well on and which ones they did not. Local JOL provide additional information regarding the ability of the student to accurately monitor their learning on specific tasks. The added specificity of local JOL would most often be of interest when trying to improve self-regulatory practices or remediate misconceptions (Vössing & Stamov-Roßnagel, 2016).

JOL measurement scales consist of the JOL prompt and the response scale. In a study investigating the effects of differing JOL prompts on JOL accuracy, Pilegard and Mayer (2015) tested a variety of prompts from metacomprehension literature. These prompts fell into

categories of "how much, how confident, how many, or how difficult" (p. 68). The 127 study participants were assigned to 1 of 4 conditions associated with each of the prompt categories. Pilegard and Mayer found that there was no significant difference between the different prompt categories and the accuracy of students' JOL.

Although this literature review did not find any CJOL research specifically addressing JOL response scales, much research has been done in the area of educational and psychological testing. DeVellis (2012) suggests that the number of response options be enough to provide variation and represent "the respondents' ability to discriminate meaningfully" (p. 90). That is, response scales must have enough options for students to be able to adequately represent their JOL but not so many that the student cannot perceive the difference between options. For example, a response scale with two options (i.e., *Learned* and *Not Learned*) may not provide enough options for students to fully represent the degree to which they feel they learned the material. On the other hand, a response scale ranging from 0 to 100 may provide too many options for the student to be able discern differences in response options (i.e., the student might not be able to discern the difference between a JOL of 99 and 100) (McKelvie, 1978; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991).

Considerations for performance measures. The quality of the performance measure is a critical component of the CJOL because the performance measure serves as the standard to which students JOL are compared. The comparison between the JOL and performance measures will be flawed if the performance measure is not accurately measuring performance. A full treatment of the development of performance measures is beyond the scope of this paper. A minimum of various descriptive statistics and estimates of reliability should be used to evaluate

the quality of performance measures (DiFrancesca et al., 2016; Stankov & Crawford, 1996; Valdez, 2013).

Considerations for calculating calibrations. Schraw (2009) conducted a review of various methods for calculating the calibration of JOL and performance measures. In his analysis Schraw identified and discussed five main methods of calibration namely, absolute accuracy, relative accuracy, bias, scatter, and discrimination. One method of calibration is not necessarily superior to another, although, each method of calibration provides a distinct form of information. Table 2 summarizes each calibration method and provides a summary of how calibration results should be interpreted.

Table 2

Five Types of CJOL Calibration Measures

Type of measure	Outcome measure	Score interpretation
Absolute accuracy	Absolute accuracy index	Discrepancy between a confidence judgment and performance. *Measures judgment precision.*
Relative accuracy	Correlation coefficient	Relationship between a set of confidence judgments and performance scores. *Measures correspondence between JOL and performance.*
Bias	Bias index	The degree of over or under confidence in judgments. *Measures direction of judgment error.*
Scatter	Scatter index	The degree to which an individual's judgments for correct and incorrect responses differs in terms of variability. *Measures differences in variability for confidence judgments for correct and incorrect items.*
Discrimination	Discrimination index	Ability to discriminate between correct and incorrect outcomes. *Measures discrimination between confidence for correct and incorrect items.*

Note. Adapted from *A Conceptual Analysis of Five Measures of Metacognitive Monitoring* by G. Schraw, 2009, *Metacognition and Learning, 4*(1), 35.

Schraw (2009) provided two main recommendations for the use of these calibration methods. First, when possible, use multiple calibration measures to gain a more complete understanding of the calibration between JOL and performance measures. Second, select calibration measures that most appropriately address the purpose for which CJOL are being used. See Schraw's (2009) review for a full mathematical treatment of each calibration measure.

Summary of Literature Review Findings

Over the past three decades, a substantial amount of research has been conducted using calibrated judgments of learning (CJOL). Although the majority of this research has largely taken place in domain-specific areas of psychology, a growing number of researchers have found evidence to support the use of CJOL in applied educational settings. The significant research linking academic self-awareness (measured by CJOL) to positive academic outcomes suggest that a possible improvement in the accuracy of students CJOL could also improve their academic performance. The linking of accurate academic self-awareness to the allocation and use of study time suggests at least one way in which increasing the academic self-awareness of students could benefit their overall academic performance. The use of CJOL results in incentivizing academic self-awareness provides another example of how information about students' academic self-awareness could be used to improve their educational performance, especially low performers.

Much work remains in researching implications of context-specific decisions that must be made in order to fully utilize CJOL as a beneficial instructional practice. More specifically, much work remains to fill the gap of literature addressing the academic self-awareness of underprepared postsecondary students enrolled in developmental courses. Baseline studies are needed in order to better understand the academic self-awareness of this demographic of student. Research is needed to apply established CJOL measurement methodologies to this specific group of students to better understand the degree to which they are academically self-aware. A major premise for being concerned with, and measuring, academic self-awareness is that an improvement in academic self-awareness could lead to an improvement of academic outcomes. To support this premise, research specifically addressing the improvement of academic self-awareness over time for underprepared students is needed.

Method

The following sections will address each of the key methodological considerations which were taken into account while conducting this study, namely participant recruitment and selection, research instrumentation, research design and procedures, and data analysis methods.

Participants

Participants for this study were recruited from the Utah Valley University Developmental Math Department. Utah Valley University (UVU) seeks to provide learning opportunities for students with a wide variety of academic and career goals through programs at the certificate, associate, baccalaureate, and graduate levels. For the past several years UVU has been one of the largest public institutions of higher education in Utah with an enrollment of approximately 32,000 students. UVU has an open admissions policy, which allows any student the opportunity to attend. However, the application process requires students to participate in placement exams (e.g., ACT, SAT, Accuplacer) in order to assist in the proper placement in math and English. Students who score low on these exams are required to take developmental courses at the university (see Table 3 for specific course cut scores).

Participants for this study included students who did not meet the minimal requirements to enroll in college algebra at the time of their application to UVU. This population was chosen based on research indicating lower performing students are less academically self-aware (Bol & Hacker, 2001; DiFrancesca et al., 2016; Dunning et al., 2004; Hacker et al., 2000; Nietfeld et al., 2005; Shake & Shulley, 2014; Valdez, 2013). With typically lower levels of academic self-awareness, there is presumed to be more potential for gains in academic self-awareness over time. In addition, this population represents a diverse student population in terms of age, year in school, and chosen field of study; while maintaining the consistent subject domain of math. The

large differences of academic rigor between the lower level developmental courses and the upper level developmental courses also provide opportunities for attempting to determine whether or not more remedial coursework is indicative of lower academic self-awareness.

Each developmental math instructor was given the opportunity to facilitate this research study in their course and was instructed on the research objectives and methodology. Nine faculty members volunteered to participate. Students with participating instructors were introduced to the research through a brief presentation detailing the benefits, risks, and procedures of the study according to the Institutional Review Board approved protocols obtained for this study. A total of 285 students participated in the study. Study participants represented students from five developmental math courses namely, Math Fundamentals, Foundations for Algebra, Introductory Algebra, Integrated Beginning and Intermediate Algebra, and Intermediate Algebra. Study participants included 185 men and 100 women with ages ranging from 18 to 61 years (all participants: $M = 23.48$, $SD = 5.95$, men: $M = 23.46$, $SD = 5.07$, women: $M = 23.52$, $SD = 7.33$). Age was non-normally distributed, with skewness of 3.42 and kurtosis of 17.27. A participant's *year in school* is determined by the number of credit hours the student has completed (Freshman < 30, Sophomore >=30 & < 60, Junior >=60 & < 90, Senior >=90). Study participants included 192 Freshmen, 62 Sophomores, 22 Juniors, 9 Seniors. Table 4 details course descriptions and student demographics for each of the individual math courses.

Table 3

Summary of Course Placement Cut Scores

Course	Accuplacer	SAT	ACT
Math Fundamentals	*AA = 20-38*	*< 400*	*<16*
Foundations for Algebra	*AA = 39-65* *AL = 25-39*	*410-460*	16
Introductory Algebra	*AA ≥ 90* *AL = 46-60*	*470-490*	*17-18*
Integrated Beginning & Intermediate Algebra	*AA = 66-89* *AL = 40-45*		
Intermediate Algebra	*AL ≥ 61* *CL = 30-59*	*≥ 500*	*≥ 19*

Note: AA = Accuplacer Arithmetic, AL = Accuplacer Elementary Algebra, CL = Accuplacer College Level Math

Table 4

Summary of Student Demographics For Those Participating In This Study

Course Course Description	Age	Gender	Year In School
Math Fundamentals Designed for students requiring basic math review. Reviews basic operations with whole numbers and fractions. Topics of study include basic operations involving decimals, percents, ratios, rates, and basic operations involving physical measurements.	*M* = 25.33 *SD* = 5.68	Men = 7 Women = 5	Fr = 9 So = 3
Foundations for Algebra Designed for students requiring basic math and pre algebra instruction. Covers basic operations for number systems up to and including real numbers. Includes fractions, ratios, proportions, decimals, exponents, roots, linear equations, and polynomial expressions.	*M* = 27.11 *SD* = 8.92	Men = 8 Women = 1	Fr = 7 Jr = 1 Sr = 1
Introductory Algebra For students who have completed a minimum of one year of high school algebra or who lack a thorough understanding of basic algebra principles. Teaches integers, solving equations, polynomial operations, factoring polynomials, systems of equations and graphs, rational expressions, roots, radicals, complex numbers, quadratic equations and the quadratic formula. Prepares students for MAT 1010, Intermediate Algebra	*M* = 24 *SD* = 7.95	Men = 21 Women = 18	Fr = 25 So = 8 Jr = 3 Sr = 3
Integrated Beginning & Intermediate Algebra Teaches Beginning and Intermediate Algebra in one semester. Includes linear, quadratic, and rational expressions, equations, and functions; systems of equations; logarithms; exponents; graphing; and problem solving.	*M* = 28.05 *SD* = 11.53	Men = 13 Women = 6	Fr = 11 So = 6 Jr = 2
Intermediate Algebra Expands and covers in more depth basic algebra concepts introduced in Beginning Algebra. Topics of study include linear and quadratic equations and inequalities, polynomials and rational expressions, radical and exponential expressions and equations, complex numbers, systems of linear and nonlinear equations, functions, conic sections, and real world applications of algebra.	*M* = 22.69 *SD* = 4.17	Men = 136 Women = 70	Fr = 140 So = 45 Jr = 16 Sr = 5

Instrumentation

Two types of data must be collected in order to measure academic self-awareness through the use of CJOL. The first type of data is students' JOL for each question on the math exam. The second type of data is students' performance on each of the questions on

the math exam. The instrumentation used to collect each of these two types of data will be described in the following sections.

Judgment of learning (JOL) instrumentation. JOL data was collected using a self-report Likert type scale. Each math question had a corresponding JOL prompt. Students responded to a JOL prompt after answering a math question and before moving on to subsequent math questions. The JOL prompt was: *How confident are you that you answered the test question correctly?* Students responded to the prompt on a five point scale representing confidence bands of 20 percent (i.e., 0-20, 20-40, 40-60, 60-80, and 80-100).

Students were informed that the JOL or its accuracy would not influence the student's grade. Students were instructed to spend no more than 5 seconds responding to each JOL prompt. Students were also instructed to respond to the JOL prompt immediately after answering a math question before moving on to the next math question. For example, if the math exam had two questions, the student would have been instructed to proceed in the following manner: answer math question 1, respond to JOL prompt 1, answer math question 2, and respond to JOL prompt 2. Figure 1 illustrates how the math question and JOL prompt appeared on math exams integrating the JOL prompt directly into the math exam. The Confidence Calibration Assessment (CCA) was used when the JOL prompt was not directly integrated into the math exam. A sample of the full CCA can be found in Appendix A.

1. Simplify the expression $(2^{x^2}-5x-12)/(2^{x^2}-4x-16)$.

a. (x-6)/2(x-2)
b. (x-6)/2(x+2)
c. (2x+3)/2(x-2)
d. (2x+3)/2(x+2)

Math Test Question #	How confident are you that you answered the test question correctly?				
	0% -20%	20% to 40%	40% to 60%	60% to 80%	80% to 100%
1					

Figure 1. JOL prompt example for a sample problem. After completing a question like the one above, students were asked to indicate how confident they were that they got the questions correct.

Math exams. The math exams were used to measure the degree to which students understand mathematical concepts. These exams were the math exams each faculty member developed for their developmental math course independent of this research study. Each exam was unique to the individual developmental math course and instructor. Some instructors personally wrote each question on the exam while others selected relevant questions from a published test bank. The types of questions that were developed or selected include multiple choice, fill in the blank, free response, and matching. The number of questions on each exam varied from instructor to instructor and from test to test; no exam exceeded 30 items. Exams questions were both dichotomously scored (i.e., right or wrong) and polytomously scored (i.e., partial credit awarded based on the correctness of the student response). The majority of exam questions were dichotomously scored.

Although some specific computed psychometric properties of these exams were unknown (such as item reliability, difficulty, and discrimination), the exams possessed several evidences of validity. Exam validity can be defined as the "degree to which

evidence and theory support the interpretation of test scores entailed by the proposed uses" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). In other words, validity is the connection between the interpretation of test results and the nature of the phenomenon it attempts to measure. The evidences for validity that are commonly present in these exams are:

1. *Content Evidences of Validity*: The match between the content of the actual exam and what should be in the exam according to experts
2. *Evidences of Face Validity*: The degree to which the exam appears to be related to what is being measured according to non-experts
3. *Evidences of Association with Other Variables*: The relationship between exam results and results from other exams

It is important to recognize a few of the assumptions that underline the data that were generated from the math exams. One of these assumptions is that there is a degree to which student responses are either correct or incorrect. This degree of correctness represents the degree to which students understand the underlying math concept. Another assumption is that the instructors are experts in their field and that they are capable of adequately judging the degree to which students understand math concepts and are capable of assigning numeric values (points), which represent the degree of understanding.

Research Design and Procedures

This study utilized a within-subjects repeated measures design. This research design was chosen in order to determine to what degree students' CJOL accuracy changes over time. The study was carried out in six main steps: (a) recruiting research participants, (b) collecting JOL

data from students, (c) collecting performance data from faculty, (d) calculating calibrations between students' JOL and performance data, (e) analyzing CJOL results and providing feedback to students and faculty, and (f) analyzing data across time in terms of demographic variables. Steps two through five were repeated each time the faculty member administers an exam during the semester.

Step 1. The first activity in this research project involved the recruitment of research participants, which began with speaking to faculty members about the research project. After consenting to facilitate the research in their class, faculty members received training concerning their role in the facilitation of the data collection process. This training consisted of proper procedures for recruiting participants, distributing and collecting research documents, and providing necessary data on student exam performance. After being trained, faculty members distributed an informed consent form to each of their students (Appendix B). The form was read and discussed in class. Prior to consenting to be part of the study, faculty addressed any questions regarding their role in the research study. Students were referred to the principal investigator for specific questions regarding the study that could not be answered by the faculty member. Students were referred to the UVU Institutional Review Board (IRB) for questions regarding their rights as research participants. Students wishing to participate in the study returned their signed consent forms to the faculty member after which the faculty member returned the consent form to the researcher.

Step 2. The second step in this research project consisted of collecting JOL data from the research participants. This stage of data collection occurred simultaneously with each math exam during the semester. Faculty members had two choices for facilitating the data collection process. Faculty members chose to either embed the JOL directly into their math exam making

JOL responses part of the exam itself, or utilize the Confidence Calibration Assessment (CCA) (Appendix A) allowing the students to report JOL responses on a separate sheet of paper. On the appropriate day, the faculty members distribute the course math exam along and CCA when appropriate. Students who chose not to participate in the study skipped the embedded JOL prompt or did not complete the CCA. Students returned JOL information to the instructor following the completion of the math exam and JOL prompts.

Step 3. The third step of the research project consisted of collecting students' exam performance data from the faculty member. This stage of data collection occurred after each exam was administered during the semester. The student's individual scores on exam questions represented the degree to which a student understands the concept being tested. After grading the exam, faculty members documented the exam scores and JOL responses using the Math Exam Score Reporting Template (Appendix C). The completed template was then emailed to the principal investigator for analysis.

Steps 4 and 5. The fourth and fifth steps of the research project consisted of analyzing JOL and math exam results and providing direct feedback to the student and instructor regarding the accuracy of students' academic self-awareness. Feedback was emailed to faculty and students university email accounts (Appendix D: Student Feedback, Appendix E: Teacher Feedback).

Step 6. The sixth and final step of the research project consisted of analyzing data across time in terms of demographic variables. The collection of participant age, gender, year in school, and course level were retrieved from the secure university academic servers. Demographic data was then combined with students CJOL results to create a final research dataset. Personally identifiable information was then removed from the dataset and be replaced

with a unique participant ID. A data crosswalk was created in the event that questions arise in the analysis phase regarding the original data. The crosswalk was stored in a secure location on university computers. The final analysis of the research data was carried out as described in the Data Analysis section.

Data Analysis

The data analysis process was divided into two main phases. The first phase consisted of the calculation of calibrated judgment of learning (CJOL) and feedback. The second phase consisted of the analysis of aggregated CJOL results. Data analysis methods for each phase will be discussed in detail in the remainder of this section.

CJOL calculations. Following the recommendations of Schraw (2009), multiple CJOL calculations were used in order to gain a greater understanding of the accuracy of students' academic self-awareness. The first step in calculating CJOL measures was to standardize both the scores for each math question and JOL responses. This standardization was necessary due to the variation in scoring methods on math exams from faculty member to faculty member. In addition, scaling was necessary to bring the exam scores and the JOL responses on the same scale in order to appropriately compare the two scores. The scores for each math question were standardized by taking four times the proportion of points earned to the points possible ($4 \times \frac{Points\ Earned}{Points\ Possible}$). This resulted in scaled exam question scores with values ranging from 0 to 4. The JOL responses were scaled by converting the five-point response scale into values ranging from 0 to 1. JOL responses were scaled using the following conversions: (0%-20%) became 0, (20%-40%) became 1, (40%-60%) became 2, (60%-80%) became 3, and (80%-100%) became 4.

The second step in calculating CJOL was to compare exam performance and JOL responses. The absolute accuracy index and the bias index (Schraw, 2009) were used to

calculate CJOL following the standardization of the math and JOL scores. The absolute accuracy index was used to measure the discrepancy between the JOL and exam performance and represented judgment precision. The absolute accuracy index was calculated for each math question by taking the absolute value of the difference between standardized math and JOL scores ($|JOL\ Score_i - Math\ Score_i|$). This analysis produced values ranging from 0 to 4 with values closer to 0 representing more accurate CJOL. Students' absolute CJOL accuracy scores were aggregated at the test level by calculating the mean of the absolute accuracy scores for all exam questions. The bias index was used to measure the degree of over or under confidence in JOL and represented the magnitude and direction of judgment error. The bias index was calculated for each math question by taking the difference between standardized math and JOL scores ($JOL\ Score_i - Math\ Score_i$). This analysis produced values ranging from -4 to 4. Negative scores corresponded to under confidence while positive scores corresponded to over confidence. Students' CJOL bias scores were aggregated at the test level by calculating the mean of the bias scores for all exam questions. It is important to note that bias scores aggregated at the test level lose their ability to represent the magnitude of the bias and only represent average over and under confidence.

Student feedback. Student feedback was generated from individual student absolute accuracy and bias CJOL results. A sample of the student feedback email can be found in Appendix D. There were five points of data included in the student feedback namely a, general academic self-awareness score, general accuracy indicator, general bias indicator, academic self-awareness graph, and question review suggestions.

General academic self-awareness feedback. The aggregated exam absolute accuracy score was transformed to produce the general academic self-awareness score for student

feedback (Figure 2). The transformation was used in order to aid the interpretability of the results for students by conforming to the common assessment practice of having a high score reflect favorable results. This transformation inverted the previously calculated CJOL score so that a high score represented a more accurate CJOL and a low score represented a less accurate CJOL. This inversion was calculated by subtracting 4 from the scaled absolute accuracy score (4 – absolute accuracy).

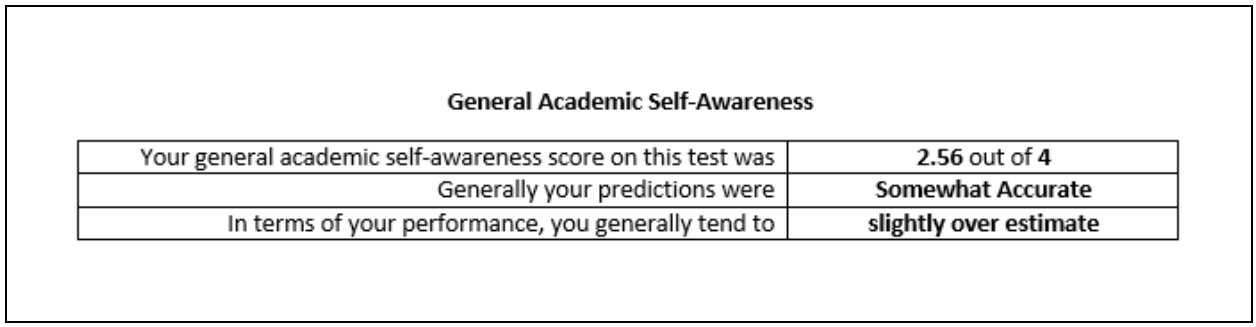


Figure 2. General academic self-awareness portion of student feedback email. This portion of the feedback gives students a quantitative representation of their self-awareness along with qualitative descriptions of their accuracy and CJOL bias.

The accuracy indicator was used to provide students with a general interpretation of the accuracy of their academic self-awareness. Categories of *accurate*, *somewhat accurate*, *somewhat inaccurate*, and *inaccurate* were used to describe the general accuracy of their academic self-awareness (Figure 2). The criteria for each of the four categories can be found in Table 5.

Table 5

Absolute Accuracy Category Criteria

General Accuracy Category	Criteria for Inclusion Absolute Accuracy Score
Accurate	Score <= 4 & Score > 3
Somewhat Accurate	Score <= 3 & Score > 2
Somewhat Inaccurate	Score <= 2 & Score > 1
Inaccurate	Score <= 1 & Score >= 0

The bias indicator was used to provide students with a general understanding of whether they tend to over or under estimate their performance. Categories of *under estimate*, *slightly under estimate*, *perfectly estimate*, *equally over and under estimate*, *slightly over estimate*, and *over estimate*. The criteria for each of these six categories can be found in Table 6. It is important to note that both the perfect estimate and equally over and under estimate categories required the inclusion of the absolute accuracy score to distinguish the difference between the two types of estimations.

Table 6

Bias Category Criteria

General Accuracy Category	Criteria for Inclusion Bias Score	Criteria for Inclusion Absolute Accuracy Score
Under Estimate	Score <= -4	
Slight Under Estimate	Score > -4 & Score < 0	
Perfectly Estimate	Bias Score = 0	Score = 0
Equally Over and Under Estimate	Bias Score = 0	Score = 0
Slight Over Estimate	Score < 4 & Score > 0	
Over Estimate	Score <= 4	

Academic self-awareness feedback graph. A graphical representation was used to provide feedback on how often and to what degree students over and under estimated their performance. The bias score for each individual exam question was used to classify CJOL results into seven categories. These categories were *large under estimate*, *under estimate*, *slight under estimate*, *accurate*, *slight over estimate*, *over estimate*, and *large over estimate*. The criteria for each of the seven categories can be found in Table 7. A bar chart was then created using frequency counts of the number of questions in each category. The bar chart was color coded according to the direction of corresponding bias scores. An example of the academic self-awareness feedback graph can be found in Figure 3.

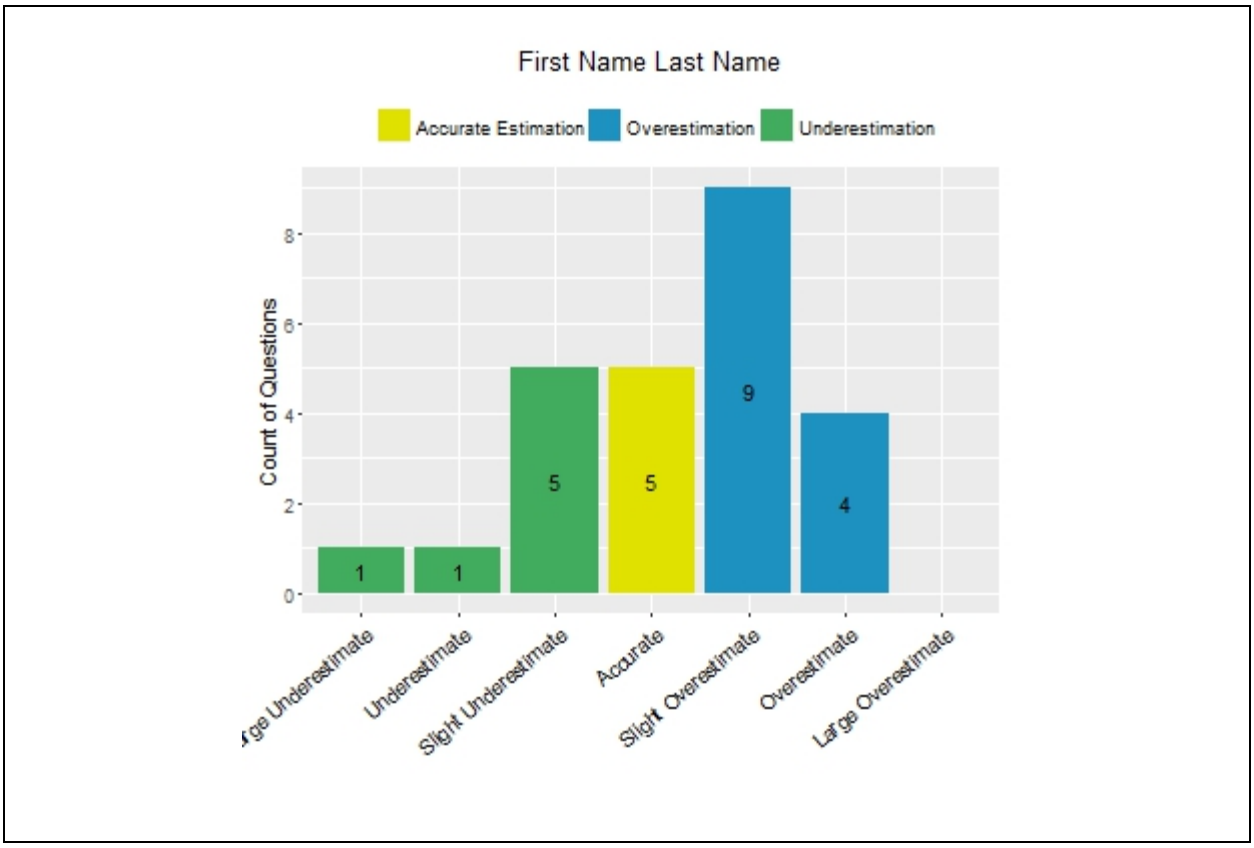


Figure 3. Academic self-awareness feedback graph. The academic self-awareness feedback graph represents the number of CJOL results in each accuracy category. The graph color corresponds to the larger categories of under, over, and accurate estimations.

Table 7

Academic Self-Awareness Graph Criteria

General Accuracy Category	Criteria for Inclusion Bias Score (range -4 to 4)
Large Under Estimate	Score < -3
Under Estimate	Score < -2 & Score >= -3
Slight Under Estimate	Score <= -1 & Score >= -2
Accurate	Score < 1 & Score > -1
Slight Over Estimate	Score >= 1 & Score <= 2
Over Estimate	Score > 2 & Score <= 3
Large Over Estimate	Score > 3

Question review feedback. The question review section of student feedback suggested exam questions for student review. Exam questions that fell into categories of *large over / under estimate* and *over / under estimate* were considered as potential questions to suggest for review.

These categories were selected because they represented the greatest opportunities for improved accuracy. From the list of potential questions for review, a maximum of five questions were randomly selected from over and under estimate categories and suggested to the student. A maximum of five questions were selected in an attempt to not overwhelm students who may have had a large number of inaccurate CJOL. Figure 4 provides an example of the question review feedback.

Greatly Under Estimated	**17, 19**
Greatly Over Estimated	**14, 7, 12, 13**

Figure 4. Question review feedback. The question review feedback suggests up to five questions for students to review from the of *large over / under estimate* and *over / under estimate* categories.

Teacher feedback. Teacher feedback was generated from both absolute accuracy and bias CJOL results for each exam question. A sample of the teacher feedback email can be found in Appendix E. There were five points of data included in the teacher feedback namely, analysis of question difficulty, graphical representation of question difficulty, analysis of question discrimination, analysis of accuracy, and a graphical representation of accuracy. Teachers were also provided with a document containing all of the student feedback reports for their particular course.

Item difficulty feedback. The item difficulty for teacher feedback was calculated as the mean student score for each item. For the purposes of graphically representing the item difficulty, the item difficulty score was transformed so that higher numbers represented more difficult items and lower numbers represented easier items (Figure 4). This transformation was done by calculating the difference of points possible and average student score for each exam question ($points\ possible_i - ave\ score_i$). The bar chart was color coded according to the

direction of the average corresponding bias scores. Teachers were also given a list of the top five most difficult items.

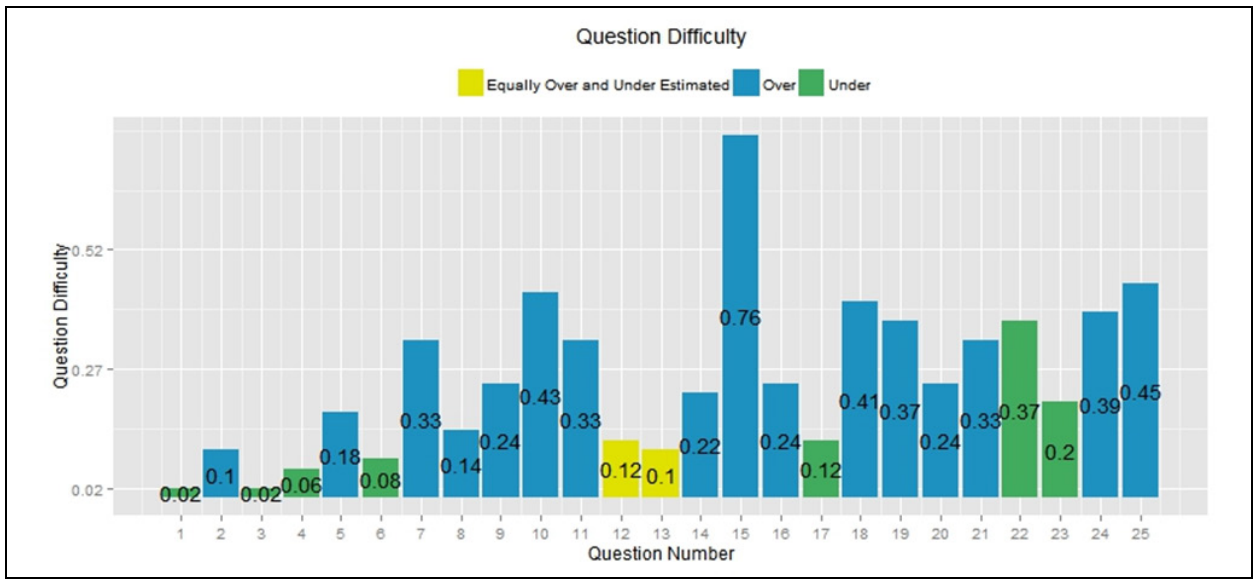


Figure 5. Question difficulty graph. The question difficulty graph represented the relative difficulty of each question. The numeric value represents the deviance of the average score from the possible score. Graph colors represent whether or not the question was on average over, under or equally over or under estimated.

Item discrimination. Item discrimination was calculated using an item to total correlation. The item to total correlation was calculated by using a Pearson product-moment correlation between the item score (e.g., 0 or 1) and the overall exam score (e.g., 87%). Item discrimination can be interpreted as the degree to which the item discriminates between high and low performing students (Miller et al., 2013). When using item to total correlations, high and low performance is evaluated on a continuous scale (0-100) corresponding to the percentage of the total points earned on the math exam. For the purpose of teacher feedback, item discrimination was only used to identify potentially poor items. A list of negatively discriminating items was given to teachers as a suggestion to review the item.

Item CJOL accuracy feedback. The accuracy of students' CJOL for each question on the exam was reported to the teacher though a bar graph (Figure 6). The item CJOL accuracy was

calculated by taking the average absolute accuracy score for each item. Lower values represented more accurate CJOL. Higher values represented less accurate CJOL (i.e., a greater discrepancy between the students' JOL and actual performance). The bar chart was color coded according to the direction of the average corresponding bias scores. A list of the five most accurate and least accurate questions was also provided.

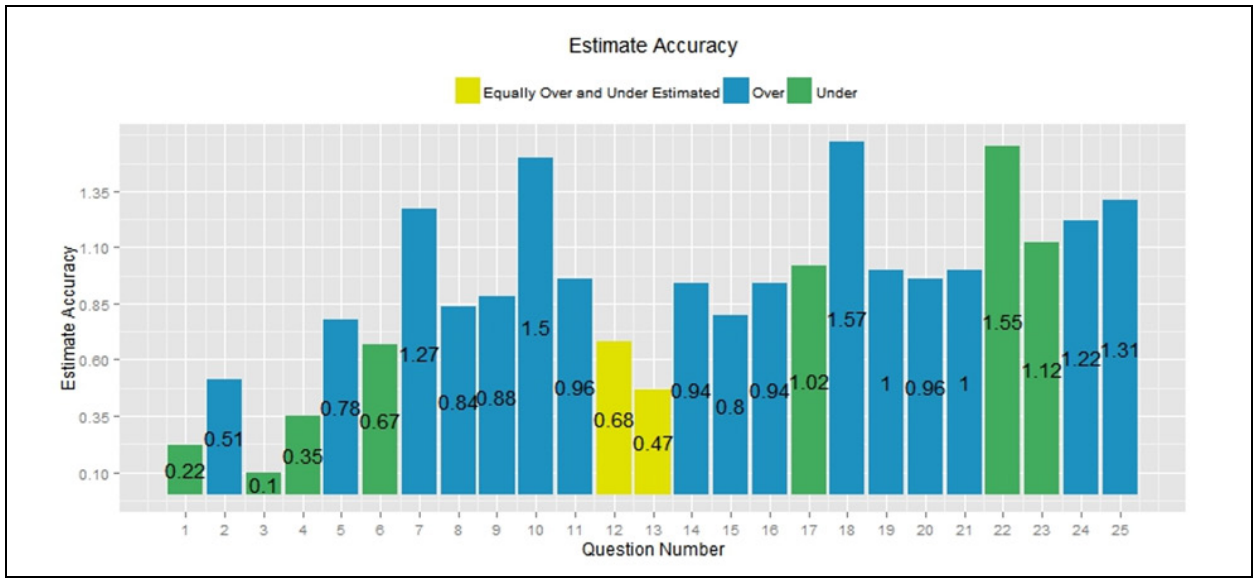


Figure 6. Estimation accuracy graph. The estimation accuracy graph represented the average absolute accuracy score for each question. The smaller the value, the more accurate student absolute accuracy scores were on that question. Graph colors represent whether or not the question was on average over, under or equally over or under estimated.

Analysis of aggregated CJOL results. Results from the CJOL calculations were combined with demographic variables into a final dataset in order to address three specific research questions. The final dataset included 285 students, 653 test questions, 36 tests, and 17,432 individual question responses. Students' performance on each exam was transformed into a z-score in order to compare test performance across tests, instructors, and courses. Item difficulty for each test question was calculated by taking the mean of scaled item scores. Test difficulty was calculated by taking the mean of scaled test scores. Students' test ability was classified into high, average, or low based on the test z-score (Table 8). Students' overall ability

was classified into high, average or low based on the mean z-score for each test (Table 8). Variables included in the final dataset included unique student identifiers, bias scores, absolute accuracy scores, test performance (transformed into a z-score), item difficulty, test difficulty, student ability classification, and overall ability classification.

Table 8

Student Test Ability Classification Criteria

Test Ability Classification	Classification Criteria
High	Test z-score > 1
Average	Test z-score <= 1 & Test z-score >= -1
Low	Test z-score < -1

The following sections will address the methods of analysis for each research question namely, (a) to what degree are developmental math students' perceptions of their performance match their actual performance accurate (CJOL accuracy), (b) to what degree do developmental math students' CJOL become more accurate over time after receiving feedback on their accuracy, and (c) to what degree do differences in CJOL accuracy exist amongst disaggregated groups of age, gender, year in school, and course level. Results were considered statistically significant, and the null hypotheses rejected, when p values were 0.05 or smaller.

Analysis of research question 1. Research question 1 was: to what degree do developmental math students' perceptions of their performance accurately match their actual performance (CJOL accuracy). Four separate points of analysis were undertaken to answer the broader research question regarding the degree to which students' CJOL are accurate. These analyses were, (a) overall absolute accuracy, (b) proportion of over, under, and perfect estimates, (c) absolute accuracy and bias with item and test difficulty, and (d) absolute accuracy and bias

for high, average, and low performing students. In all four of these analyses all 17,432 points of data were used. Any missing data was removed on a pairwise basis.

First, the analysis of absolute accuracy across all data points was calculated by taking the mean value for all absolute accuracy scores. The purpose of this analysis was to investigate the overall absolute accuracy of student responses. Second, the proportion of over, under and perfect estimates was calculated by dividing the number of estimates per category by the total number of estimates. The bias score was used to classify each CJOL into the over, under or perfect categories. Negative bias scores were classified as under estimates. Positive bias scores were classified as over estimates. Bias scores of zero were classified as perfect estimations. The purpose of this analysis was to determine whether or not students had a tendency to over, under, or perfectly estimate their performance.

The third point of analysis compared item and test difficulty to absolute accuracy and bias scores. The purpose of this analysis was to investigate whether or not the difficulty of the item or test influenced the accuracy of absolute accuracy and bias scores. To investigate the relationship between absolute accuracy and item difficulty, an aggregate absolute accuracy score was computed for each test item by taking the mean value of all absolute accuracy scores derived from that particular test item. Item difficulty was calculated as the mean standardized math score for each item. A Pearson product moment correlation was then calculated to correlate the mean absolute accuracy scores and the item difficulty. A linear regression analysis was calculated in order to determine the degree to which variation in students' absolute accuracy scores can be accounted for by item difficulty. The assumptions of linear regression were checked namely, linearity of data, independence of observations, normality of distribution, and the equality of

variances. Similar methods of analysis were used to analyze the relationship between absolute accuracy and test difficulty by aggregating data at the test level rather than the item level.

Percentages of over, under, and perfect estimations were used to investigate the relationship between bias scores and item difficulty. This was done by taking the percent over, under, and perfect estimations for each test item. These percentages were then correlated with item difficulty using the Pearson product moment correlation. A linear regression analysis was calculated in order to determine the degree to which variation in students' over, under, and perfect estimations could be accounted for by item difficulty. The assumptions of linear regression were also checked.

The fourth point of analysis compared the absolute accuracy and bias scores for high, average, and low ability students. The purpose of this analysis was to investigate whether or not students CJOL accuracy differed amongst ability levels. To carry out this analysis, absolute accuracy scores were aggregated for each student by taking the mean of all the student's absolute accuracy scores. Students' overall ability was then dummy coded and regressed on students' average absolute accuracy scores. To investigate how students' overall ability related to whether or not they over under or perfectly estimated, students' bias scores were aggregated at the student level by calculating the percent over, under and perfect estimations for all items to which the student responded. A linear regression analysis was calculated in order to determine the degree to which variation in students' over, under, and perfect estimations could be accounted for by the overall ability of the student.

Analysis of research question 2. Research question 2 is: to what degree do developmental math students' CJOL become more accurate over time after receiving feedback on their accuracy. To address this research question, students' absolute accuracy scores were

compared across their first three test occasions. The decision to only include data from the first three test occasions was made because it maximized both the number of students and test occasions. The nature of the data was examined to determine if it met the assumptions of a one-way repeated measure ANOVA which are: (a) the dependent variable is measured at the continuous level, (b) the independent variable consists of at least two categorically related groups, (c) there are no significant outliers, (d) the distribution of dependent variables are approximately normally distributed, and (e) there is sphericity in the data. For this analysis the dependent variable was the absolute accuracy score and the independent variable was time with three levels of the variable representing the three test occasions.

Analysis of research question 3. Research question 3 is – to what degree do differences in CJOL accuracy exist amongst disaggregated groups of age, gender, year in school, and course level. The nature of the data was examined to see if it meets the assumptions of a one-way ANOVA test. These assumptions are (a) the dependent variable is measured at the continuous level, (b) the independent variable consist of at least two categorically related groups, (c) the measures in each group represent independent observations (d) there are no significant outliers, (e) the distribution of the dependent variable is approximately normally distributed, (f) there is homogeneity of variances. For this analysis the dependent variable will always be the computed absolute accuracy score. The independent variables are the disaggregated groups of age, gender, year in school, and course level. Age values were aggregated into 5 groups shown in Table 9.

Table 9

Age Groupings

Group Number	Age Range
1	Age <= 20
2	Age > 20 & Age <= 25
3	Age > 25 & Age <= 30
4	Age > 30 & Age <=35
5	Age > 35

Results

Research Question 1

Research question 1 was: to what degree do developmental math students' perceptions of their performance accurately match their actual performance (CJOL accuracy). After accounting for missing data, the total number of observations with complete absolute accuracy and bias data was 17,091 observations. The mean absolute accuracy score across all observations was $M = 1$, $SD = 1.21$. The proportion of over, under and perfect estimations, as indicated by the bias score, were as follows: 19.1% of students' estimations were over estimations of performance, 34.0% of students' estimations were under estimations of performance, and 46.9% of students' estimations were perfect estimations of performance.

The average absolute accuracy scores were significantly related to the both item and test difficulty. The Pearson correlation coefficient calculating the relationship between the average absolute accuracy score per item and item difficulty resulted in a significant correlation (r (741) = -.60, $p < .001$). The correlation between the average absolute accuracy score per test and test difficulty also resulted in a significant correlation (r (39) = -.49, $p < .001$). A simple linear regression was calculated to predict students' absolute accuracy scores based on the difficulty of item and test difficulty. A significant regression equation was found when absolute accuracy was regressed on item difficulty (F (1,741) = 425.6, p < .001) with an R^2 of .36. Students

predicted absolute accuracy was equal to 2.18 – 1.55(item difficulty). The resulting equation was inspected for violation of assumptions (Figure 7). The linearity of the data was assessed by plotting the residual values vs. the fitted values. The *Residuals vs Fitted Plot* graphs the regression residuals vs. the fitted values with a lowess line and 95% confidence interval. The relatively straight lowess line suggests that the data meets the linearity assumption. The variance in residuals also suggests that the assumption of homoscedasticity has not been violated due to the lack of an extreme *fan* shaped distribution. The *Normal Q-Q Plot* further suggests that the assumption of homoscedasticity has not been violated due to closeness of the expected vs. actual plotted residuals and the diagonal straight line. A significant regression equation was also found when absolute accuracy was regressed on test difficulty (F (1,39) = 12.17, p < .01) with an R^2 of .24. Students' predicted absolute accuracy is equal to 2.01– 1.34 (test difficulty). The resulting equation met assumption criteria for linear regression (Figure 8).

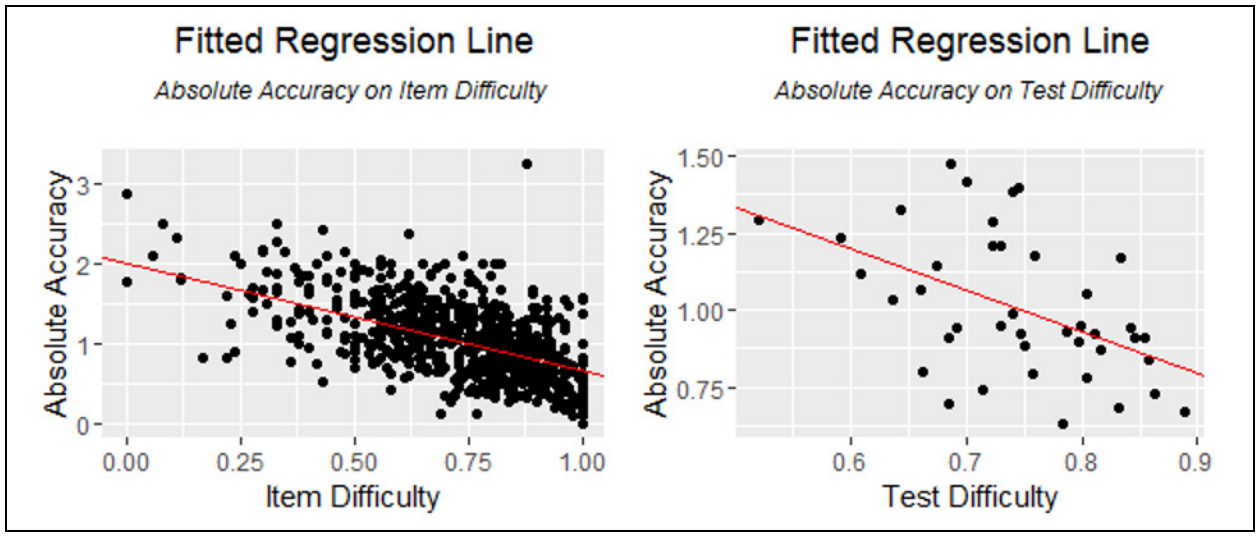


Figure 7. Fitted regression lines for absolute accuracy scores and difficulty. Graphs provide a visual representation of the relationship between absolute accuracy scores and difficulty by using scatter plots with regression lines. The graph on the left represents item difficulty and the graph on the right represents test difficulty.

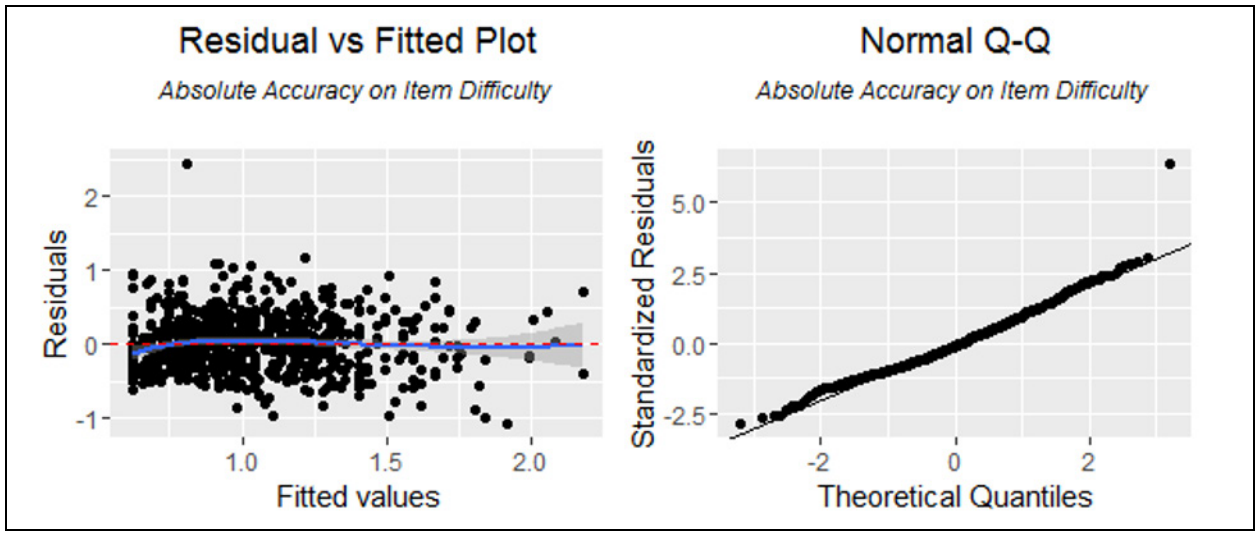


Figure 8. Plots for assessing the assumptions of the regression of absolute accuracy scores on item difficulty. Left: a residual vs fitted values plot was used to inspect linearity in the data. Right: a normal Q-Q plot was used to assess the homoscedasticity of the data by comparing the probability distributions of the data vs. a theoretical probability distribution.

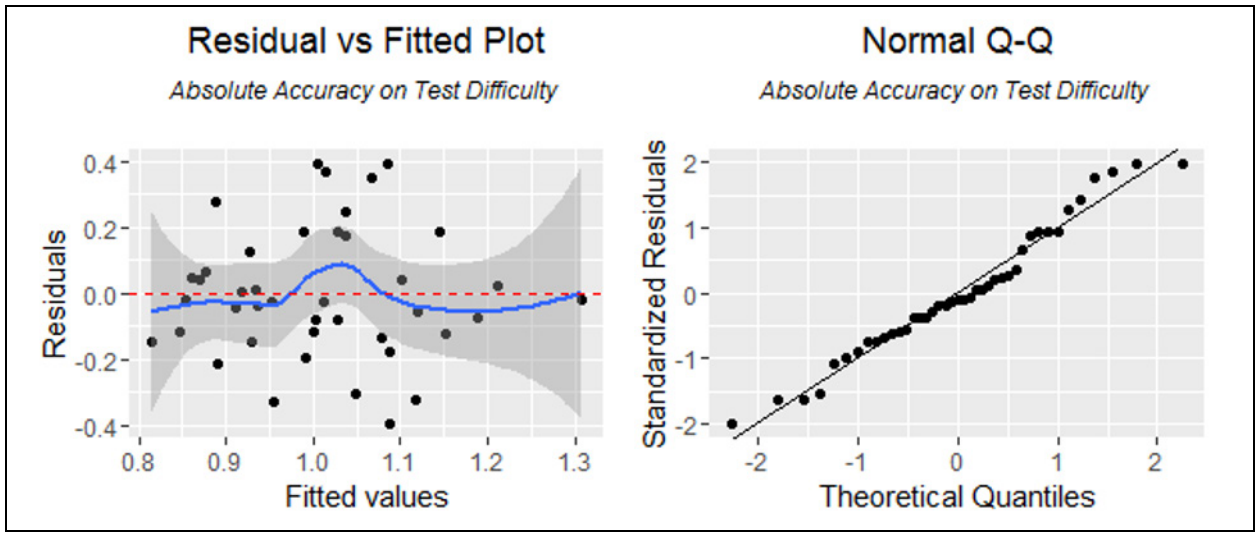


Figure 9. Plots for assessing the assumptions of the regression of absolute accuracy scores on test difficulty. Left: a residual vs. fitted values plot was used to inspect linearity in the data. Right: a normal Q-Q plot was used to assess the homoscedasticity of the data by comparing the probability distributions of the data vs. a theoretical probability distribution.

The relationship between percentages of over, under and perfect estimations was significantly related to item difficulty. The correlation between the percent of over estimations and item difficulty resulted in a strong correlation (r (741) = -.82, p < .001). The correlation between the percent of under estimations and item difficulty resulted in a weak yet statistically

significant correlation (r (741) = .13, $p < .001$). The correlation between the percent of perfect estimations and item difficulty resulted in a moderate correlation (r (741) = .59, $p < .001$). The item difficulty was also a strong predictor of whether or not a student would over, under, or perfectly estimate their performance. There was a statistically significant relationship between item difficulty and the percentage of over (F(1,741) = 1578, p < .001, R^2 = .68), under (F (1,741) = 11.81, p < .001, R^2 = .02), and perfect (F (1,741) = 400.4, p < .001, R^2 = .35) estimations. Of the three types of estimation, the item difficulty was the greatest predictor of the percentage of overestimation.

There was a statistically significant relationship between students' ability (overall exam performance) and their absolute accuracy scores. When students' average absolute accuracy was regressed on their ability (average test z-scores) a significant relationship was found (F (2,282) = 34.9, p < .001, R^2 = .20). The regression coefficients for high, average, and low ability were .50, .52 and .87 respectively. These results indicate that low performance was a better predictor of a students' accuracy than high or average performance. High, average, and low performing students differed overall in the percent of questions that were over, under and perfectly estimated. High ability students over estimated only 5.0% of the time, under estimated 25.4% of the time, and perfectly estimated 70.0% of the time. Average ability students over estimated only 19.0% of the time, under estimated 35.8% of the time, and perfectly estimated 45.2% of the time. Low ability students over estimated only 34.3% of the time, under estimated 33.4% of the time, and perfectly estimated 32.4% of the time. Students' ability was regressed on the percent over, under and perfect estimations of the student. The results indicate that student ability is a statistically significant predictor (p < .001) of whether or not the student will over, under, or

perfectly estimate their performance (Table 10). The difference in percent over, under, and perfect estimates amongst the differing ability groups is illustrated in Figure 10.

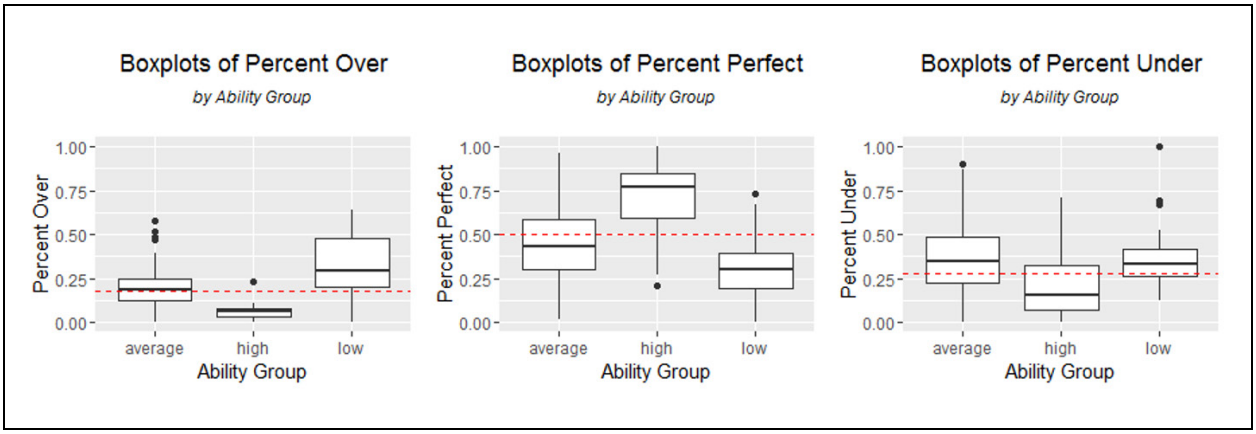


Figure 10. Boxplots of the percent over, under, and perfect estimations by ability group. The dotted red line represents the average median percent over, under, and perfect by each ability group.

Table 10

Student Ability Regressed On Percent Over, Under, And Perfect Estimations

Regression Model	Ability	Coefficients	Std. Error	*p*
Perfectly estimated on Overall ability	High	.69	.04	< .001
	Average	-.25	.04	< .001
	Low	-.39	.05	< .001
Over estimated on Overall ability	High	.07	.02	< .01
	Average	.13	.02	< .001
	Low	.27	.03	< .001
Under estimated on Overall ability	High	.24	.04	< .001
	Average	.12	.04	< .001
	Low	.12	.05	< .05

Significant relationships were found at all four of the analyses carried out to address research question 1 namely, (a) overall absolute accuracy, (b) proportion of over, under, and perfect estimates, (c) absolute accuracy and bias with item and test difficulty, and (d) absolute accuracy and bias for high, average, and low performing students.

Research Question 2

Research question 2 considered the degree to which developmental math students' CJOL become more accurate over time after receiving feedback on their accuracy. Only complete cases with student absolute accuracy scores on three test were selected for this analysis. After accounting for missing data, 204 complete student observations remained. The mean and standard deviation for absolute accuracy scores at each test occasion were: test 1 $M = .85$, $SD = .50$, test 2 $M = .86$ $SD = .47$, test 3 $M = 1.06$ $SD = .52$. A one-way repeated measures ANOVA was used to determine the significance of the increases of absolute accuracy scores across the three tests. The assumption of normality for a one-way repeated measures ANOVA was met. However, Mauchly's test of sphericity indicated that significant sphericity existed within the data. That is, the variance between test 1, 2, and 3 were unequal. Significant sphericity in the data can cause the statistical test to be inflated, resulting in a greater risk of Type 1 error. To account for this sphericity the Greenhouse-Geisser correction factor was used. The results of the one-way repeated measures ANOVA comparing students' absolute accuracy scores at three different times showed a significant effect ($F(1.89,406) = 14.53$, $p < .001$). A protected t test revealed that there was no significant change from test 1 ($M = .85$, $SD = .50$) to test 2 ($M = .86$ $SD = .47$). A significant increase did exist between test 2 ($M = .86$ $SD = .47$) to test 3 ($M = 1.06$ $SD = .52$).

Research Question 3

Research question 3 looked at the degree to which differences in CJOL accuracy exist amongst disaggregated groups of age, gender, year in school, and course level.

Differences by age. A one-way ANOVA was carried out to assess the differences in absolute accuracy scores amongst these disaggregated groups. The mean absolute accuracy for

each age group is shown in Table 11. Due to a violation of homogeneity of variances, a Welch ANOVA test was carried out to analyze differences in absolute accuracy scores between age groups ($F(4,1795.5) = 8.948$, $p < .001$). A Games-Howell post-hoc test indicated a statistically significant mean differences between group l (Age <= 20) and group 2 (Age > 20 & Age <= 25) of -.12. A statistically significant mean difference was also found between age group 2 (Age > 20 & Age <= 25) and age group 5 (Age > 35) of .19.

Table 11

Mean Absolute Accuracy by Age Group

Age Group	Mean	Std. Dev.
Less than 20	1.01	1.2
Between 20 & 25	0.96	1.2
Between 26 & 30	1.13	1.3
Between 31 & 35	1.05	1.1
Greater than 35	1.00	1.2

Differences by gender. No statistically significant differences were found between disaggregated groups of gender.

Differences by year in school. Due to a violation of homogeneity of variances, a Welch ANOVA test was carried out to analyze differences in absolute accuracy scores between year in school groups ($F(3,1492.9) = 22.02$, $p < .001$). A Games-Howell post-hoc test indicated a statistically significant mean differences between freshman and sophomores (MD = .-12), juniors (MD = .-17), and seniors (MD = .-40). A statistically significant mean difference between sophomores and seniors was also found (MD = .-28). The means of juniors and seniors differed significantly (MD = .-23).

Differences by course. Due to a violation of homogeneity of variances, a Welch ANOVA test was carried out to analyze differences in absolute accuracy scores amongst differing course levels. Overall, a significant relationship was found between absolute accuracy and course level ($F(4,1997.37) = 45.39$, $p < .001$). Table 12 highlights significant differences between course as indicated by a Games-Howell post-hoc test.

Table 12

Post-Hoc Test of Differences of Absolute Accuracy Between Course Level

Course 1	Group Mean	Course 2	Group Mean	Mean Difference	Std. Error	*p*
Math Fundamentals	1.09	Foundations for Algebra	1.28	-.19	.07	$p < .045$*
		Integrated Beg. & Inter. Algebra	.69	.41	.06	$p < .000$***
Foundations for Algebra	1.28	Introductory Algebra	1.09	.19	.06	$p < .019$*
		Integrated Beg. & Inter. Algebra	.69	.59	.06	$p < .000$***
		Intermediate Algebra	.99	.29	.06	$p < .000$***
Introductory Algebra	1.09	Integrated Beg. & Inter. Algebra	.69	.40	.04	$p < .000$***
		Intermediate Algebra	.99	.10	.03	$p < .026$*
Integrated Beg. & Inter. Algebra	.69	Intermediate Algebra	.99	-.31	.03	$p < .000$***

Note: * significant at $p < 0.05$, ** significant at $p < 0.005$, *** significant at $p < 0.001$

Discussion

The results of this study indicate that overall developmental math students tended to be quite accurate in their CJOL. This section will address the degree to which study results support each of the three research questions namely: (a) To what degree are developmental math

students' perceptions of their performance match their actual performance accurate (CJOL accuracy)?, (b) To what degree do developmental math students' CJOL become more accurate over time after receiving feedback on their accuracy?, and (c) To what degree do differences in CJOL accuracy exist amongst disaggregated groups of age, gender, year in school, and course level?

To What Degree Are Developmental Math Students' CJOL Accurate?

One of the primary objectives of this study was to determine the degree to which developmental math students are academically self-aware as measured by their CJOL accuracy scores. The first analysis resulted in an average absolute accuracy score of 1 (n= 17,091, $M = 1$, $SD = 1.21$). Differing from how absolute accuracy was reported to students in their feedback emails, absolute accuracy scores represent the deviance from a perfectly accurate score. That is, an absolute accuracy score of 0 would represent perfect accuracy and an absolute accuracy score of 4 would represent complete inaccuracy. These results indicate that students were, on average, one confidence interval away from being perfectly accurate. For example, on average, a student who received 100% of the credit for the math exam question would have reported that they were 60% to 80% sure that they were going to get the question correct. In a situation where the math exam question was dichotomously scored, which represented a strong majority of exam questions in the study, the only way a student could improve upon an absolute accuracy of 1 would be for them to indicate they were 100% confident that they got the item correct (and get the item correct) or indicate that they were 0% confident that they would get the item correct (and get the item incorrect).

This study found that in the context of individual developmental math courses, students tend to be quite accurate in their CJOL. They seemed to know when they knew and did not

know the answer to individual exam items. At first glance, this finding appears to contradict research that reports that students at lower ability (like those taking developmental math) have less accurate CJOL (Bol & Hacker, 2001; DiFrancesca et al., 2016; Dunning et al., 2004; Hacker et al., 2000; Nietfeld et al., 2005; Shake & Shulley, 2014; Valdez, 2013). However, upon closer inspection of these studies, the ability referred to is not a general ability (e.g., students in developmental math have lower math abilities). Rather, the ability referred to in these studies was based on the students' performance on the same task from which the JOL was made (e.g., a student's ability on a specific math question). The results of this study indicate that the low general math ability of developmental math students, as measured by placement tests, is not indicative of their CJOL accuracy within the context of individual developmental math courses. Within a developmental math course there exists a normal distribution of ability levels. This normalization of ability at the course level might allow developmental math students' CJOL to be compared to CJOL from non-developmental student populations.

To determine the nature of the bias found in developmental math students' CJOL the proportion of over, under and perfect estimations were calculated. The results indicate that 19.1% of students' estimations were over estimations of performance, 33.9% of students' estimations were under estimations of performance, and 46.9% of students' estimations were perfect estimations of performance. The percentages of over, under, and perfect estimations were also quite different than expected. Much research has shown that when students err in their JOL they most commonly err in over estimation (Blackwood, 2013; Metcalfe, 2009; Miller & Geraci, 2011a, 2011b). Considering the developmental nature of the course, it was presumed that the students would follow a similar pattern and largely over estimate their performance. However, the results of this study strongly indicate that this was not the case. Of the three

categories, students were most often perfectly accurate (46.9%). Students' percent of under estimations (34.0%) and perfect estimations (46.9%) together accounted for 80.9% of the CJOL results.

The statistically significant relationship between the absolute accuracy scores and item difficulty (r (741) = -.60, $p < .001$) suggests that students' ability to accurately estimate their performance is strongly influenced by the difficulty of the item they are considering. The negative correlations between absolute accuracy scores and item difficulty indicate that as the difficulty of the items on a test increases, the absolute accuracy decreases (i.e., as difficulty scores increases the absolute accuracy scores approach 4).

When viewed in the context of prediction, item difficulties were significant predictors of students' absolute accuracy (F (1,741) = 425.6, $p < .001$, $R^2 = .36$). In other terms, a prediction of absolute accuracy would be 36.0% more accurate than predicting the mean absolute accuracy score, if the item difficulty was known. There are two ways in which future research could study students' CJOL accuracy while accounting for item difficulty. First, research into the development of new CJOL measurement methods could help more fully determine students' actual academic self-awareness after removing the effects of item difficulty from the equation. Second, future research could investigate the possible intervention of informing students of the item difficulty as part of the exam. Providing students with this additional information might heighten their awareness on more difficult items and possibly serve to improve their CJOL accuracy.

The statistically significant relationship between the percent over, under, and perfect estimations and question difficulty (r_{over} (741) = -.82, $p < .001$; r_{under} (741) = .13, $p < .001$; $r_{perfect}$ (741) = .59, $p < .001$) suggests that the difficulty of the question strongly influences whether or

not a student would over, under, or perfectly estimate their performance. Most notable is the correlation of -.82 between the percent of over estimations and the question difficulty. This strong correlation suggests that as the difficulty of the item increases (very few individuals get the item correct) the percent of students who over estimate their performance increases. In other words, the more difficult the item, the more likely a student will think that they answered correctly when in reality they answered incorrectly. The positive correlation of .59 between item difficulty and the percent of perfect estimations suggests that as the test difficulty decreases (very few students get the item wrong) the percent of students who perfectly estimate their ability also increases. Item difficulty was also a significant and meaningful predictor of absolute accuracy, explaining 68.0% of the variance in students over estimating and 35.0% of the variance of students perfectly estimating their performance.

A student's math ability, as represented by an average of the student's exam z-scores, was a significant predictor of the student's absolute accuracy (F (2,282) = 34.9, $p < .001$, $R^2 = .20$). That is, knowing whether a student was a high, average, or low performing student in a class increased the prediction of students' absolute accuracy by 20.0%. In the present study, there was both a statistically significant and meaningful difference between students' math ability and their CJOL accuracy. Students who were in the high ability group had, on average, absolute accuracy scores that were 50% better than students with average ability. These high achieving students were 64% more accurate than students with low ability. Future research should investigate the fundamental differences between how JOL are made at each ability level. Having a better understanding of how and why students in each ability group make their JOL could help provide insights into the creation of interventions targeted for each ability group.

To What Degree do CJOL Becoming More Accurate Over Time?

Another objective of this study was to determine the degree to which developmental math students' CJOL changed over time after receiving feedback on their accuracy. A simple analysis of the mean of absolute accuracy scores at each test interval indicated that students decreased very slightly in their accuracy from test 1 (M = .85, SD = .50) to test 2 (M = .86 SD = .47) with a mean difference of .01. A more distinct change occurred between test 2 and test 3 (M = 1.06 SD = .52) with a mean difference of .20. The change in mean absolute accuracy scores across the three test indicates that students seemed to be getting less accurate in their CJOL. A one-way repeated measures ANOVA with a Greenhouse-Geisser correction factor indicated that this change was significant (F(1.89,406) = 14.53, $p < .001$). The protected t test revealed that the change from test 2 to test 3 was statistically significant. At face value it appears that the process of repeatedly providing JOL and receiving feedback on the CJOL results did not improve students' academic self-awareness accuracy.

However, an alternative explanation emerges when taking into account the test difficulty across the three testing periods. As expected in any courses, a follow-up analysis found that tests were getting more difficult over time ($M_{test\ 1}$ = .80, $SD_{test\ 1}$ = .06; $M_{test\ 2}$ = .75, $SD_{test\ 2}$ = .08; $M_{test\ 3}$ = .71, $SD_{test\ 3}$ = .10). A one-way repeated measures ANOVA with a Greenhouse-Geisser correction factor found that this change in test difficulty was statistically significant (F(1.77,406) = 146.7, $p < .001$). Due to the nature of this study and the data collected, the change in absolute accuracy over time is confounded by a factor of test difficulty. Anecdotal evidence provided by faculty members also suggested that some students might be inclined to not respond to JOL prompts for difficult items. Faculty members also noted that in some circumstances some

students appeared to withhold JOL rather than indicating that they were not confident that their math response was correct.

Future studies should take into consideration the strong relationship between item and test difficulty and the accuracy of CJOL. A more in-depth analysis of missing data might also provide insights into student behavior and the accuracy of their CJOL. One way to account for the confounding aspect of difficulty would be to structure a study in such a way that the assumptions of a multilevel longitudinal model would be met (i.e., more than 20 different courses with a minimum of 2 instructors per course and at least 3 tests administered at multiple points in time). In addition to accounting for item and test difficulty, future studies should more thoroughly investigate the influence and use of the feedback by the students. In the present study it is unknown how much attention the students devoted to the feedback and whether or not their interpretation of the feedback was correct. Future studies could also investigate more thorough interventions designed to improve students' academic self-awareness. With the present study the feedback intervention focused solely on the students' CJOL accuracy. Future interventions might provide students with CJOL feedback and require students to identify the reasons for their miss-calibrations (i.e., what made them think they were right when they were actually wrong). An intervention of this type follows a similar logic to that found in more traditional math performance remediation where students are asked to identify and correct mistakes rather than simply being told whether or not they answered the question correctly. This enhanced intervention may prove to be more effective in improving students' academic self-awareness because it may help them identify and remedy the source of their miss-calibration rather than simply identifying where their CJOL were inaccurate.

To What Degree do CJOL Differ Amongst Disaggregated Groups?

The final objective of this study was to determine the degree to which differences in CJOL absolute accuracy exist amongst disaggregated groups of age, gender, year in school, and course level. Analyses investigating the absolute accuracy scores between age groups revealed statistically significant results ($F(4,1795.5) = 8.948$, $p < .001$). Post-hoc analyses showed that the most meaningful difference occurred between students who age 21-25 and students in age groups of 26-30 and over 35. One possible explanation for these differences could be the time since the student last participated in a math course. Students age 26 or older, taking entry level remedial math courses, have a higher chance of being non-traditional students who have been out of a formal math education experience for several years. This potential gap between these students' last formal math educational experience and their experience in this study could be one possible contributing factor that influenced the accuracy of their academic self-awareness.

This presumption can be partially supported by the results of examining the differences between absolute accuracy and year in school. The results of disaggregating absolute accuracy by year in school revealed that students became less accurate in their CJOL as their year in school increased ($M_{fr} = .95$, $SD_{fr} = 1.18$; $M_{so} = 1.07$, $SD_{so} = 1.26$; $M_{jr} = 1.12$, $SD_{jr} = 1.27$; $M_{sr} = 1.35$, $SD_{sr} = 1.40$). While it is unclear why this occurs, the overall significance of the differences between year in school groups ($F(3,1492.9) = 22.02$, $p < .001$) might suggest that students are less academically self-aware in their remedial math courses when they wait until later on in their educational experience to take these courses or they might choose wait to take their math courses because they struggle with the topic and therefore are already less aware.

The analysis of differences between absolute accuracy in a student's academic self-awareness and math courses were also statistically significant ($F(4,1997.37) = 45.39$, $p < .001$).

Post-hoc tests indicated that there were significant mean differences between all courses with the exception of Math Fundamentals and Introductory Algebra. Although there were significant differences amongst the different courses, there was no distinguishable pattern in mean differences. The average mean difference between Integrated math course (integrating Beginning and Intermediate Algebra) and all other math courses was quite large (Ave. *MD* = -.43) denoting that students in the Integrated course were almost twice as accurate than students in all other courses. This large difference in absolute accuracy could possibly be explained by the accelerated nature of the Integrated course (i.e., the course was harder but typically only students with a propensity to do well take the course). The Integrated course is a fast paced course that condenses the content of two semester courses into one. Students' choice to take this more rigorous course may indicate that they are already more academically self-aware, as evidenced by their self-assessment of their preparation for the course.

Conclusions

The present study investigated the use of CJOL in university developmental math courses. More specifically, this study sought to answer the following research questions: (a) to what degree are developmental math students' perceptions of their performance match their actual performance accurate (CJOL accuracy), (b) to what degree do developmental math students' CJOL become more accurate over time after receiving feedback on their accuracy, and (c) to what degree do differences in CJOL accuracy exist amongst disaggregated groups of age, gender, year in school, and course level.

One of the main findings was that these developmental math students generally were quite academically self-aware. When developmental math students were inaccurate in their CJOL they tended to slightly under estimate their performance rather than over estimate their

performance. This might simply be a response set issue (i.e., students are unwilling to select an extreme position on the response scale indicating they are 100% sure they were correct). The general exception to this is when math items had a high level of difficulty. Students tended to more frequently over estimate their performance on difficult items. In the context of specific developmental math courses, high performing students were consistently more accurate than lower performing students. Students' CJOL accuracy decreased over the course of the study as the difficulty of the items and tests increased. This lack of improved academic self-awareness merits further investigation through the use of more robust methodologies that take into account the increasing difficulty of the exams being taken in a course. Intervention methods, such as students' use of the feedback provided, should also be more thoroughly investigated. Although much work remains in researching the use of CJOL in applied educational context, the potential for improving academic outcomes through increasing student academic self-awareness remains. By continuing to bring research on CJOL and academic self-awareness out of the lab and into applied settings, students will have increased opportunities to develop their academic self-awareness and become more successful in their academic pursuits.

References

Act. (2014). *The Condition of College & Career Readiness 2014*. *The Condition of College & Career Readiness 2014 Report*. Retrieved from www.act.org/readiness/2014

Adams, C. (2014, October 29). State efforts fuel ACT, SAT growth. *Education Week,* pp. 10-11.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educatiaonl and Psychological Testing*. Washington D.C.: American Educational Research Association.

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*(3), 432–447. http://doi.org/10.1037/a0015928

Blackwood, T. (2013). Business undergraduates' knowledge monitoring accuracy: How much do they know about how much they know? *Teaching in Higher Education*, *18*(1), 65–77.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2005). *Handbook of self-regulation*. Burlington, MA: Academic Press.

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

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

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28,* 2, 353-375. http://doi.org/10.1007/s10648-015-9311-9

Chung, M. K. (2000). The development of self-regulated learning. *Asia Pacific Education Review*, *1*(1), 55–66.

Cohen, M. (2012). The importance of self-regulation for college student learning. *College Student Journal*, *46*(4), 892–902.

Credé, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, *3*(6), 425–453. http://doi.org/10.1111/j.1745-6924.2008.00089.x

de Bruijn-Smolders, M., Timmers, C. F., Gawke, J. C. L., Schoonman, W., & Born, M. P. (2016). Effective self-regulatory processes in higher education: Research findings and future directions. A Systematic Review. *Studies in Higher Education*, *41*(1), 139–158.

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*(3), 347–374.

DeVellis, R. F. (2012). *Scale development: Theory and applications*. Thousand Oaks, CA: SAGE.

DiFrancesca, D., Nietfeld, J. L., & Cao, L. (2016). A comparison of high and low achieving students on self-regulated learning variables. *Learning and Individual Differences, 45,* 228-236. http://doi.org/10.1016/j.lindif.2015.11.010

Dignath, C., Buettner, G., & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review*, *3*(2), 101–129.

Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition*, *25*(5), 691–700. http://doi.org/10.3758/BF03211311

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*(3), 69–106. http://doi.org/10.1111/j.1529-1006.2004.00018.x

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*(3), 199–212. http://doi.org/10.1002/bdm.495

Fitzgerald, J. T., White, C. B., & Gruppen, L. D. (2003). A longitudinal study of self-assessment accuracy. *Medical Education*, *37*(7), 645–649.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. http://doi.org/10.1037/0033-295X.98.4.506

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*(3), 334–372. http://doi.org/10.3102/003465430303953

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101–121.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170. http://doi.org/10.1037/0022-0663.92.1.160

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

Hartwig, M. K., Was, C. A., Isaacson, R. M., & Dunlosky, J. (2012). General knowledge monitoring as a predictor of in-class exam performance. *British Journal of Educational Psychology*, *82*(3), 456–468.

Kitsanas, A. (2002). Test preparation and performance: A self-regulatory analysis. *Journal of Experimental Education*, *70*(2), 101–113.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. http://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, *123*(3), 297–315. http://doi.org/10.1037/0096-3445.123.3.297

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517. http://doi.org/10.1037/0033-295X.103.3.490

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118. http://doi.org/10.1037/0278-7393.6.2.107

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*(6), 1624–1637.

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology*, *80*(3), 325–340. http://doi.org/10.1348/000709910X485719

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006, July). What matters to student success: A review of the literature. In *Commissioned report for the national symposium on postsecondary student success: Spearheading a dialog on student success*. Retrieved from http://nces.ed.gov/npec/pdf/kuh_team_report.pdf

Ley, K., & Young, D. B. (1998). Self-regulation behaviors in underprepared (developmental) and regular admission college students. *Contemporary Educational Psychology*, *23*(1), 42–64. http://doi.org/10.1006/ceps.1997.0956

Lin, L.-M., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology*, *22*(2), 111–128. http://doi.org/10.1080/027027101300213083

Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*(4), 345–391. http://doi.org/10.1006/ceps.1998.0972

Luftenegger, M., Schober, B., van de Schoot, R., Wagner, P., Finsterwald, M., & Spiel, C. (2012). Lifelong learning as a goal—Do autonomy and self-regulation in school result in well prepared pupils? *Learning and Instruction*, *22*(1), 27–36.

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. J. Perfect, B. L. Schwartz (Eds.), *Applied metacognition.* (pp. 39–67). New York, NY: Cambridge University Press. http://doi.org/10.1017/CBO9780511489976.004

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*(4), 723–731.

McKelvie, S. J. (1978). Graphic rating scales: How many categories? *British Journal of Psychology*, *69*(2), 185–202. http://doi.org/10.1111/j.2044-8295.1978.tb01647.x

Meier, B., von Wartburg, P., Matter, S., Rothen, N., & Reber, R. (2011). Performance predictions improve prospective memory and influence retrieval experience. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *65*(1), 12–18. http://doi.org/10.1037/a0022784

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159–163. http://doi.org/10.1111/j.1467-8721.2009.01628.x

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

Metcalfe, J., & Finn, B. (2008b). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1084–1097.

Mihalca, L., Mengelkamp, C., Schnotz, W., & Paas, F. (2015). Completion problems can reduce the illusions of understanding in a computer-based learning environment on genetics. *Contemporary Educational Psychology*, *41*, 157–171. http://doi.org/10.1016/j.cedpsych.2015.01.001

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Boston, MA: Pearson.

Miller, T. L., Duffy, S. E., & Zane, T. (1993). Improving the accuracy of self-corrected mathematics homework. *The Journal of Educational Research*, *86*(3), 184–189. http://doi.org/10.1080/00220671.1993.9941157

Miller, T. M., & Geraci, L. (2011a). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning*, *6*(3), 303–314.

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

Mishkind, A. (2014). Overview: State Definitions of College and Career Readiness. *College and Career Readiness and Success Center*. Retrieved from http://eric.ed.gov/?id=ED555670

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*(4), 207–213. http://doi.org/10.1111/j.1467-9280.1994.tb00502.x

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *Journal of Experimental Education*, *74*(1), 7.

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *Journal of Educational Research*, *95*(3), 131–142.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, *36*(2), 89–101. http://doi.org/10.1207/S15326985EP3602_4

Perels, F., Gurtler, T., & Schmitz, B. (2005). Training of self-regulatory and problem-solving competence. *Learning and Instruction*, *15*(2), 123–139.

Pilegard, C., & Mayer, R. E. (2015). Adding judgments of understanding to the metacognitive toolbox. *Learning and Individual Differences*, *41*, 62–72. http://doi.org/10.1016/j.lindif.2015.07.002

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. *Reading Research Quarterly*, *22*(2), 219–236. http://doi.org/10.2307/747666

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford, M. C. O'Connor, B. R. (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction.* (pp. 37–75). New York, NY: Kluwer Academic/Plenum Publishers.

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Educational Psychology*, *79*(4), 749–767.

Roth, A., Ogrin, S., & Schmitz, B. (2015). Assessing self-regulated learning in higher education: a systematic literature review of self-report instruments. *Educational Assessment, Evaluation and Accountability*, 1-26. http://doi.org/10.1007/s11092-015-9229-2

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology*, *19*(2), 143–154. http://doi.org/10.1006/ceps.1994.1013

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33–45.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460–475. http://doi.org/10.1006/ceps.1994.1033

Schraw, G., & Others, A. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology*, *18*(4), 455–463.

Schunk, D. H., & Zimmerman, B. J. (1997). Social origins of self-regulatory competence. *Educational Psychologist*, *32*(4), 195–208. http://doi.org/10.1207/s15326985ep3204_1

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales. *Public Opinion Quarterly*, *55*(4), 570–582.

Shake, M. C., & Shulley, L. J. (2014). Differences between functional and subjective overconfidence in postdiction judgments of test performance. *Electronic Journal of Research in Educational Psychology*, *12*(2), 263–282.

Shaughnessy, J. J. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality, 13,* 4, 505-14.

Shivpuri, S., Schmitt, N., Oswald, F. L., & Kim, B. H. (2006). Individual differences in academic growth: Do they exist, and can we predict them? *Journal of College Student Development*, *47*(1), 69–86.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204–221. http://doi.org/10.1037/0278-7393.26.1.204

Sparks, D., & Malkus, N. (2013). First-year undergraduate remedial coursetaking: 1999–2000, 2003–04, 2007–08, (January). Retrieved from http://files.eric.ed.gov/fulltext/ED538339.pdf

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*(6), 971–986. http://doi.org/10.1016/S0191-8869(96)00130-4

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review*, *12*(4), 437–475.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. http://doi.org/10.1037/0022-0663.95.1.66

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://doi.org/10.1037/0278-7393.25.4.1024

Valdez, A. (2013). Student metacognitive monitoring: predicting test achievement from judgment accuracy. *International Journal of Higher Education*, *2*(2), 141–146.

Van Overschelde, J. P., & Nelson, T. O. (2006). Delayed judgments of learning cause both a decrease in absolute accuracy (calibration) and an increase in relative accuracy (resolution). *Memory & Cognition*, *34*(7), 1527–1538. http://doi.org/10.3758/BF03195916

Vössing, J., & Stamov-Roßnagel, C. (2016). Boosting metacomprehension accuracy in computer-supported learning: The role of judgment task and judgment scope. *Computers in Human Behavior*, *54*, 73–82. http://doi.org/10.1016/j.chb.2015.07.066

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology*, *81*(3), 435–437. http://doi.org/10.1037/0022-0663.81.3.435

Was, C. A. (2014). Discrimination in measures of knowledge monitoring accuracy. *Advances in Cognitive Psychology*, *10*(3), 104–112. http://doi.org/10.5709/acp-0161-y

Zimmerman, B. J., & Kitsantas, A. (1996). Self-regulated learning of a motoric skill: The role of goal setting and self-monitoring. *Journal of Applied Sport Psychology*, *8*(1), 60–75. http://doi.org/10.1080/10413209608406308

Zimmerman, B. J., & Kitsantas, A. (1997). Developmental phases in self-regulation: Shifting from process goals to outcome goals. *Journal of Educational Psychology*, *89*(1), 29–36. http://doi.org/10.1037/0022-0663.89.1.29

Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, *23*(4), 614–628. http://doi.org/10.2307/1163093

Zimmerman, B. J., & Schunk, D. H. (Eds.) (2001). *Self-regulated learning and academic achievement: Theoretical perspectives*. Mahwah, NJ: L. Erlbaum.

Zimmerman, B., & Schunk, D. (2011). Self-regulated learning and performance: An introduction and an overview. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance: Educational psychology handbook series* (pp. 1–32). New York, NY: Routledge.

APPENDIX A: Confidence Calibration Assessment (CCA)

Instructions

1. Write your UVU Student ID in the space above.
2. Answer the math question on your exam.
3. Mark an **X** in the space provided below to indicate how confident you are that you answered the math question correctly.
4. Repeat steps 2 & 3 until you have answered all the questions on your math exam.
5. Turn in your math test and this sheet of paper to your instructor.

EXAMPLE TEST QUESTION: What is 1+1? **EXAMPLE ANSWER:** 7,240

Math Test Question #	How confident are you that you answered the test question correctly?				
	0% - 20%	20% - 40%	40% - 60%	60% - 80%	80% - 100%
Example	X				

Math Test Question #	How confident are you that you answered the test question correctly?					
	0% - 20%	20% - 40%	40% - 60%	60% - 80%	80% - 100%	
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						

Math Test Question #	How confident are you that you answered the test question correctly?					
	0% - 20%	20% - 40%	40% - 60%	60% - 80%	80% - 100%	
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						

APPENDIX B: Consent to be a Research Subject

Introduction – What this study is about:
This research study is being conducted by Brian Jones, a UVU Institutional Research Analyst and Instructional Psychology graduate student. The purpose of this study is to research how accurate students are when asked to rate how confident they are in their answers on math tests. Participation in this study will require approximately 2-5 minutes each time you take a math test (depending on the length of your math test).

Procedures – What we are asking you to do:
If you agree to participate in this research study, you will be asked to complete a survey that will accompany your math tests. The survey will consist of rating how confident you are in your answer to each of the questions on your math test (example shown below).

Math Test Question #	How confident are you that you answered the test question correctly?				
	0% -20%	20% to 40%	40% to 60%	60% to 80%	80% to 100%
1				X	
2					X
3		X			

As shown in the example above, if your math test has three questions on it, you would be asked to answer the survey question three times. It is expected that it will take approximately 10 seconds to answer each survey question.

Use of Your Data
All survey data will be identified using your UVU ID #. Once the survey results have been collected they will be combined with your scores on math tests, your UVU demographic information (gender, age, year in school, course level), and your UVU entrance exam scores (ACT and or Accuplacer). Your information regarding your test scores and demographics are protected under the Family Educational Rights and Privacy Act (FERPA). By consenting to participate in this study you are consenting to allow researchers to access and use this protected educational information as part of the research study.

Confidentiality
All data from this study will be confidential. This means that your identity will be known to researchers but this information will only be available and used for research purposes. Any data reported publicly will be in a summarized form (e.g., class averages and totals) and will not include any identifiable information.

Risks & Benefits
The risks for participating in this study include a potential breach of confidentiality or intrusion to private information. This risk will be minimized by storing physical copies of survey results in a locked file and by keeping all personally identifiable information on UVU's secure servers. All personally identifiable information will be removed and destroyed immediately following the linking of survey results to test scores, demographic information, and entrance exam scores.

Previous research has shown that one of the possible benefits to participating in this study is the potential improved performance on your math assessments. You will also be helping to contribute to a better understanding of student success interventions, contributing to the development of early warning systems that help to provide academic supports to students, and the development of systems for providing more meaningful feedback to students.

Voluntary Participation
You have been invited to participate in this study solely upon the basis of your enrollment in a developmental math course at UVU and your instructor's willingness to facilitate your participation. Participation in this study is voluntary. You have the right to withdraw at any time or refuse to participate entirely without any risk to your current or future relationship with your instructor or Utah Valley

University. Under no circumstances will your grade be influenced by your choice to either consent or decline participation in the study.

Questions about the Research

If you have any questions regarding this study, you may contact Brian Jones at ********@uvu.edu or ###-###-#### for further information.

Questions about Your Rights as Research Participants

If you have any questions regarding your rights as a research participant contact UVU IRB at ###-###-#### Room ##. Reference UVU IRB Tracking #####.

Statement of Consent

I have read the above information and give my consent to participate in this study.

(Individuals must be 18 years or older to participate)

_____

First & Last Name (Print)

_____

UVU ID #

_____

First & Last Name (Signature)

_____

Date

APPENDIX C: Math Exam Score Reporting Template

Math Exam Score Reporting Example						
Survey Scoring	Survey Values	0% - 20%	20% - 40%	40% - 60%	60% - 80%	80% - 100%
	Reporting Values	0	1	2	3	4
Student Name or UVU ID	Exam Question #	Math Q1	Survey R1	Math Q2	Survey R2	Math Total
	Possible Points Per Question	4	(0-4)	1	(0-4)	5
18279164	[redacted]	3	1	1	0	4
John Doe UVU Student		3	3	1	1	4
12095755		3	2	1	0	4
17362878		0	1	1	0	1
Jane Doe UVU Student		2	2	1	0	3
11789986		1	0	1	1	2
19047281		1	0	1	0	2
John Doe UVU Student		0	2	1	2	1
19433147		3	0	1	0	4
15762728		1	0	1	1	2
10344920		2	0	1	4	3
18367640		2	3	1	3	3
11623571		0	3	1	4	1
19792247		0	4	1	1	1
Jane Doe UVU Student		0	1	1	0	1
16653100		1	3	1	4	2
16277639		0	0	1	0	1
14485694		1	2	1	1	2

APPENDIX D: Student Feedback

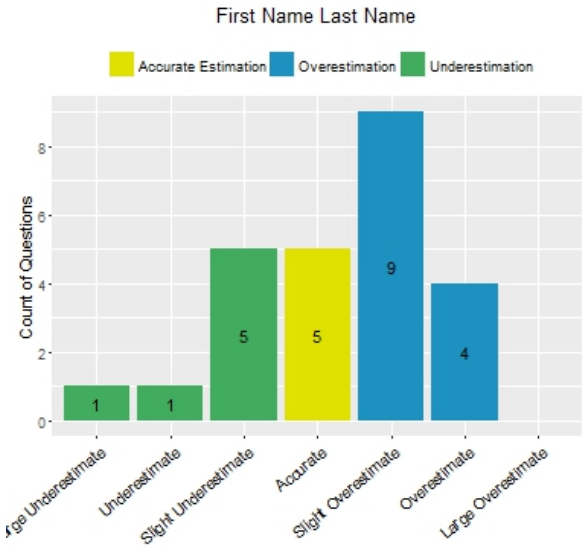
Math Course Exam # Academic Self-Awareness Feedback

Dear First Name Last Name,

Academic Self-Awareness is a measurement of how well you predicted you would do on your math test compared to how well you actually did. General Academic Self-Awareness Scores range from 0 to 4. The higher your score the more accurate your predictions were.

General Academic Self-Awareness

Your general academic self-awareness score on this test was	**2.56 out of 4**
Generally your predictions were	**Somewhat Accurate**
In terms of your performance, you generally tend to	**slightly over estimate**



The graph above shows how many times your predictions were over or under your actual performance.

Questions to Review

To improve your accuracy, review the following test questions with some of the most inaccurate predictions.

Greatly Under Estimated	**17, 19**
Greatly Over Estimated	**14, 7, 12, 13**

APPENDIX E: Teacher Feedback

Course Instructor Exam # Report

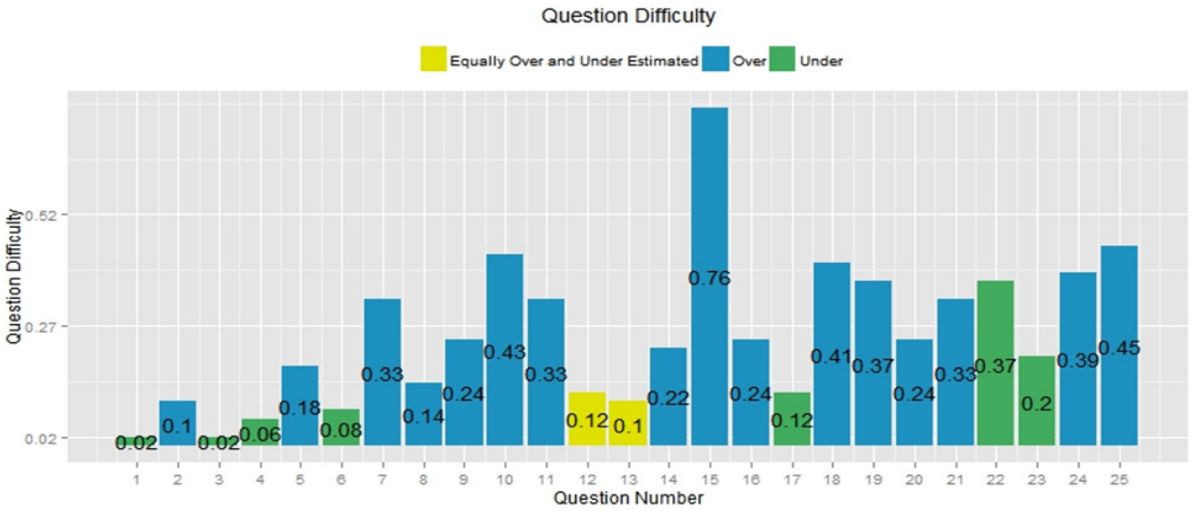
This document contains an item analysis for both Exam 1 math questions and student confidence estimations.

Question Difficulty

- Question Difficulty = Points Possible – Average Student Score
- A higher number represents a more difficult question
- The color on the graph indicates if students on average over or under estimated their performance

The Graph below indicates the difficulty of each of the math questions.

The top 5 most difficult questions were: 10, 15, 18, 24, 25.



Item Discrimination

- Item discrimination can help to identify poorly constructed questions
- Negative discrimination values are not desired. They mean that students who received a high score on the test got the question wrong and that students who got low scores on the test got the question right.
- Negative discrimination values typically mean that something was wrong with the question or the scoring of the question.

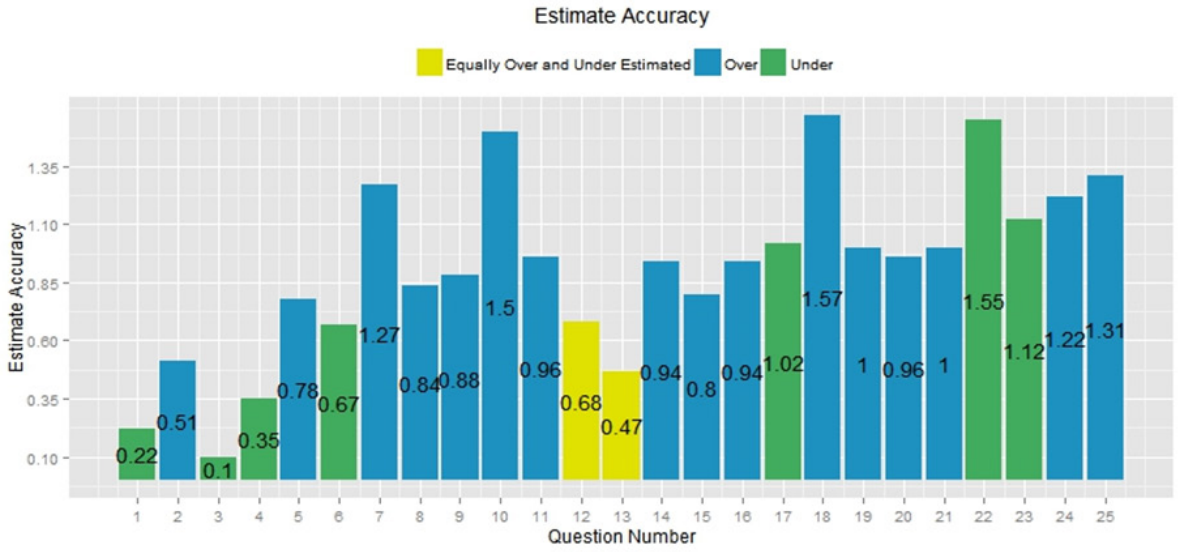
The following questions had negative discrimination on Exam #: **There were no negatively discriminating questions on this test.**

Accuracy of Student Estimations

- The accuracy of estimation is based on how close student estimates of performance were to their actual performance.
- The values on the graph below represent how many confidence intervals off student's estimates were from being perfectly accurate.
    - ❖ High values = less accurate estimations
    - ❖ Example: A question with an estimation accuracy of 1 means that students estimations were on average 1 confidence interval above or below their actual performance
- The color on the graph indicates if students on average over or under estimated their performance

Students were most accurate on questions: 1, 2, 3, 4, 13

Students were least accurate on questions: 7, 10, 18, 22, 25



Individual Student Performance

Attached to this email is a copy of the feedback that was sent to each of your students.